The Dissertation Committee for Austin Madison Mulloy certifies that this is the approved version of the following dissertation:

A Monte Carlo Investigation of Multilevel Modeling

in Meta-Analysis of Single-Subject Research Data

Committee:

Mark O'Reilly, Co-Supervisor

Susan (Tasha) Beretvas, Co-Supervisor

Nina Zuna

Terry Falcomata

Keenan Pituch

A Monte Carlo Investigation of Multilevel Modeling in Meta-Analysis of Single-Subject Research Data

by

Austin Madison Mulloy, B.A.; M.Ed.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree

Doctor of Philosophy

The University of Texas at Austin

August 2011

A Monte Carlo Investigation of Multilevel Modeling

in Meta-Analysis of Single-Subject Research Data

Austin Madison Mulloy, Ph.D.

The University of Texas at Austin, 2011

Co-Supervisors: Mark O'Reilly and Susan (Tasha) Beretvas

Multilevel modeling represents a potentially viable method for meta-analyzing single-subject research, but questions remain concerning its methodological properties with regard to characteristics of single-subject data. For this dissertation, Monte Carlo methods were used to investigate the properties of a 3 level model (i.e., with a quadratic equation at level 1), and three different level 1 error specifications (i.e., different variance components and covariances of 0, lag-1 autoregressive covariance structures, and separate error terms for each phase, with different variance components and covariances of 0). Data for simulated subjects were generated to have characteristics typical of published single-subject data (e.g., typical variances and magnitudes of effect). Samples were simulated for conditions which varied in number of data points per phase, number of subjects per study, number of studies meta-analyzed, level of autocorrelation in residuals, and continuity of variance across phases. Outcome variables examined included rates of convergence of analyses, power for statistical tests of fixed effects, and relative parameter bias of estimates of fixed effects, random effects' variance components, and autocorrelation estimates.

Convergence rates were found to be 100% for all level 1 error specifications and data conditions. Power for statistical tests of fixed effects was observed to be adequate when 10 or more data points were generated per phase and 60 or more total subjects were included in meta-analyses. The relative biases of estimates of fixed effects were found to have limited associations with numbers of data points per phase, levels of autocorrelation, and the continuity/discontinuity of variance across phases. Random effects' variance components were observed to be frequently biased. Associations between relative bias and data conditions were found to vary by random effect. Finally, autocorrelation estimates were found to be biased in all conditions for which autocorrelation was generated. Results are discussed with regard to study strengths and limitations, and their implications for the meta-analysis of single subject data and primary single subject research.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Evidence-Based Practice

In the past two decades, special educators have become increasingly attentive to the presence or absence of research support for classroom practices. The growing concern for evidence has shaped special education-related legislation (e.g., No Child Left Behind Act [NCLB], 2001; Individuals with Disabilities Education Improvement Act [IDEIA], 2004), teacher education curricula (Eren & Brucker, 2011; Kutash, Duchnowski, & Lynn, 2009), and practices at the school and classroom levels (Burns & Ysseldyke, 2009). While many in the field agree that evidence on educational practices should be carefully considered prior to making curricular decisions, dialogue on how to conduct such consideration is on-going (Gersten, Fuchs, Compton, Coyne, Greenwood, & Innocenti 2005; Horner, Carr, Halle, McGee, Odom, & Wolery, 2005; Mayton, Wheeler, Menendez, & Zhang 2010; Odom, Brantlinger, Gersten, Horner, Thompson, & Harris, 2005; Shadish, Rindskopf, & Hedges, 2008).

Much research on special education populations is performed using single-subject experimental designs (SSED). At present, consensus does not exist regarding how to summarize and synthesize data from multiple single-subject experiments. SSEDs make use of an inductive experimental method. Consequently, findings only provide insight on the single individual studied. In order to make inferences about what effects educational practices will have on other students in the population, SSED results must be synthesized (Van den Noortgate & Onghena, 2003a). Various authors have proposed procedures for synthesizing research findings (e.g., Beretvas & Chung, 2008b; Center, Skiba, & Casey, 1985–1986; Faith, Allison, & Gorman, 1996; Lundervold, & Bourland, 1988; Parker & Vannest, 2009; Scotti, Evans, & Meyer, 1991; Scruggs, Mastropieri, & Casto, 1987; Van den Noortgate & Onghena, 2003a). However, the synthesis

procedures possess a number of limitations (e.g., conditional applicability; Allison & Gorman, 1994; Beretvas & Chung, 2008b) and, in certain instances, flaws (e.g., loss or misrepresentation of information; Salzberg, Strain, & Baer, 1987; White, 1987). In order to continue to improve special education programs, the limitations and flaws must be overcome. Researchers must develop robust synthesis techniques that produce accurate and nuanced understandings of the efficacy of educational practices.

Importance of evidence-based practice. Knowledge of the efficacy of educational practices is of critical importance to the field of special education. For one, schools and special education departments have limited resources and limited time with students. To best serve students with special needs, resources should be spent on practices confirmed to be most effective (NCLB, 2001). Unfortunately, many invalid, as well as dubious and untested educational practices are commonly employed (Jacobson, Foxx, & Mulick, 2005; Green, 2007; Green, Pituch, Itchon, Choi, O'Reilly, & Sigafoos, 2006). Knowledge of evidence-based practice (EBP) should be expanded and disseminated to avoid squandering resources on ineffective methods.

Further, the growth trajectories of students with special needs can be greatly enhanced by early, effective intervention. Infants, toddlers, and preschoolers who are at risk for developmental delay and receive high quality early intervention services typically attain higher levels of functionality later in life than their peers who do not (Brown, Odom, & Conroy, 2001; Guralnick, 2004; Shonkoff & Phillips, 2000). In the early years of children's development, the brain's rapid growth and ability to self-correct offers a window of opportunity to reverse or minimize the effects of such conditions as brain injury, chromosomal anomalies, and environmental stress (Diamond & Hopson, 1999; McCain & Mustard, 1999). To maximize the impact of early intervention, educators must understand which practices are most effective, for whom they work

best, and what types of additional supports or variations on the practice are necessary for different categories of students.

The established methods of summarizing and synthesizing single-subject research data are limited in their ability to compare interventions, explain variations across studies, and accurately represent data phenomena. To best serve students with special needs, researchers must continue to explore and test various ways of summarizing and synthesizing single-subject research data.

History of evidence-based practice. Despite the logical appeal of grounding important decisions in research data, the consultation of evidence is a relatively recent development. The EBP movement in special education traces its roots to the field of medicine and the scourge of scurvy in the mid-18th century (Singh & Ernst, 2008).

In 1747, a Scottish naval surgeon, James Lind, performed the first controlled clinical trial as part of his effort to treat scurvy in sailors under his care (*ibid*). In those days, the cause of scurvy was still unknown. Lind had the bright idea to give different sailors different treatments and compare the results. The surgeon gathered 12 sailors who had similar symptoms and arranged identical sleeping and diet conditions for each. Then Lind divided the sailors into 6 pairs, gave each pair a different treatment, and made daily observations of their health. On a hunch, he included lemons and oranges as a treatment alongside five in vogue, but ultimately ineffective treatments. After just 6 days, the results were clear. The lemons and oranges had relieved the symptoms of a pair of sailors, while the other 10 remained in poor health. Lind's novel exercise of experimental control over symptoms, sleeping arrangements, diet, and treatment type allowed him to confidently conclude that the difference in health outcomes was due to the treatments.

A half century later, in 1809, a Scottish military surgeon, Alexander Hamilton, advanced Lind's method and performed the first randomized clinical trial (*ibid*). Hamilton was stationed in

a battlefield medical tent and cared for wounded soldiers. At the time, bloodletting was a popular panacea endorsed by many reputable physicians. Hamilton doubted the efficacy of bloodletting and sought to prove his position. To do so, he devised a plan to assign new patients indiscriminately and alternately to treatment involving bloodletting and treatment not involving bloodletting. Hamilton then made efforts to otherwise standardize the care and comforts provided to all patients. Since bloodletting was touted as a panacea, Hamilton included patients with all forms of medical need in his study. Over the next several months, the surgeon kept records of the death rate of the hundred or so soldiers assigned to each condition. Hamilton eventually obtained the proof he sought. Bloodletting was associated with roughly ten times as many deaths as treatment not involving bloodletting. Hamilton's novel use of random assignment to treatment conditions, in addition to his exercise of control over other aspects of patient care, allowed him to conclude, with greater confidence than Lind, that the difference in death rates was attributable to bloodletting. Random assignment represented an improvement in that it prevented the existence of systematic differences between and within treatment groups (Kazdin, 2003).

Over the next 150 years, research methods and scientific knowledge developed considerably. Several generations of scientists elaborated and improved upon Lind and Hamilton's methodologies, and used them to produce new understandings. Beginning in the mid-1900s, a movement for evidence-based medicine began to coalesce (*ibid*). Physicians and researchers united in an effort to close the gap between research knowledge and common practice (Odom et al., 2005). The medical professionals had much success over the next several decades in reforming both medical education and practice. Then, in 1992, the term "evidence-based medicine" was coined and first appeared in print (Guyatt, Cairns, Churchill, Cook, Haynes, Hirsh, et al., 1992).

Developments in the field of medicine, as well as parallel progress made in psychology, educational psychology, sociology, and anthropology, stimulated developments in the field of special education (Odom et al., 2005). The medical and social sciences established the concern for evidence, as well as generated a number of research designs and analysis techniques, which were gradually adopted by special education researchers. These adopted methodologies included experimental and quasi-experimental group designs, SSED, qualitative designs, univariate and multivariate statistical procedures, and meta-analysis.

Around the mid-1990s, special education researchers began a focused campaign to identify evidence-based practices for use with students with disabilities (Kutash, Duchnowski, & Lynn, 2009). Academics and research institutes performed many syntheses and meta-analyses on a variety of topics (e.g., Forness, Kavale, Blum, & Lloyd, 1997; Gersten, Schiller, & Vaughn, 2000; Odom & Wolery, 2003). Such efforts are on-going today (e.g., National Autism Center, 2009). As in medicine, the goal of identifying EBP has been part of a larger effort to close the gap between research and practice (Greenwood, 2001).

The term "evidence-based practice" first appeared in an education-related journal in 1999 (Richman, Reese, & Daniels, 1999). The phrase was introduced by a team of researchers from a medical school with expertise in both developmental pediatrics and applied behavior analysis. As such, the team served as one of many conduits for philosophy and practice from the field of medicine to the field of education. Over the next several years, the field of special education embraced the term EBP and worked toward establishing an operational definition. One general version was offered by Dunst, Trivette, and Cutspec in 2002. Their definition states evidence-based practices are "informed by research, in which the characteristics and consequences of environmental variables are empirically established and the relationship directly informs what a practitioner can do to produce the desired outcome" (p. 3).

The movement for EBP in education was substantially accelerated by the No Child Left Behind Act of 2001 (NCLB, 2001). The law mandated that public schools must use scientifically validated educational practices. In the act, the term "evidence-based practice" is used 110 times in discussions of how to improve the education offered to students (Slavin, 2002).

Today, EBP constitutes a "buzz word" and a priority for teachers, administrators, teacher educators, and researchers alike (Burns & Ysseldyke, 2009). Policy makers regularly invest large amounts of time and money in efforts to determine and support the use of EBP (Kutash, Duchnowski, & Lynn, 2009). Additionally, various organizations now exist for the explicit purpose of furthering the EBP movement (e.g., Center for Evidence-Based Practice, Campbell Collaborative, What Works Clearinghouse). Together, the organizations and education professionals have created a framework for engendering and implementing EBP. The process is now understood to involve (a) primary research, (b) synthesis of primary research, (c) model building for translation of research knowledge to practice, and (d) information dissemination and training (Pucketts Institute, 2009).

Determining Evidence-Based Practice from Single-Subject Research

Determining EBP from single-subject research involves the phases of primary research and synthesis of primary research. In response to the EBP movement, a number of authors and organizations have recently defined quality indicators for primary research using SSED. Across the authors and organizations, much agreement exists regarding how to best structure SSED. However, as stated above, little to no consensus exists regarding how to best synthesize data from SSED.

Standards for primary research. Representatives of the American Psychological Association (APA) and Council for Exceptional Children have defined standards for primary research using SSED (Kratochwill & Stoiber, 2002; Smith, Strain, Snyder, Sandall, McLean,

Boudy-Ramsey, et al., 2002; Horner et al., 2005). The authors and their organizations commonly identify a set of 5 standards that primary research must meet in order for research findings to inform notions of EBP. In brief, these standards require (a) definition of dependent and independent variables, participants, and settings with sufficient precision to allow replication, (b) repeated measurement of dependent variables with quantifiable indices, (c) assessment of the reliability of dependent variable measurements and fidelity of implementation of independent variables, (d) experimental control over threats to interval validity, and (e) replication of results within and/or across participants, settings, and/or materials.

While data from SSED may be interpreted with statistical analyses (e.g., randomization tests; Edgington, 1996), researchers traditionally analyze data visually (Horner, 2005; Kennedy, 2005). Such analysis involves systematic visual comparison of levels, trend, and variability in performance during baseline and intervention conditions. Researchers also visually judge the immediacy of effects following implementation or withdrawal of interventions, the proportion of overlap of data in adjacent phases, the magnitude of changes in the dependent variable, and the consistency of data patterns across multiple baseline and/or intervention phases. The main goal of analysis is to appraise whether or not change in the dependent variable is a function of the independent variable. Additional information is gleaned from data phenomena, when possible. Due to the singular focus of SSED, results of single subjects do not generalize to populations. Consequently, the synthesis of individual outcomes constitutes an important step in the process of determining EBP.

Narrative review as a synthesis method. Until roughly twenty years ago, findings from single-subject research were always synthesized in narrative reviews (Salzberg, Strain, & Baer, 1987). Use of the review method remains popular today (e.g., Chan, Lang, Rispoli, O'Reilly, Sigafoos, & Cole 2009; Schreiber, 2011). Narrative reviews involve descriptions of primary

research outcomes, based on visual analysis, and discussions of the patterns and exceptions in outcomes across studies. Frequently, research outcomes are systematically extracted and pooled using qualitative methodology. When doing so, researchers often make use of coding tables, in which features of studies and their outcomes are categorized or summarized descriptively. However, analysis and synthesis techniques vary widely across researchers.

Narrative review methods tend to work well when synthesizing small numbers of data sets that have fairly clear and undifferentiated patterns. In these conditions, the procedures can produce accurate and nuanced understandings of the efficacy of educational practices (Salzberg, Strain, & Baer, 1987). However, as the number of subjects grows and/ or data patterns become less clear and differentiated, visual analysis and descriptive synthesis become inadequate tools. The limitations of our working memory and the crudeness of eye-balling techniques can lead to inaccurate and unreliable conclusions, omission of relevant information, and/or obscuring of systematic relationships between outcomes and participant or study variables.

Meta-analysis as a synthesis method. Meta-analysis of data from SSED can provide unique opportunities to develop knowledge when reviewing evidence on educational practices (Beretvas & Chung, 2008a; Jenson, Clark, Kircher, & Kristjansson, 2007). In contrast to traditional narrative review methods, meta-analysis' reliance on quantitative metrics can allow for the drawing of more firm and definitive conclusions. Quantitative synthesis offers enhanced objectivity via aggregation of individual summary statistics, statistical testing, and lack of opportunities for authors' possible biases to wield influence. In ideal circumstances, meta-analysis allows researchers to (a) estimate an overall treatment effect, (b) establish confidence intervals for the estimate, (c) test the estimate for statistical significance, (d) compare the estimate to those for other treatments, and (e) answer questions related to variability across studies and moderators of effect (Cooper & Hedges, 1994).

During the previous 20 years, meta-analysis of SSED has represented a point of controversy among researchers. Arguments abound concerning which of the variety of methods to use, the extent of the methods' respective validities, and if quantitative synthesis is appropriate at all (e.g., Allison & Gorman, 1993; Beretvas & Chung, 2008a; Ferron, 2002; Salzberg, Strain, & Baer, 1987; Scruggs, Mastropieri, & Casto, 1987).

At present, the typical SSED meta-analysis (e.g., Shogren, Faggella-Luby, Bae, & Wehmeyer, 2004) involves use of one or more of a variety of flawed non-parametric summary statistics (Beretvas & Chung, 2008a; e.g., Percentage of Non-Overlapping Data [PND; Scruggs, Mastropieri, & Casto, 1987], Percentage of Zero Data [PZD; Scotti, Evans, & Meyer, 1991], Standardized Mean Difference [SMD; Busk & Serlin, 1992], Nonoverlap of All Pairs [NAP; Parker & Vannest, 2009]). The statistics' flaws pertain to their susceptibility to bias and inability to account for common single-subject data phenomena.

For example, the PND and NAP statistics have inverse relationships with the number of baseline data points, such that higher PND and NAP values are probabilistically associated with lower numbers of baseline data points (Allison & Gorman, 1994). The PZD statistic has an inverse relationship with the length of treatment phases past the first zero data point, such that longer treatment phases are probabilistically associated with lower PZD scores. Slow acquisition rates can lead to low PND, NAP, and PZD scores, despite eventual success of treatments in changing or eliminating behaviors (Allison & Gorman, 1994; Scotti, Evans, & Meyer, 1991; White, 1987). Trends in data confound SMD values by creating error in variance estimates and skewing means (Marquis, Horner, & Carr, 2000). Further, PND, NAP, PZD, and SMD statistics only describe level change. Their use brings about a loss of all information related to incremental change and variability.

Additional limitations of the summary statistics concern (a) error resulting from combining statistics across subjects and studies, (b) error resulting from comparing statistics for different treatments or subject groups, and (c) summary statistics lack of utility in moderator analyses. When confounding variables, such as variations in numbers of baseline data points, auto-correlation, or learning curves or other trends are present, accuracy in combining statistics can only be achieved when the individual statistics result from cases with identical circumstances (e.g., same number of baseline data points, same number of treatment data points; Allison & Gorman, 1994; Salzberg, Strain, & Baer, 1987). Similarly, comparing statistics for different treatments or subject groups requires that all cases from which individual statistics are drawn have identical circumstances. In moderator analyses, assessment of parametric and non-parametric correlations between summary statistics and values for hypothesized moderator variables is highly tenuous, and in many cases inappropriate due to (a) the unknown, and likely not normal, underlying distributions of the statistics (Beretvas & Chung, 2008a; Clark-Carter, 2004), (b) auto-correlation, which can be present in single-subject data (Busk & Marascuilo, 1988), and (c) confounding variables, such as those mentioned above.

Despite the popularity of SSED summary statistics, their tendencies to misrepresent and obscure data phenomena make their use in synthesizing research and identifying evidence-based practices inappropriate. The statistics' limitations regarding their combining, comparison, and use in moderator analysis render them inadequate tools for extracting additional insights from a body of research. Pursuit of more sound alternatives is imperative to valid and fruitful practice of SSED meta-analysis.

Multilevel Modeling in Meta-Analysis of Single-Subject Research

Recently, authors have proposed use of multi-level modeling (MLM) techniques in the meta-analysis of SSED (e.g., Van den Noortgate & Onghena, 2003, 2008). The techniques

potentially offer solutions to problems encountered with other methods (Beretvas & Chung, 2008b). For example, MLM estimation is not biased by differences across subjects in numbers of data points collected in each phase, and it can model learning curves and other meaningful fluctuations in time-series data (Raudenbush & Bryk, 2002; Singer & Willet, 2003). Initial inspections of MLM's properties with regard to SSED suggest it is robust to the presence of auto-correlation with regards to type I and type II errors (Jenson et al., 2007). If need be, terms for auto-correlation can be added to models to attenuate the error induced in estimations (Raudenbush, Bryk, Cheong, & Congdon, 2004).

Many of the advantages of MLM result from the procedures' sensitivity to differential effects (Raudenbush & Bryk, 2002). Multilevel models comprise a number of regression equations, which are organized into hierarchical levels and nested within each other to create an overall model. The multiple levels and nesting of equations can allow for modeling of the within-subgroup similarities and between subgroup differences that occur in research contexts. At the lowest level of a model, regression equations calculate the expected dependent variable scores for subjects. At higher levels, the regression equations estimate expected values for regression coefficients from lower levels. Via opportunities to include random effects and predictor variables at each level of a model, dependencies and variation can be accounted for within and between subgroups (e.g., students of the same teacher or school, persons with the same disability diagnosis).

With regard to single-subject data, regression equations at the lowest level can be used to describe changes in subjects' dependent measurements across time (e.g., slopes, curves, and level changes seen in graphs of data). Higher level equations can then be used to describe differences in such changes across subjects, disability groups, and/ or treatments. Due to MLM's sensitivity to variation within and between groups, the procedure has the potential to achieve accuracy in

estimation of treatment effects and impacts of moderators. Additionally, MLM allows determination of how well the model, at each level, adequately describes the phenomena captured in the data, thus providing a check of its precision of analysis. In contrast to the crude lumping and averaging, and lack of statistical accountability inherent in non-parametric SSED summary statistics, MLM could represent an elegant solution to data synthesis, should it stand up to tests of its validity.

To assess the validity of MLM meta-analyses, researchers must (a) apply the techniques to published data and inspect models' fit to the data, (b) explore properties of models across various data conditions with large, simulated samples of data (i.e., Monte Carlo methods), and (c) compare various MLM approaches with each other, and to other means of data synthesis. At present, 7 studies have applied MLM techniques to SSED data (Adams, 2009; Hurwitz, 2008; Miller, 2006; Morgan & Sideridis, 2006; Terrazas Arellanes, 2009; Wade, Ortiz, & Gorman, 2007; Wang, Cui, & Parrila, 2011). However, only Adams (2009) critically assessed models' fit to the data (i.e., all others did not mention considering alternative model structures or assessing multiple models for their quality of fit). Six studies have examined MLM's methodological properties with regard to single-subject data using Monte Carlo methods (Beretvas & Wang, 2011; Ferron et al., 2009; Ferron, Farmer, & Owens, 2010; Jenson et al., 2007; Van den Noortgate & Onghena, 2011). However, the studies have explored a limited number of models and data conditions. Much remains to be examined. No studies systematically compared MLM meta-analysis to other means of data synthesis.

Research Questions

This dissertation seeks to answer the following research questions:

When MLM is used to meta-analyze single-subject research,

1. What levels of power are achieved for statistical tests of fixed effects?

2. How accurate are estimates of fixed and random effects, and autocorrelation levels in terms of relative parameter bias across conditions examined?
3. What patterns of differences exist in convergence rates, power rates, and relative bias in estimates of fixed effects and random effects' variance components across (a) specifications for model errors at level 1, (b) numbers of data points per experimental phase, (c) numbers of participants per study, (d) numbers of studies meta-analyzed, (e) degrees of autocorrelation in individuals' data, and (f) continuity of level 1 variance across phases?

CHAPTER 2

Review of Literature on Use of Multilevel Modeling with Single-Subject Experimental Data

To ensure that this study builds upon and is situated in the context of prior research, a literature review was first conducted on use of MLM with SSED. A number of authors have explored and/or commented on methodological issues related to using MLM with SSED (Bell, Morgan, Zhu, & Schoeneberger, 2011, Beretvas, 2011; Beretvas & Wang, 2011; Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens, 2010; Jenson et al., 2007; Nugent, 1996; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008, 2011). These authors have delineated procedures for using MLM with SSED and determined various methodological properties of the procedures. This study extends the work of these authors by exploring research questions beyond the scope of their articles, but within the same line of reasoning/inquiry. A separate group of authors have employed MLM with SSED (Adams, 2009; Hurwitz, 2008; Miller, 2006; Morgan & Sideridis, 2006; Terrazas Arellanes, 2009; Wade, Ortiz, & Gorman, 2007; Wang, Cui, & Parrila, 2011). The work of these authors illustrates the pertinent methodological issues and provides a rationale for pursuit of related research. This study seeks to identify best practice (or at least better practice) with regard to the many (potential, yet probable) methodological flaws in these applied studies.

This chapter is organized into several sections. First, in the Methods section, the literature search and coding processes are explained. Next, in the Results section, the results of the literature search and coding processes are described. Results are offered separately for articles which comment on the use of MLM with SSED and those that apply MLM to SSED. Finally, in the Discussion section, a critique of the use of MLM with SSED is made and future research questions are identified.

Methods

Search procedures. Systematic searches were conducted in three electronic databases: PsycINFO, Psychology and Behavioral Sciences Collection, and Educational Resources Information Clearing House (ERIC). On all three databases, the Boolean term "([single case] or [single subject]) and ([multilevel] or [multi-level] or [hierarchical linear])" was typed in the search field without specification of a search domain (e.g., keywords, abstract, title). No additional restrictions were placed on search results (e.g., publication year, language, peer-reviewed). The abstracts of the resulting 28 articles were reviewed to identify studies for inclusion (see Selection Criteria below). Following this initial search, ancestry searches were performed to identify additional articles for possible inclusion. First, the electronic databases were searched for other papers by authors of selected articles. Abstracts of the resulting papers were reviewed for inclusion. Then, reference lists of all articles meeting the selection criteria were reviewed. Finally, for all relevant citations, abstracts were reviewed for inclusion.

Additionally, conference proceedings for the most recent meeting of the American Educational Research Association (i.e., April, 2011) were searched for relevant presentations. Four additional posters and papers were identified, the contents of which had yet to be published in journals. Reports were requested and obtained from study authors.

Selection criteria. To be included in the review, an article had to meet one of three criteria. Included articles (a) described procedures for applying MLM to SSED, (b) examined the methodological properties of MLM with regard to SSED, or (c) applied MLM to data from SSED.

Coding and summary of selected articles. Coding of the selected articles pertained to methodological assertions and procedures applied. Each article was summarized in terms of (a)

model specifications, (b) unit counts at each level, (c) data extraction, (d) data standardization method, (e) treatment of autocorrelation, and (f) analysis process.

Results

Methodological commentary. The literature search yielded 12 articles that comment on methodological issues related to using MLM with SSED (Bell, Morgan, Zhu, & Schoeneberger, 2011, Beretvas, 2011; Beretvas & Wang, 2011; Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens, 2010; Jenson et al., 2007; Nugent, 1996; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008, 2011). Six articles described procedures for applying MLM to SSED (Beretvas, 2011; Nugent, 1996; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). The other six studies examined the methodological properties of MLM with regard to SSED (Beretvas & Wang, 2011; Ferron et al., 2009; Ferron, Farmer, & Owens, 2010; Jenson et al., 2007; Van den Noortgate & Onghena, 2011). Table 1 presents summaries of the articles' methodological assertions. Below, the assertions are described.

Model specifications. As mentioned above, multilevel models comprise several nested regression equations. The specifications of models are important because components of the equations determine which data phenomena can be summarized and how accurately and precisely the phenomena are represented. Authors of the reviewed articles suggested use of a variety of models. In all cases, authors noted that the choice of specifications should be based on patterns observed in the data and/or results of analyses (e.g., statistical tests of variance components). In the following sub-sections, the varieties of specifications suggested by authors are reviewed.

Mean change model. The model most often commented upon was the mean change model (Ferron et al., 2009; Ferron, Farmer, & Owens, 2010; Jenson et al., 2007; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). In this model, repeated measurements for subjects are organized and analyzed at level 1. Means are calculated for each phase and

contrasted. The resulting effect measure is the difference between phase means. Van den Noortgate & Onghena (2003a, 2008) clarified that the mean change model should only be used when linear or curvilinear trends are not present in data.

The authors suggested use of the following regression equation at level 1:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{phase})_{ijk} + e_{ijk} \quad (1)$$

where Y_{ijk} is the dependent score at time *i*, for subject *j*, from study *k*; π_{0jk} is a regression coefficient which represents the mean of the subject's baseline data; π_{1jk} is a regression coefficient which serves as the treatment effect measure for the subject; $(\text{phase})_{ijk}$ is a dummy variable that indicates whether the dependent measurement occurred (i.e., phase = 1) or did not occur (i.e., phase = 0) during the treatment phase; and e_{ijk} is a random effect that accounts for the deviation of measurement *i* from its expected value in the model. For a two level model, that excludes studies as a clustering variable, the subscript *k* is dropped from the regression equations' notation. Should a researcher wish to analyze data from more than two phases, additional regression coefficients and dummy variables can be added to the model (Van den Noortgate & Onghena, 2007). For example, a regression equation for the analysis of a baseline phase followed by two different treatments could take the form:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{phase2})_{ijk} + \pi_{2jk}(\text{phase3})_{ijk} + e_{ijk} \quad (2)$$

where π_{2jk} is a regression coefficient that represents the difference in means between treatment 1 (i.e., phase 2) and treatment 2 (i.e., phase 3), and symbols previously defined have the same meaning.

Linear growth model. Four articles described use of a linear growth model (Bell et al., 2011; Van den Noortgate & Onghena, 2003a, 2008, 2011). Additional articles alluded to their utility, but did not describe how to structure linear growth models (Ferron et al., 2009; Jenson et al., 2007; Van den Noortgate & Onghena, 2003b, 2007). In these models, as above, repeated

measurements for subjects are organized and analyzed at level 1. Intercepts and slopes of regression lines are estimated for each phase. The effect measures that result from the model describe the immediate effect (i.e., the difference between expected dependent scores at the end of baseline phases and the beginning of treatment phases) and the gradual effect (i.e., the difference between slopes of baseline and treatment phase regression lines).

The authors suggested use of the following regression equation at level 1:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{phase})_{ijk} + \pi_{2jk}(\text{time})_{ijk} + \pi_{3jk}(\text{phase})_{ijk}(\text{time in treatment})_{ijk} + e_{ijk} \quad (3)$$

where Y_{ijk}, $(\text{phase})_{ijk}$, and e_{ijk} have the same meanings as before; π_{0jk} is a regression coefficient that represents the intercept of the baseline regression line; π_{1jk} is a regression coefficient that serves as a measure of the immediate effect; π_{2jk} is a regression coefficient that represents the slope of the baseline regression line; $(\text{time})_{ijk}$ is a numeric variable that denotes the number of dependent measurements that have been taken at time *i* (i.e., if session number = 4, then time = 4); π_{3jk} is a regression coefficient that serves as a measure of the gradual effect; and $(\text{time in treatment})_{ijk}$ is a numeric variable, centered at the final baseline time point, that charts how many dependent measurements have been taken during the treatment phase (i.e., the total number of baseline sessions plus one, subtracted from the session number). As above, for a two level model, that excludes studies as a clustering variable, the subscript *k* is dropped from the regression equations' notation. The authors did not comment on how to structure linear growth models for the analysis of more than two phases. However, the models can easily be extended to include additional regression coefficients, dummy variables, and time variables for additional phases, as in equation 2.

Polynomial growth models. One article described use of polynomial growth models (Nugent, 1996). Additional articles briefly mentioned their utility, but did not explicitly state how to structure such models (Jenson et al., 2007; Van den Noortgate & Onghena, 2003b). In these

models, as above, repeated measurements for subjects are organized and analyzed at level 1. Polynomial growth models extend linear growth models by adding curvilinear elements, such as squared or cubed terms. As such, polynomial growth models allow the fitting of curved regression lines to subjects' data that have one or more prominent bends. The effect measures that result describe the immediate effect (i.e., the difference between expected dependent scores at the end of baseline phases and the beginning of treatment phases) and components of the gradual effect (e.g., the instantaneous linear slope, acceleration).

Nugent (1996) suggested use of regression equations of the following format at level 1:

$$Y_{ij} = \pi_{0j} + \pi_{1j}(\text{time})_{ij} + \pi_{2j}(\text{time})_{ij}^{2} + \ldots + \pi_{pj}(\text{time})_{ij}^{p} + e_{ij} \quad (4)$$

where Y_{ij}, $(\text{time})_{ij}$, and e_{ij} have the same meanings as before; π_{0j} is a regression coefficient that represents the intercept of the treatment phase regression line; π_{1j} is a regression coefficient that serves as a first measure of the gradual effect (i.e., the instantaneous linear slope); π_{2j} is a regression coefficient that serves as a second measure of the gradual effect (i.e., the acceleration); π_{pj} is a generic form of a regression coefficient that could serve as an additional measure of the gradual effect (i.e., the rate at which dependent scores increase when time, raised to the power of *p*, increases by 1). Nugent (1996) recommended assessing models' fit to data before committing to the use of a particular level 1 regression equation. He suggested visually inspecting subjects' data for apparent patterns, estimation of several models that seem to be appropriate for the data, and testing of obtained values (e.g., autocorrelation, fixed effects).

The approach suggested by Nugent (1996) only involves analysis of treatment phase data. To include baseline data in a polynomial model, a regression equation, such as the following, could be used:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{phase})_{ijk} + \pi_{2jk}(\text{time})_{ijk} + \pi_{3jk}(\text{phase})_{ijk}(\text{time in treatment})_{ijk} + \pi_{4jk}(\text{phase})_{ijk}(\text{time in treatment})_{ijk}^{2} + e_{ijk} \quad (5)$$

SMD-based models. One article described use of SMD-based models (Van den Noortgate & Onghena, 2003a). The authors suggested calculating SMDs from subjects' repeated measurements using the formula presented by Cohen (1969). The effect measures equal the difference between treatment and baseline means, divided by the standard deviation of baseline data. Van den Noortgate and Onghena (2003a) state the statistics should then be incorporated in multilevel models as dependent outcomes in level 1 regression equations.

The authors suggested use of the following regression equation at level 1:

$$d_j = \delta_j + e_j \quad (6)$$

where d_j is the SMD value for subject *j*; δ_j is the "true" SMD value for subject *j* (i.e., an estimate of an unbiased statistic); and e_j is the deviation of d_j from δ_j. Should a researcher wish to include data from more than two experimental phases, the additional information can be (a) included in SMD calculations via aggregation of data from like-phases or (b) used to calculate additional SMD statistics, which would then be combined within-participants in an additional level of analysis. Van den Noortgate & Onghena (2003a) stated this model should only be used when linear and curvilinear trends are not present in data.

OLS regression coefficient-based models. Three articles described use of ordinary least squares (OLS) regression coefficient-based models (Van den Noortgate & Onghena, 2003a, 2008, 2011). An additional article briefly mentioned their utility, but did not explicitly state how to structure such models (Van den Noortgate & Onghena, 2003b). The authors suggest first analyzing subjects' repeated measurements using OLS regression techniques. Next, the resulting regression coefficients should be standardized (see ***Standardization of data on different metrics*** below). The new values should then be used as dependent outcomes in level 1 regression equations. The resulting effect measures are the same as those of the mean change, linear, and polynomial models, depending on the format of the OLS regression equation.

The authors suggested use of the following regression equations at level 1, when using equation 3 for the OLS regression analysis:

$$\hat{\pi}_{1j} = \pi_{1j} + e_{1j} \quad (7)$$

$$\hat{\pi}_{3j} = \pi_{3j} + e_{3j} \quad (8)$$

where $\hat{\pi}_{1j}$ and $\hat{\pi}_{3j}$ are OLS regression coefficients representing level and slope change; π_{1j} and π_{3j} are the "true" coefficient values for subject *j*; and e_{1j} and e_{3j} are the deviation of $\hat{\pi}_{1j}$ and $\hat{\pi}_{3j}$ from π_{1j} and π_{3j}. Should a researcher wish to include data from more than two experimental phases, the additional information can be included in OLS regression analyses via additional regression coefficients and variables. Comparable phases' regression coefficients could then be combined within-participants in an additional level of analysis, or additional equations would be employed at levels 1 and above to separately synthesize coefficients across participants.

Nonlinear logistic models. One article described use of a nonlinear logistic model (Beretvas, 2011). The author suggested use of a nonlinear logistic models when data appears to form asymptotes or is subject to floor or ceiling levels (e.g., at y = 0 or y = 100). The following two models allow the fitting of S-like curves to data that form asymptotes at user defined levels. Beretvas (2011) suggests use of the following model when baseline data does not contain trends:

$$Y_{ij} = (1 - \text{phase})_{ij}(\pi_{0j}) + (\text{phase})_{ij}\left(\alpha_1 + \frac{\alpha_2 - \alpha_1}{\{1+(\pi_{1j})\exp[-(\pi_{2j})(\text{time in treatment}_{ij})]\}} + (\pi_{0j})\right) + e_{ij} \quad (9)$$

where Y_{ij} is the outcome measure at time *i* for subject *j*, "phase" is a dummy variable indicating a measurement took place during the treatment phase, α_1 and α_2 are the lower and upper asymptotes, respectively, π_{0j} represents the baseline intercept, and π_{1j} and π_{2j} determine the horizontal position and rate of vertical rise of the logistic function, respectively. When baseline data is linearly trended, Beretvas (2011) suggests use of the following model:

$$Y_{ij} = (1 - \text{phase})_{ij}[\pi_{0j} + \pi_{1j}(\text{time})_{ij}] +$$

$$(\text{phase})_{ij}\left(\alpha_1 + \frac{\alpha_2 - \alpha_1}{\{1+(\pi_{2j})\exp[-(\pi_{3j})(\text{time in treatment}_{ij})]\}} + \left[\pi_{0j} + \pi_{1j}(\text{time in treatment})_{ij}\right]\right) + e_{ij} \quad (10)$$

where symbols defined immediately above have the same meaning, except π_{1j} now represents the slope fit to baseline data, and π_{2j} and π_{3j} determine the horizontal position and rate of vertical rise of the logistic function, respectively.

Models employing log-link functions. One article described use of a model employing a log-link function (Beretvas & Chung, 2011). Due to the logarithmic character of models employing log-link functions, these models match single-subject data well due to the impossibility of prediction of negative values. Beretvas & Chung (2011) suggest use of the following model, in place of the previously described linear model:

$$\text{Log}\,(Y_{ij}) = \pi_{0j} + \pi_{1j}(\text{time})_{ij} + \pi_{2j}(\text{phase})_{ij} + \pi_{3j}(\text{phase})_{ij}(\text{time in treatment})_{ij} \quad (11)$$

where symbols defined for the linear model have the same meaning, except coefficients are interpreted in terms of log(Y), instead of simply Y.

When data are collected on multiple dependent variables for each subject, the authors recommend the following expansion of the previous model:

$$\text{Log}\,(Y_{ij}) = \pi_{0j} + \pi_{1j}(\text{time})_{ij} + \pi_{2j}(\text{phase})_{ij} + \pi_{3j}(\text{phase})_{ij}(\text{time in treatment})_{ij} +$$
$$\pi_{4j}(\text{setting2})_j + \pi_{5j}(\text{setting3})_j + \pi_{6j}(\text{setting2})_j(\text{phase})_{ij} +$$
$$\pi_{7j}(\text{setting3})_j(\text{phase})_{ij} + \pi_{8j}(\text{setting2})_j(\text{time in treatment})_{ij} +$$
$$\pi_{9j}(\text{setting3})_j(\text{time in treatment})_{ij} \quad (12)$$

where "setting2" is a dummy variable indicating a measurement took place in setting 2, as opposed to setting 1; "setting3" is also a dummy variable indicating a measurement took place in setting 3; π_{4j} and π_{5j} represent the difference between baseline intercepts in setting 1 and settings 2 and 3, respectively; π_{6j} and π_{7j} represent the difference between immediate treatment effects in

setting 1 and settings 2 and 3, respectively; and π_{8j} and π_{9j} represent the difference between gradual treatment effects in setting 1 and settings 2 and 3, respectively.

Multiple error terms at level 1. One article described use of multiple error terms at level 1 (Van den Noortgate & Onghena, 2003b). An additional article briefly mentioned the feasibility of the practice, but did not go into detail of how to include the multiple terms (Van den Noortgate & Onghena, 2007). The authors stated error variance may not be the same across phases (i.e., heteroscedastic variance). Should this be true, inclusion of a single error term can introduce bias into estimates of level 1 variance and other parameters. To avoid bias, Van den Noortgate & Onghena (2003b) recommend including separate error terms for each phase.

The authors suggested use of regression equations of the following format at level 1:

$$Y_{ij} = \pi_{0j} + \pi_{1j}(phase)_{ij} + e_{1ij}(phase1)_{ij} + e_{2ij}(phase2)_{ij} \quad (13)$$

where Y_{ij}, π_{0j}, π_{1j}, and $(phase)_{ij}$ have the same meanings as symbols defined for the mean change model; e_{1ij} is the error term for baseline measurements; $(phase1)_{ij}$ is a dummy variable indicating a measurement did (i.e., phase1 = 1) or did not (i.e., phase1 = 0) take place during the baseline phase; e_{2ij} is the error term for treatment phase measurements; and $(phase2)_{ij}$ is a dummy variable indicating a measurement did (i.e., phase2 = 1) or did not (i.e., phase2 = 0) take place during the treatment phase. Should a researcher wish to analyze data from more than two experimental phases, additional error terms and dummy variables can be added to level 1 equations.

Higher levels of models. In higher levels of models, results of level 1 analyses are synthesized. All authors were in agreement on how to structure regression equations at higher levels. They suggested using coefficients from lower levels (e.g., π_{0j}, π_{1j}) as the dependent outcomes of higher level equations. Specifically, the authors recommended use of equations of the following formats at levels 2 and 3, respectively:

$$\pi_{pjk} = \beta_{pk} + r_{pjk} \quad (14)$$

$$\beta_{pk} = \gamma_{p0} + u_{pk} \quad (15)$$

where π_{pjk} is the p^{th} regression coefficient from the level 1 equation for subject *j* from study *k*; β_{pk} is the average of π_p parameters from study *k*; r_{pjk} is the deviation of π_{pjk} from β_{pk}; γ_{p0} is the overall average of β_p parameters; and u_{pk} is the deviation of β_{pk} from γ_{p0}. If the model only comprises two levels, the *k* subscripts are excluded and β_p becomes the overall average of π_p parameters.

Inclusion of predictor variables. To assess the role of potential mediators of effect, predictor variables can be added to models. All authors agree on how to include predictor variables in models. They suggested including variables in level 2 and 3 equations according to the following format:

$$\pi_{pjk} = \beta_{p0jk} + \beta_{p1jk}(X_1)_{jk} + \ldots + \beta_{pqjk}(X_q)_{jk} + r_{pjk} \quad (16)$$

$$\beta_{pqk} = \gamma_{p0} + \gamma_{p1}(Z_1)_k + \ldots + \gamma_{pq}(Z_q)_k + u_{pqk} \quad (17)$$

where symbols defined immediately above have the same meanings; the subscript *q* is a label that differentiates parameters and variables at a given level; X_1 through X_q are person-level variables; and Z_1 through Z_q are study-level variables. As before, if the model comprises only two levels, the *k* subscripts are excluded.

Unit counts. A second issue related to the use of MLM with SSED regards the number of units analyzed at each level of a model. In the models suggested by the authors, and described above, level 1 units are dependent variable measurements or measures of effect, level 2 units are subjects, and level 3 units are studies. The sample sizes of these units have a direct relationship with the precision and reliability of parameter estimates (Raudenbush & Bryk, 2002). Larger samples of measurements and subjects are associated with greater precision and reliability of parameter estimates.

A variety of variables influence what constitutes an optimal sample size (Spybrook, Raudenbush, Congdon, & Martinez, 2009). These include the magnitude of effects, intra-class correlation, proportion of variance explained by level 2 predictor variables, and the power of analysis procedures. Optimal sample sizes can be determined using mathematical formulae (Raudenbush, 1997; Raudenbush & Liu, 2000).

Six articles commented on unit counts (Bell et al., 2011; Beretvas & Wang, 2011; Ferron, Farmer, & Owens, 2010; Jenson et al., 2007; Van den Noortgate & Onghena, 2003b, 2007). Van den Noortgate and Onghena (2003b) asserted that 30 or more units should be included at each level to obtain precise parameter estimates. The authors added that the number of measurements within subjects may be very small, as long as the analysis includes an adequate number of subjects. In their 2007 paper, Van den Noortgate and Onghena stated that "at least about 20" units at level 2 should be included, or more if predictor variables are added at level 2. In both papers, the authors do not provide justification or references for their claims.

Jenson et al. (2007) inspected the rates of type I error and power for various sample sizes at level 1 and 2 in a Monte Carlo simulation study. Using a computer program, the researchers simulated 1,000 samples of data sets for 90 conditions. Each condition was a unique combination of (a) numbers subjects (i.e., 15, 40, or 80), (b) numbers of baseline and treatment data points (i.e., 5/10 or 10/20), (c) levels of autocorrelation (i.e., high, low, none), and (e) effect size (1 standard deviation, 0.5 standard deviations, or no effect). Data was not simulated to contain trends. Samples of data sets were then analyzed using the mean change model described above. Jenson and colleagues produced results that show small numbers of level 2 units (i.e., $n = 15$ and 40) can interact with high autocorrelation to decrease power to undesirable levels (i.e., $.20 \leq \beta \leq .73$). The researchers also found that small samples of data points (i.e., $n = 15$ data points) are associated with low power (i.e., $.20 \leq \beta \leq .76$), except when subject samples are large (i.e., $n =$

80; β = .82). It should be noted that the results of Jenson and colleagues' study do not necessarily generalize to meta-analyses of data with trends or those that employ models other than the mean change variety.

Other simulation studies showed similar relationships between MLM performance and unit counts (Bell et al., 2011; Beretvas & Wang, 2011; Ferron, Farmer, & Owens, 2010). Generally, as the numbers of data points per phase and subjects per study increased, MLM performance improved. Bell et al. (2011) provided data that showed large samples of subjects (i.e., ≥ 16) were necessary to achieve adequate power for gradual treatment effects in a linear model when effects were small (i.e., 0.5 *SD*). However, when gradual treatment effects were large (i.e., ≥ 1.0 *SD*), smaller samples of subjects (i.e., ≥ 8) were also associated with adequate power. Bell et al. (2011) also showed that adequate power for tests of immediate effects in the linear model could only be achieved when effects were very large (i.e., 1.75 *SD*; increases in sample sizes did not improve power). Beretvas & Wang's (2011) obtained conflicting findings regarding subject sample sizes, although their results should be received with caution. They found increases in numbers of subjects were not associated with reductions in relative bias in estimates of fixed effects and random effects' variance components. The researchers only generated samples with 3 and 5 subjects and thus may have obtained their conflicting results due to the restriction of the range of numbers of subjects. Beretvas & Wang (2011) also found that relative biases of estimates of fixed effects, but not variance components, were less when measure sample sizes numbered 30 than when they numbered 10. Ferron, Farmer, & Owens (2010) found that increases in numbers of data points were associated with reductions in bias of fixed effect estimates. Their results suggest that when OLS procedures are used to estimate fixed effects and 30 data points were collected for each subject, fixed effects are accurate.

Data extraction. When conducting a meta-analysis, researchers must decide which data from primary research to include. SSED designs often comprise multiple phases in which the experimental conditions vary. Typically, data collection begins with baseline measurements and is followed by one or more treatment phases, and possibly additional baseline phases (Kennedy, 2005). At a minimum, a meta-analysis must draw on data from one baseline and one treatment phase (i.e., AB pair) for each subject. However, data from all phases can be incorporated in multilevel models.

Each author or author group gave different suggestions regarding data extraction. Nugent (1996) implied in his example analyses that data should be extracted only from treatment phases and aggregated prior to analysis. However, Nugent did not explicitly recommend this practice or offer guidelines for how to combine data from multiple treatment phases. Jenson et al. (2007) simulated data sets with only one baseline and one treatment phase in their Monte Carlo study. The authors commented that they viewed inclusion of single and/or multiple AB pairs in meta-analyses as viable practices. Van den Noortgate and Onghena (2003a, 2003b, 2007, 2008) offer several recommendations for extraction. They state that data can be extracted from single AB pairs or longer strings of phases. When data from like-phases is similar in terms of level, trend, and variability, the authors recommend aggregating the data prior to modeling. Alternatively, when data from like-phases is not similar, Van den Noortgate and Onghena recommend modeling each phase separately. To illustrate, the authors offered the following example in their 2003a paper of a level 1 model for an ABAB design:

$$Y_{ij} = \pi_{0j}(\text{pair 1})_{ij} + \pi_{1j}(\text{phase})_{ij}(\text{pair 1})_{ij} + \pi_{2j}(\text{pair 2})_{ij} + \pi_{3j}(\text{phase})_{ij}(\text{pair 2})_{ij} + e_{ij} \quad (18)$$

where Y_{ij} and e_{ij} have the same meanings as before, π_{0j} and π_{2j} are regression coefficients which represent the means from the first and second baseline phases, respectively; π_{1j} and π_{3j} are

regression coefficients which represents the differences between treatment and baseline phase means for the first and second AB pairs, respectively.

Standardization of data on different metrics. Synthesis of SSED study results often involves data on different metrics. SSED typically make use of one of a variety of methods for assessing dependent variables (Kennedy, 2005). Such methods include, but are not limited to, frequency counts, partial interval recording with continuous intervals, and partial interval recording with interrupted intervals. Further variation in partial interval recording systems pertains to lengths of intervals (e.g., 5 seconds, 10 seconds, 15 seconds). Each method for assessing dependent variables results in data on different metrics that is not directly comparable. Additionally, time scales used for measuring independent variables may vary across studies. While most researchers use session number as the time scale, some use other scales (e.g., time of day). In order to synthesize data on different metrics, meta-analysts must first standardize the data.

Three articles suggested procedures for standardizing data (Van den Noortgate & Onghena, 2003a, 2003b, 2008). In the articles, the authors recommended several methods.

Standardization of time scales. The first method involves standardization of time variables (Van den Noortgate & Onghena, 2003a). The authors suggested converting time variables to the same scale. Such conversions consist of simple algebraic transformations.

Reliance on OLS regression coefficients. The second method addresses diversity in dependent variable metrics. Van den Noortgate and Onghena (2003a, 2003b) recommend using standardized OLS regression coefficients as the unit of analysis at level 1. According to their prescriptions, subjects' data sets are first analyzed using OLS regression procedures. The resulting coefficients are then divided by the root mean square error (RMSE) of the regression model. Finally, the corresponding covariance matrices are made comparable with division by the

mean square error (MSE). Alternatively, in their 2011 work, Van den Noortgate & Onghena suggest inputting OLS regression coefficients as level 1 data without first dividing coefficients and covariance matrices by the RMSE and MSE. Their 2011 work provides data that shows this method is associated with unbiased estimates of fixed effects and accurate confidence intervals.

Division of scores by RMSE. Van den Noortgate & Onghena (2008) recommend a mathematically equivalent process to the use of standardized coefficients described above. They assert that division of all dependent scores by the RMSE of an OLS regression analysis, followed by synthesis of the standardized data with MLM, gives the same synthesis outcomes as above. However, in their 2011 work, the authors showed this method is associated with biased estimates of fixed effects and inaccurate confidence intervals.

Approximations of z-scores. A final method, suggested by Van den Noortgate & Onghena (2003b), involves an adaptation of the formula for z-scores. The authors suggest subtracting the baseline mean from all dependent scores, and dividing the difference by the within-phase standard deviation.

Treatment of autocorrelation. Repeated measurements, and their residuals in regression models, are autocorrelated, or serially dependent, if the value of one figure depends on the value of one or more of the immediately preceding figures (Verbeke & Molenberghs, 1997). Autocorrelation represents a problem because MLM procedures are based on the assumptions that residuals of the model are uncorrelated (Raudenbush & Bryk, 2002). Violations of this independence assumption can lead to biased estimates of standard errors and variances, and thus biased tests of parameters (*ibid*). In group design research, the assumption of independence is often met. For example, in a study of a pain medication, the pain ratings given by subject A likely do not depend on the pain ratings given by subject B. Should the pain ratings depend on each other somehow (e.g., via common influence of a variable or the communication of expectations

from one subject to another) researchers can implement experimental controls to prevent systematic dependence. In single-subject research, measurements are theoretically serially dependent (Borckardt, Nash, Murphy, Moore, Shaw, & Neil, 2008). For example, an individual's pain rating at three o'clock is dependent on his pain rating from two o'clock, if for no other reason because it proceeds from the two o'clock experience of pain. Similarly, repeated measurements of the weather and stock market are dependent upon previous measurements due to their continuity and gradualism (*ibid*). Additional factors that can contribute to dependence and autocorrelation of measurements include cyclical and random context variables that influence successive observations (Van den Noortgate & Onghena, 2003b).

Assessment of the problem of autocorrelation. Although SSED data is theoretically autocorrelated, observed levels of autocorrelation may not result in biased estimates and statistical tests. In the reviewed articles, the authors mention three strategies for determining when autocorrelation is present or represents a problem.

As described above, Jenson and colleagues (2007) performed a simulation study which assessed the relationship between autocorrelation and type I error and power. Simulation of an extremely large number of data sets allowed the authors to empirically identify conditions in which various levels of autocorrelation in data may induce substantial bias and confound statistical tests. When researchers wish to assess the impact of autocorrelation on the modeling of data with specific characteristics, simulation studies are viable strategy.

Van den Noortgate and Onghena (2003a, 2003b, 2007, 2008) describe a different approach useful for applied meta-analysts. When preparing an analysis of raw data, researchers can request for software packages to model a first-order autocorrelation of residuals within subjects. The software then uses all data to estimate an autocorrelation coefficient. The coefficient can be tested to determine if the level of autocorrelation is statistically significant. In

contrast to the method employed by Jenson et al. (2007), identification of a significant level of autocorrelation does not provide information regarding if and how estimates of standard error and variance, as well as statistical tests, are substantially biased.

Nugent (1996) and Van den Noortgate and Onghena (2003a) comment on a third approach. The authors assert that individual data sets can be tested for significant levels of autocorrelation of successive residuals resulting from a regression model using Durbin-Watson statistics. Similar to above, identification of significant levels of autocorrelation with Durbin-Watson statistics does not provide information on if and how MLM results may be biased.

Solutions for problematic levels of autocorrelation. Nugent (1996) and Van den Noortgate & Onghena (2003a, 2003b, 2007) recommended solutions for problematic levels of autocorrelation. Nugent (1996) suggested use of a more complex model (e.g., quadratic model, as opposed to mean change) when residuals are found to be autocorrelated. He notes that misspecification of models are consistently associated with autocorrelated residuals (Greene, 1990; Harvey, 1990). Van den Noortgate and Onghena (2003a, 2003b, 2007) recommend specification of an autoregressive covariance structure, as opposed to the default, simple covariance structure, when significant autocorrelation is identified. They claim autoregressive covariance structures attenuate or eliminate bias induced by autocorrelation. Ferron et al. (2009) found that when errors at level 1 are autocorrelated, specification of an autoregressive covariance structure and use of the Kenward-Rogers method for approximating degrees of freedom is associated with accurate estimation of confidence intervals. However, Ferron et al. (2009) also found that the practice did not reduce relative bias levels below the threshold of acceptability.

Analysis process. When performing a multilevel meta-analysis, researchers should proceed systematically through a series of steps (Raudenbush & Bryk, 2002). Generally, the process involves (a) preliminary analyses that assess how well a data sample meets the

assumptions of MLM, (b) recursive estimation of models and statistical testing of parameters that allow gradual determination of appropriate model specifications, and (c) estimation and testing of a final model. Omission of steps in the process risks model misspecification and/or biased results (*ibid*). Collectively, the reviewed articles addressed all aspects of the systematic analysis process. However, the guidelines offered by authors were limited in detail and no single article addressed all aspects of the process.

Preliminary analyses and the checking of assumptions. Four assumptions of MLM should be checked prior to meta-analysis of SSED data. The modeling procedures are based on assumptions that (a) outcomes are linear functions of the regression coefficients, (b) residuals are normally distributed, (c) residuals are independent, and (d) variance is homoscedastic (Raudenbush & Bryk, 2002). Also, analysts should examine the frequency distribution of each variable with attention to the shape, scale, existence of outliers, and possible needs for variable transformations (*ibid*).

Three articles addressed preliminary analyses (Nugent, 1996; Van den Noortgate & Onghena, 2003a, 2003b). Nugent (1996) noted the assumptions of MLM, but did not discuss how or why to check assumptions. Van den Noortgate and Onghena (2003a, 2003b) addressed the importance of checking assumptions and suggested procedures for doing so. With regard to the assumption of normally distributed residuals, the authors recommended making normal probability plots of residuals and inspecting for outliers (2003b). To assess the independence of residuals, the authors recommended modeling a first-order autocorrelation within subjects and testing the resulting autocorrelation parameter for significance (2003a, 2003b). For the assumption of homoscedasticity, the authors recommended estimation of separate error terms for each experimental phase and comparing outcomes (2003b). Van den Noortgate and Onghena, and

Nugent did not comment on the assumption that outcomes are a linear function of regression coefficients, nor the need to examine the distributions of each variable.

Recursive model estimation. All articles reviewed commented to some degree on the process of building an appropriate model. Authors of each article stated that unconditional models, which do not contain any predictor variables, should be estimated first. Nugent (1996) additionally stated that unconditional models should be evaluated for their goodness of fit to the data. He suggested that researchers visually analyze the quality of models' representation of patterns in individual data sets. Should the model appear inappropriate, Nugent (1996) recommended alteration of the model (e.g., inclusion of linear or curvilinear parameters), re-estimation, and visual analysis of the new model's goodness of fit to the data.

Each article similarly stated that predictor variables can be added to level 2 equations following estimation of an unconditional model. Van den Noortgate & Onghena (2003a, 2003b, 2007) and Jenson et al. (2007) noted that predictor variables should be included when, and only when, significant variance in parameters exists across cases.

As described above, Van den Noortgate and Onghena (2003a, 2003b, 2007) discussed inclusion of multiple error terms and specification of an autoregressive covariance structure. The authors recommended that researchers explore the need for and impact of including these elements during the recursive process of model building.

Additionally, Van den Noortgate and Onghena (2003a, 2003b, 2007) state that when more than one dependent variable is analyzed, multivariate MLM methods should be employed.

Analysis with a final model. All articles reviewed commented on the estimation and testing of a final model. Each author or author group suggested that estimation of a final model should include statistical tests of the fixed and random effects at the highest level of the model. Van den Noortgate and Onghena (2003b) mentioned an alternative to the typical χ^2 test for the

variance of random effects. They recommended testing the difference between fit statistics for models with and without random effects.

In addition to analysis with statistical tests, Nugent (1996) and Van den Noortgate (2007) recommend visual analysis of individual data sets. The authors mention visual analysis can qualify and supplement the findings of statistical tests.

Application of MLM to single-subject experimental data. A separate group of authors have conducted studies in which they applied MLM to single-subject experimental data. The literature search yielded 7 such studies (Adams, 2009; Hurwitz, 2008; Miller, 2006; Morgan & Sideridis, 2006; Terrazas Arellanes, 2009; Wade, Ortiz, & Gorman, 2007; Wang, Cui, & Parrila, 2011). Three studies were peer-reviewed journal articles (Morgan & Sideridis, 2006; Wade, Ortiz, & Gorman, 2007; Wang, Cui, & Parrila, 2011) and the remaining 4 were dissertations (Adams, 2009; Hurwitz, 2008; Miller, 2006; Terrazas Arellanes, 2009). Four of the studies were meta-analyses (Hurwitz, 2008; Miller, 2006; Morgan & Sideridis, 2006; Wang, Cui, & Parrila, 2011) and 3 were primary studies which aggregated and statistically tested results from individual participants using MLM (Adams, 2009; Terrazas Arellanes, 2009; Wade, Ortiz, & Gorman, 2007). Table 2 presents summaries of the articles' methods. Below, the methods are described.

Model specifications. The applied researchers made use of a variety of models at level 1 and higher.

Level 1 models. Generally, three types of models were specified at level 1: mean change models (Adams, 2009; Miller, 2006; Morgan & Sideridis, 2006; Wang, Cui, & Parrila, 2011), the SMD-based model (Hurwitz, 2008), and linear growth models (Terrazas Arellanes, 2009; Wade, Ortiz, & Gorman, 2007).

Researchers who used mean change models structured level 1 equations in several different ways. Miller (2006) and Wang, Cui, and Parrila (2011) employed simple mean change models. The researchers specified the following equation at level 1:

$$Y_{ij} = \pi_j(\text{phase})_{ij} + e_{ij} \quad (19)$$

where Y_{ij} and e_{ij} have similar meanings as before; $(\text{phase})_{ij}$ is a dummy variable indicating a measurement took place during a treatment phase (i.e., phase = 1) or baseline phase (i.e., phase = 0); and π_j is a regression coefficient which serves as the treatment effect measure (i.e., the difference between phase means). The model used by Wang, Cui, and Parrila (2011) consisted of 3 levels, and thus each term in the level 1 equation included a *k* subscript. Neither Miller (2006) nor Wang, Cui, and Parrila (2011) included an intercept in their level 1 models. The omissions were feasible due to their standardization methods, which set the mean of baseline measurements to zero (see *Standardization of data on different metrics* below).

The mean change model employed by Adams (2009) included a linear slope, continuous across all phases, and predictor variables. Adams (2009) specified the following equation at level 1:

$$Y_{ij} = \pi_{0j} + \pi_{1j}(\text{wear duration})_{ij} + \pi_{2j}(\text{day of week})_{ij} + \pi_{3j}(\text{time})_{ij} + \pi_{4j}(\text{phase})_{ij} + e_{ij} \quad (20)$$

where Y_{ij} and e_{ij} have similar meanings as before; π_{0j} is the overall mean of the baseline and withdrawal (i.e., second baseline) phases, when controlling for trend; π_{1j} is a regression coefficient that serves as a measure for the incremental effect associated with an increase of 1 in wear duration; $(\text{wear duration})_{ij}$ is the number of hours the independent variable was implemented on a given day, centered on the grand mean; π_{2j} is a regression coefficient that represents the difference between means of the dependent variable on weekdays and weekends during intervention, when wear duration equals the overall average and trend is controlled; $(\text{day of week})_{ij}$ is a dummy variable indicating the measurement took place on a weekday (i.e., day of

week = 1) or weekend (i.e., day of week = 0); π_{3j} is a regression coefficient that represents the linear slope, continuous across all phases; $(\text{time})_{ij}$ is the number of days lapsed since the beginning of data collection; π_{4j} is a regression coefficient which represents the difference between phase means, when wear duration equals the overall average, day of the week equals weekend, and trend is controlled; and $(\text{phase})_{ij}$ is a dummy variable indicating the measurement took place during baseline or withdrawal (i.e., phase = 0), or during treatment (i.e., phase = 1).

Morgan and Sideridis (2006) structured their model to depict the percent change between phase means. Their model additionally included a linear slope, continuous across phases. The researchers specified the following equation at level 1:

$$Y_{ij} = \pi_{0j} + \pi_{1j}(\text{time})_{ij} + \pi_{2j}(\text{baseline } \bar{Y})_{ij} + e_{ij} \quad (21)$$

where Y_{ij} and e_{ij} have similar meanings as before; π_{0j} is the model intercept, which represents the mean of baseline measurements when controlling for trend; π_{1j} is a regression coefficient that represents the continuous linear slope; $(\text{time})_{ij}$ is the session number; π_{2j} is a regression coefficient that serves as the treatment effect measure (i.e., the percent change between baseline and treatment phase means, after controlling for trend, expressed as a decimal); and $(\text{baseline } \bar{Y})_{ij}$ is the mean of baseline measurements.

The SMD-based model employed by Hurwitz (2008) seemed to be identical to that suggested by Van den Noortgate and Onghena (2003a). Hurwitz did not explicitly define the equations she used in her model. Based on descriptions of her methods and results, she appeared to use the following equation at level 1:

$$\overline{\text{SMD}}_j = \beta_k + r_j \quad (22)$$

where $\overline{\text{SMD}}_j$ is an effect size for subject *j*; β_k is the average effect size for consultant *k*; and r_j is the deviation of $\overline{\text{SMD}}_j$ from β_k.

The final model type specified at level 1 was the linear growth model. Terrazas Arellanes (2009) and Wade, Ortiz, and Gorman (2007) made use of models that accounted for differences between phases in means and slopes.

Terrazas Arellanes (2009) employed the following equation at level 1:

$$Y_{ij} = \pi_{0j} + \pi_{1j}(\text{time})_{ij} + \pi_{2j}(\text{phase})_{ij} + \pi_{3j}(\text{phase})_{ij}(\text{time in treatment})_{ij} + e_{ij} \quad (22)$$

where Y_{ij} and e_{ij} have similar meanings as before; π_{0j} is the model intercept which represents the expected value of the first baseline measurement; π_{1j} is a regression coefficient that represents the slope of a regression line for baseline data; $(\text{time})_{ij}$ is the session number; π_{2j} is a regression coefficient that represents the difference between the expected y-values for the final baseline measurement and the intercept of the treatment regression line, which has an x-value equal to the final baseline session number (i.e. the immediate effect, specified to take place at the end of the baseline phase); $(\text{phase})_{ij}$ is a dummy variable that indicates a measurement took place during the treatment phase (i.e., phase = 1) or baseline phase (i.e., phase = 0); π_{3j} is a regression coefficient that represents the difference between slopes of the baseline and treatment regression lines; $(\text{time in treatment})_{ij}$ is the session number minus the number of baseline sessions.

Wade, Ortiz, & Gorman (2007) employed the following equation at level 1:

$$Y_{ij} = \pi_{1j}(\text{baseline})_{ij} + \pi_{2j}(\text{treatment})_{ij} + \pi_{3j}(\text{follow-up})_{ij} + \pi_{4j}(\text{baseline time})_{ij} + \pi_{5j}(\text{treatment time})_{ij} + \pi_{6j}(\text{follow-up time})_{ij} + e_{ij} \quad (23)$$

where Y_{ij} and e_{ij} have similar meanings as before; π_{1j}, π_{2j}, and π_{3j} are regression coefficients that represent the mean of baseline, treatment, and follow-up phase measurements, respectively; $(\text{baseline})_{ij}$, $(\text{treatment})_{ij}$, and $(\text{follow-up})_{ij}$ are dummy variables that indicate a measurement took place during the phase (i.e., variable = 1) or not (i.e., variable = 0); π_{4j}, π_{5j}, and π_{6j} are regression coefficients that were meant to represent the slopes of regression lines for baseline, treatment, and follow-up phases, respectively (however, since the coefficients are not selectively "turned off" by

phase designating dummy variables, they do not serve as accurate estimates of within-phase slopes); and $(\text{baseline time})_{ij}$, $(\text{treatment time})_{ij}$, and $(\text{follow-up time})_{ij}$ are variables that indicate the number days that have passed since the phase began (i.e., day 1 = 0, day 2 = 1, etc).

Higher level models. Final models estimated by authors were both unconditional (i.e., did not include predictor variables at the highest level) and conditional (i.e., did include predictor variables at the highest level).

Adams (2009), Miller (2006), Terrazas Arellanes (2009), and Wade, Ortiz, and Gorman (2007) estimated final models without predictor variables at the highest level. Miller (2006) explored inclusion of a number of predictor variables at level 2, but found each to be insignificant. Adams (2009) tested the variance at level 2 and found that subjects' level 1 parameters did not vary significantly. Consequently, Adams decided not to add level 2 predictors to his model. Terrazas Arellanes (2009) and Wade, Ortiz, and Gorman (2007) did not report considering or testing predictor variables.

Hurwitz (2008), Morgan and Sideridis (2006), and Wang, Cui, and Parrila (2011) estimated final models with predictor variables at the highest level. Hurwitz (2008) stated that she tested parameters for a number of level 2 predictor variables and only retained those that were statistically significant. Morgan and Sideridis (2006) and Wang, Cui, and Parrila (2011) did not report estimating an unconditional model or testing the variance at level 2 in preparation of adding predictor variables. Further, the authors did not remove insignificant predictors from their final model.

Hurwitz (2008) added two predictor variables to her model's second level. Although she didn't define her level 2 equation, her methods and results suggested she used the following model:

$$\beta_k = \gamma_0 + \gamma_1(\text{study})_k + \gamma_2(\#\text{ of completed cases})_k + u_k \quad (24)$$

where β_k is the average SMD produced by consultant *k*; γ_0 is a regression coefficient that represents the hypothetical overall average SMD for subjects from study 1, whose consultant completed 0 cases; γ_1 is a regression coefficient that represents the hypothetical difference in effects associated with study 1 and 2, when $(\text{\# of completed cases})_k = 0$; $(\text{study2})_k$ is a dummy variable that indicates a consultant was a member of a study 1(i.e. study = 0) or study 2 (i.e. study = 1); γ_2 is a regression coefficient that represents the incremental effect associated with an increase of 1 in the number of cases completed by a consultant, when controlling for study; $(\text{\# of completed cases})_k$ is the number of cases completed by a consultant; and u_k is the deviation of expected values for β_k from the values estimated at level 1.

Morgan and Sideridis (2006) added 4 predictor variables to 2 of their model's second level equations. They specified the following 3 equations at level 2:

$$\pi_{0j} = \beta_{01}(\text{sex})_j + \beta_{02}(\text{age})_j + \beta_{03}(\text{placement})_j + \beta_{04}(\text{treatment1})_j + \beta_{05}(\text{treatment2})_j + \beta_{06}(\text{treatment3})_j + \beta_{07}(\text{treatment4})_j + \beta_{08}(\text{treatment5})_j + \beta_{09}(\text{treatment6})_j + \beta_{010}(\text{treatment7})_j + r_{0j} \quad (25)$$

$$\pi_{1j} = \beta_{11}(\text{sex})_j + \beta_{12}(\text{age})_j + \beta_{13}(\text{placement})_j + \beta_{14}(\text{treatment1})_j + \beta_{15}(\text{treatment2})_j + \beta_{16}(\text{treatment3})_j + \beta_{17}(\text{treatment4})_j + \beta_{18}(\text{treatment5})_j + \beta_{19}(\text{treatment6})_j + \beta_{110}(\text{treatment7})_j + r_{1j} \quad (26)$$

$$\pi_{2j} = \beta_{20} + r_{2j} \quad (27)$$

where π_{0j}, π_{1j}, and π_{2j} are level 1 parameters for subject *j*; β_{01} through β_{010} and β_{11} through β_{110} are regression coefficients that represent the effects associated with each variable, when controlling for other variables; $(\text{sex})_j$ is a dummy variable that indicates subject *j* is male (i.e., sex = 1) or female (i.e., sex = 0); $(\text{age})_j$ is a dummy variable that indicates the subject is in grades K – 4 (i.e., age = 0) or 5 – 12 (i.e., age = 1); $(\text{placement})_j$ is a dummy variable that indicates the subject receives special education services (i.e. placement = 0) or general education services (i.e., placement = 1); $(\text{treatment1})_j$ through $(\text{treatment7})_j$ are dummy variables that comprise a single

categorical variable and indicate a subject received a type of treatment (treatment# = 1) or did not (treatment# = 0); β_{20} is the average of π_2 parameters across subjects; and r_{0j}, r_{1j}, and r_{2j} are random effects that represent the deviation of expected values from the estimates of level 1 equations. Morgan and Sideridis (2006) stated they omitted intercepts from the level 2 equations in order to make the β parameters directly comparable.

Wang, Cui, and Parrila (2011) added 2 variables and an interaction term to their model's third level equation. They specified the following equation at level 3:

$$\beta_k = \gamma_0 + \gamma_1(\text{treatment})_k + \gamma_2(\text{age})_k + \gamma_3(\text{age*treatment})_k + u_k \quad (28)$$

where β_k is the average treatment effect measure for subject *k* (i.e., β_k is the outcome of level 2 equations, which synthesize the *j* multiple effect measures for a single participant); γ_0 is a regression coefficient that represents the hypothetical, overall average effect measure for subjects who received treatment 1 (i.e. treatment = 0), and had an age of 0; γ_1 is a regression coefficient that represents the difference in effects associated with treatment 1 and treatment 2, when controlling for age and the interaction of age and treatment type; $(\text{treatment})_k$ is a dummy variable that indicates a subject received treatment 1 or treatment 2 (i.e., treatment = 1); γ_2 is a regression coefficient that represents the incremental effect associated with an increase of 1 in age, when treatment equals 0; $(\text{age})_k$ is the age of subject *k* in years; γ_3 is a regression coefficient that was meant to represent the incremental effect associated with the interaction of age and treatment (however, it only represents the interaction effect when $(\text{treatment})_k = 1$; when $(\text{treatment})_k = 0$, the interaction term equals zero and γ_3 is "turned off"); $(\text{age*treatment})_k$ is the product of a subject's age and the dummy code for the treatment they received; and u_k is the deviation of expected values for β_k from those estimated at level 1.

Unit counts. Among the reviewed studies, only the meta-analyses (i.e., Hurwitz, 2008; Miller, 2006; Morgan & Sideridis, 2006; Wang, Cui, & Parrila, 2011) appeared to have adequate unit counts at each level. The studies involved between 43 and 202 units at the highest levels of models, and between 650 and 1796 units at level 1. The primary studies, in which MLM was used to synthesize and test results from multiple participants (i.e., Adams, 2009; Terrazas Arellanes, 2009; Wade, Ortiz, & Gorman, 2007), consistently involved inadequate numbers of units at level 2. The studies analyzed data for 5, 5, and 12 cases, respectively. Evaluation of the adequacy of unit counts is based on the rough estimates offered by Van den Noortgate and Onghena (i.e., 20+ or 30+ units at each level; 2003b, 2007).

Data extraction. Authors of the reviewed articles employed a variety of data extraction procedures. Morgan and Sideridis (2006) and Wang, Cui, and Parrila (2011) limited data collection to subjects' first baseline and treatment phase pairs only (i.e. AB pairs). All other authors extracted all data available. In the cases of Hurwitz (2008) and Terrazas Arellanes (2009), this involved extraction of one AB pair per subject. Adams (2009) collected data from all phases of an ABA design. For Wade, Ortiz, and Gorman (2007), extraction involved data from one baseline, treatment, and follow-up phase for each subject. Miller (2006) collected data from various numbers of phases per subject. She extracted data from AB pairs and withdrawal designs (i.e. ABA, ABAB). Authors who collected data from more than one baseline and/or treatment phase (i.e., Adams, 2009; Miller, 2006) aggregated the data for like-phases prior to analysis without first examining the consistency of data phenomena.

For several studies, authors extracted data on multiple dependent variables (Adams, 2009; Morgan & Sideridis, 2006; Terrazas Arellanes, 2009; Wade, Ortiz, & Gorman, 2007; Wang, Cui, and Parrila, 2011). Adams (2009), Terrazas Arellanes (2009), and Wade, Ortiz, and Gorman (2007) analyzed data for multiple dependent variables separately, in different multilevel

models. Morgan and Sideridis (2006) treated the multiple dependent variables as different cases at level 2, the highest level of their model. The authors made use of weights to equalize the influence of subjects who contributed multiple data sets with subjects who contributed 1 data set. In contrast, Wang, Cui, and Parrila (2011) combined effect measures for multiple dependent variables by subject in the second level of their 3 level model. Consequently, the authors did not use weights in their analysis.

Standardization of data on different metrics. Two studies standardized data on different metrics prior to analysis (Miller, 2006; Wang, Cui, & Parrila, 2011). Morgan and Sideridis (2006) meta-analyzed data on different metrics, but did not standardize the data. The remaining authors did not need to standardize data, because their data were on consistent metrics.

Miller (2006) standardized her data by transforming dependent measurements into z-scores. In contrast to the recommendations of Van den Noortgate and Onghena (2003b), Miller (2006) calculated z-scores by subtracting subjects' within-phase means from measurements and dividing the differences by the within-phase standard deviations. Subsequently, Miller (2006) subtracted subjects' mean baseline z-scores from all standardized scores.

Wang, Cui, and Parrila (2011) also standardized their data by transforming dependent measurements into z-scores. In contrast to both the recommendations of Van den Noortgate and Onghena (2003b) and the technique of Miller (2006), the researchers calculated z-scores by subtracting subjects' overall means from measurements and dividing the differences by overall standard deviations. Wang, Cui, and Parrila (2011) also subtracted the mean baseline z-score from all standardized scores.

Treatment of autocorrelation. One study involved treatment of autocorrelation (Adams, 2009). Adams (2009) assessed the level of autocorrelation by modeling a first-order autocorrelation of residuals within-subjects in an unconditional model. After finding a statistically

significant level of autocorrelation in his sample, Adams specified a heterogeneous autoregressive covariance structure for subsequent models. No other studies reported attending to autocorrelation issues.

Analysis process. Generally, authors of the reviewed articles did not proceed systematically through analyses according to the standards for MLM (Raudenbush & Bryk, 2002). Often, processes that were undertaken were conducted in a flawed or limited manner. In most papers, analysis choices were not explained or justified with rationale or references.

Preliminary analyses and the checking of assumptions. Only 1 study involved a preliminary analysis (Adams, 2009). As described above, Adams (2009) assessed the level of autocorrelation in his sample. No other preliminary analyses were performed by Adams (2009) or other authors.

Recursive model estimation. Authors of the reviewed articles based very few decisions of model specifications on data. As mentioned above, only one author referenced data in his consideration of level 1 models (Adams, 2009). Adams (2009) estimated several models, assessed their fit to the data with statistical and visual analyses, and selected the best fitting model. Authors who chose level 1 models that ignored trend (i.e., Hurwitz, 2008; Miller, 2006; Wang, Cui, and Parrila, 2011) did not report inspecting data for an absence of trends. With regard to level 2 models, only 3 authors based their selection on data (Adams, 2009; Hurwitz, 2008; Miller, 2006). As described above, Adams (2009) decided not to include predictor variables at level 2 after finding level 1 parameters did not significantly vary across subjects (i.e., level 1 variance components were not significant). Hurwitz (2008) and Miller (2006) tested the significance of predictor variables included at level 2. After finding some or all of the variables were insignificant, the authors omitted the insignificant predictors from their final models.

Analysis with a final model. Authors committed a number of errors in their final models and analyses. Ideally, the final model should only include significant parameters and the associated variables (Raudenbush & Bryk, 2002). Also, final analyses should involve statistical tests of both fixed and random effects (*ibid*). For applications of MLM in primary research, final analyses should additionally incorporate visual analyses (Nugent, 1996; Van den Noortgate & Onghena, 2007). Only the analyses by Adams (2009), Hurwitz (2008), and Miller (2006) met these ideals. Several authors omitted tests of random effects (i.e., Morgan & Sideridis, 2006; Terrazas Arellanes, 2009; Wang, Cui, & Parrila, 2011). Also, several authors failed to remove insignificant predictors from final models (Morgan & Sideridis, 2006; Wade, Ortiz, & Gorman, 2007; Wang, Cui, & Parrila, 2011). On a positive note, all applications of MLM in primary research did incorporate visual analyses along with statistical analyses (Adams, 2009; Terrazas Arellanes, 2009; Wade, Ortiz, & Gorman, 2007).

Discussion

This literature review summarizes the methodological content of 6 commentary, 6 experimental, and 7 applied articles. Review of the articles suggests the use of MLM with SSED is viable. However, much remains unknown about the methodological properties of MLM with regard to SSED. It appears certain practices and data characteristics can invalidate the use of MLM. Additional research is needed to clarify when the use of MLM with SSED is and is not appropriate. In the sub-sections below, a critique is offered on the use of MLM with SSED and the needs for future research are delineated.

Critique of Application of MLM with SSED. The following critique is organized in parallel with the sub-sections of the Results section. The methodological issues of using MLM with SSED are discussed with regard to (a) model specifications, (b) unit counts, (c) data

extraction, (d) standardization of data on different metrics, (e) treatment of autocorrelation, and (f) the analysis process.

Model specifications. The models used in analyses should represent data phenomena with accuracy and precision. Construction of models with good fit to data involves, at a minimum, (a) perusing graphs of data with an eye for general patterns, (b) formulation and estimation of one or more unconditional models that summarize observed patterns, and (c) comparison and evaluation of the fit of each model variety using statistical tests and visual analysis, prior to selection of a final model (Nugent , 1996; Raudenbush & Bryk, 2002; Van den Noortgate & Onghena, 2003a, 2003b, 2007). Achieving good model fit may also require inclusion of predictor variables at level 2 and/or higher (Raudenbush & Bryk, 2002).

Level 1 specifications. Level 1 models that do not fit data well fail to do so for a number of reasons. At times, models may omit necessary elements. For example, when trended data is analyzed with mean change and SMD-based models, the trends confound estimates of means and standard deviations. Compared to models that incorporate trend, the residuals are exaggerated, model fit is diminished, and estimates of fixed effects are less accurate (Van den Noortgate & Onghena, 2003a). Similarly, use of single error terms can confound estimation of variance components. Treatments regularly induce changes in the variance of measurements from baseline levels. When this occurs, and additional error terms are not included, variance components may be biased (Van den Noortgate & Onghena, 2003b). Perusing graphs of data with an eye for general patterns, as well as evaluation of models' fit with statistical tests and visual analyses can help analysts recognize omission of necessary elements.

Also, models may produce error as artifacts. For example, linear and polynomial growth models force trend components (e.g., parameters for linear and curvilinear slopes) to have constant values within phases. However, trends in single-subject data are at times discontinuous

and vary within phases. Implementation of a treatment may be first followed by a sudden behavior spike, then a gradual behavior reduction, and, eventually, the leveling off of behavior. When such phenomena are summarized with linear or polynomial growth models, the expected values of models can deviate greatly from actual observed data. Models may misrepresent some or all fluctuations in the data, and they may produce unrealistic values (e.g., expected final outcomes in the negative range or beyond ceilings, such as 150% of intervals). Again, perusing graphs of data with an eye for general patterns, as well as evaluation of models' fit with statistical tests and visual analyses can help analysts recognize artifactual error.

Incorrect formulation of level 1 equations can also lead to poor model fit. For example, Wade, Ortiz, and Gorman (2007) appeared to incorrectly specify the terms for within-phase slopes in their model (see Equation 23). In their model, the terms are not "turned off" by dummy variables (indicating measurements took place during a particular phase), and thus are formulated to be constant across all phases. Visual inspection of the graphs provided in their paper shows the slopes did vary across phases, and thus the model fit was likely diminished by the misspecification. Similarly, Terrazas Arellanes (2009) likely obtained relatively poorer model fit due to her choice of level 1 model formulation (see Equation 22). Terrazas Arellanes (2009) specified the term for change in treatment slope (i.e., $[\text{phase}]_{ij}$*$[\text{time in treatment}]_{ij}$) to include the difference between the session number and the number of baseline sessions (i.e., time in treatment = $n_t - n_b$). As a result, for the final baseline and first treatment data points, the term equals 0 and 1, respectively. Thus, the gradual effect begins at the end of baseline and first registers in the first treatment data point. Visual inspection of the graphs provided in her paper suggests that the formulation should have involved adding 1 to the number of baseline sessions (i.e., $n_t - [n_b + 1]$; Huitema & McKean, 2000). Doing so would have placed the intercept of the treatment regression lines at the first treatment data point, causing the gradual effect to first

register in the second treatment data point, and avoiding confounding of the immediate effect. In each of these cases, estimation and comparison of multiple models may have helped identify the possible errors in formulation.

Models may also fail to fit well due to high variance in data (Adams, 2009; Van den Noortgate & Onghena, 2003b). Taking many repeated measurements within-subjects can result in increased exposure to the effects of extraneous variables (e.g., history effects, setting events; Kennedy, 2005). The presence of extraneous variables is theoretically one source responsible for high variance in single-subject data.

Given the common presence in SSED data of between-phase level changes and within-phase linear and curvilinear trends, the following model has the potential to fit data well (Nugent, 1996; Van den Noortgate & Onghena, 2003a, 2008):

$$Y_{ij} = \pi_{0j} + \pi_{1j}T_{ij} + \pi_{2j}(\text{treatment})_{ij} + \pi_{3j}(T_{ij} - [n_t + 1])(\text{treatment})_{ij} + \pi_{4j}(T_{ij} - [n_t + 1])^2(\text{treatment})_{ij} + e_{ij} \quad (29)$$

where T_{ij} represents the session number at time i for subject j and n_t represents the total number of baseline data points (see Figure 1 and the text below for additional clarification). Use of this equation allows modeling of deceleration or acceleration of treatment phase behavior, as well as changes in direction of behavior trends during treatment phases (e.g., behavior spikes, followed by gradual declines of behavior). Both types of phenomena are common in single-subject data. Additionally, the model formulates the gradual linear and curvilinear effects of treatment as beginning at the start of the treatment phase (as opposed to the end of the baseline phase), which is consistent with field-wide expectations for the onset of treatment effects (Kennedy, 2005). Should a sample of data not be characterized by particular trends (i.e., linear or curvilinear) or changes between phases (i.e., slope change or immediate level change), estimates of the fixed effects will approximate zero. In such scenarios, inclusion of the extraneous parameters does not

compromise the model's fit to data. However, when tests of the fixed effects yield insignificant results, analysts may choose to remove the insignificant parameters. Unless visual analyses of data samples suggest otherwise, model building should begin with either an assessment of this model's fit to data, or a variety of this model that includes separate error terms for each phase.

Figure 1 visually illustrates the meaning of each parameter in Equation 29. The "actual data" (in blue, connected with a hatched line) was taken from Turner, et al. (1996). The data consist of 5 baseline and 5 treatment phase data points. The "polynomial model" (in green, with a solid line) was estimated using OLS regression procedures. Due to the inclusion of the dummy variables $(\text{treatment})_{ij}$, two distinct regression lines were estimated simultaneously during the regression analysis. When analyzing the baseline data, the dummy variables equaled 0 and effectively "turned off" the third, fourth, and fifth terms in the model. As a result the first two terms alone described baseline data, and the last three terms described the changes that take place in treatment phase data. In the model, π_{0j} serves as the intercept and represents the expected level of behavior during the first baseline session (i.e., the y-value when x=1). The parameter π_{1j} describes the linear trend in baseline data (i.e., the slope of the baseline regression line). The treatment effects are captured in π_{2j}, π_{3j}, and π_{4j}. The first of these, π_{2j} represents the difference between the expected level of behavior in the first treatment session and the expected level of behavior

had the baseline phase continued (i.e., the vertical distance between the intercept of the treatment regression line and the baseline regression line, when extended into the treatment phase). The parameters π_{3j} and π_{4j} describe the changes between trends in treatment phase and baseline data. The difference between the linear slope component of the treatment phase regression line and the slope of the baseline regression line is represented by π_{3j}. Due to the fact that π_{1j} remains "on" during the treatment phase and is not "turned off" by a dummy variable, the linear slope

component of the treatment phase regression line equals π_{1j} plus π_{3j}. The curvilinear slope of the regression line is represented by π_{4j}. Together, π_{1j}, π_{3j}, and π_{4j} shape the curvature of the treatment phase regression line. Finally, e_{ij} serves as the error term or residual of the model (i.e., the vertical distance between each "actual data" point and the model's expected values).

Incorporation of predictor variables. The incorporation of predictor variables can improve model fit, as well as parameter accuracy and precision. Although predictor variables may be included at any level of the model, typical practice involves including predictors at levels 2 and higher (Raudenbush & Bryk, 2002).

Traditionally, single-subject researchers make efforts to hold all variables constant for the duration of each phase (Kennedy, 2005). This exercise of experimental control is important for achieving internal validity in SSED. MLM and other statistical modeling procedures create an opportunity for a shift in practice. Instead of holding all variables constant, researchers could benefit from allowing certain variables to vary across time and measuring their levels during each session. After data collection is complete, researchers could then assess the relationships between the time-varying variables and dependent scores. Instead of relying on experimental controls, statistical control would grant internal validity to findings. The drawback to such an approach is the need for large samples of data points and subjects. As predictor variables and their parameters are added to models, successful estimation requires increasingly large samples.

Several practices may limit the benefit of including predictor variables. For one, modeling outcomes simply as the linear sum of effects of predictor variables can confound results. Often, variables interact. In such cases, the overall effect of multiple variables can be deconstructed into interaction effects and unique effects. To illustrate, Morgan and Sideridis (2006) may have confounded their analysis by not including parameters for interaction effects in their model. Their level 2 outcome, the percent change in phase means, may not be a linear

function of the unique effects of subjects' sex, age, educational placement, and treatment received. The variables could have interacted in such a way that each combination of conditions is uniquely associated with a particular level of percent change. If this was true, Morgan and Sideridis' (2006) model would have been improved by including several interaction terms. Unfortunately, the addition of interaction terms can quickly make estimation of models infeasible, due to sample size requirements.

Incorrect formulation of models can also limit the benefit of including predictor variables. For example, Wang, Cui, and Parrila (2011) may have incorrectly specified the interaction term in their conditional model (see Equation 28). The researchers sought to estimate the interaction between age and treatment type (i.e., a dummy variable). However, their formulation of the interaction involved multiplication of the dummy variable and numeric variable. Thus, it appears that when the treatment type equaled the reference category (i.e., treatment = 0), the interaction term equaled zero, and the term was effectually deleted from the model. Consequently, estimation of the effect of the interaction may have only drawn on data involving the comparison treatment (i.e., when treatment = 1 and the interaction term is retained in the model).

Additionally, not centering numeric variables can complicate interpretation of estimates of intercepts and parameters representing level change. parameter estimates. In their models, Hurwitz (2008) and Wang, Cui, and Parrila (2011) both left variables associated with level change parameters uncentered. As a result, the intercepts and level change parameters in their models have no practical meaning. As stated above, certain parameters represent hypothetical, average outcomes for subjects whose age is 0 or whose consultant had completed 0 cases. Centering the numeric variables on average values would have resulted in practically meaningful

estimates (i.e., the average outcome for subjects of the average age or whose consultant had completed the average number of cases).

Unit counts. While the recommended minimum unit counts of Van den Noortgate and Onghena (2003b, 2007) are reasonable rules of thumb, the adequacy of sample sizes depends on a variety of variables (Spybrook, Raudenbush, Congdon, & Martinez, 2009; Raudenbush, 1997; Raudenbush & Liu, 2000). The magnitude of effects, intra-class correlations, proportions of variance explained by level 2 predictor variables, and the power of analysis procedures interact to create sample size requirements unique to each analysis. To be confident that sample sizes are adequate, additional research will need to be done that clarifies the relationships between SSED data characteristics, sample sizes, precision and reliability of parameter estimates, and the validity of statistical tests.

Jenson et al. (2007), Bell et al. (2011), Ferron and colleagues (2009 and 2010), and Van den Noortgate & Onghena (2011) have begun the process of clarifying such relationships. However, the relationships have not been explored for many common conditions typically encountered in single subject research.

Unfortunately, research conducted on sample size requirements for the analysis of group design data cannot provide guidelines for the meta-analysis of SSED. Various common characteristics of single-subject data, such as very large effects, high variability, and presence of autocorrelation, are typically not shared by group design data. As a result, the power of MLM analyses of single-subject data is likely different than the power achieved with group design data. Also, the proportion of variance explained by level 2 predictor variables is likely smaller when analyzing SSED data (Adams, 2009; Van den Noortgate & Onghena, 2007).

In the applied studies reviewed, unit counts likely represented a problem for the primary studies. The numbers of subjects included in the analyses (i.e., 5, 5, and 12; Adams, 2009; Wade,

Ortiz, & Gorman, 2007; and Terrazas Arellanes, 2009, respectively) are typical for SSEDs (Horner et al., 2005). Should research confirm that these sample sizes are associated with poor precision, reliability, and validity, MLM should not be regularly used in the analysis of primary studies' data.

Data extraction. The choice of which data to extract for analysis should vary from sample to sample. Ideally, all data available for each subject would be extracted and incorporated in models (Van den Noorgate & Onghena, 2003a, 2007). However, conditions may make this undesirable or unfeasible. For example, a sample may contain diverse designs, such as multiple baseline designs made up of single AB pairs, withdrawal designs made up of multiple AB pairs, and alternating treatment designs in which AB pairs are followed by additional treatment phases (e.g., ABCD). In order to extract and make use of all data for subjects, the sample must be limited to a single design of consistent format (e.g., withdrawal designs with 2 AB pairs). Unfortunately, this option requires exclusion of data sets not meeting the design criteria. Alternatively, extraction could be limited to the first AB pairs of all designs (Morgan & Sideridis, 2006; Wang, Cui, & Parrila, 2011). With this option, wide samples of data may be collected. Although information from later phases is lost, the effect of interest is often captured in the first AB pair, with later phases serving to demonstrate experimental control. When making the choice of which data to extract, researchers should consider options that maximize the information included in analyses and allow them to most comprehensively answer their research questions.

Should a researcher extract data from multiple like-phases, the data from each phase should be modeled with separate parameters, unless analysis supports doing otherwise (Van den Noortgate & Onghena, 2003a). For example, if a researcher extracts data from ABAB designs, the level 1 model initially estimated should include distinct parameters for the first and second baseline phases, as well as the first and second treatment phases. Estimation of separate

parameters will allow researchers to observe if phenomena are consistent or vary across like-phases. Statistical tests can be then be performed to determine if phenomena differ significantly or if like-phase data may be aggregated. Assumption that phenomena are similar across like-phases and aggregation of data without supporting evidence represents poor practice (e.g., Adams, 2009; Miller, 2006).

Should a researcher extract data on multiple dependent variables, multivariate MLM procedures should be used in analyses (Van den Noortgate & Onghena, 2003a, 2003b, 2007). However, when dependent variables are theoretically similar, incorporation of multiple effect measures per subject in models is feasible via use of weights and/or additional levels in models (Morgan & Sideridis, 2006; Wang, Cui, & Parrila, 2011).

Standardization of data on different metrics. Little is known about the need for standardization of single-subject data on different metrics. The various methods of assessing dependent variables certainly do produce data on different metrics. However, the degree to which data on different metrics is different is not known. For example, data taken on a particular subject using partial interval recording with continuous intervals may or may not be significantly different than data taken during the same time using partial interval recording with interrupted intervals. Further, the difference between using 5 second intervals and 15 second intervals has not been empirically tested in large samples. At the moment, standardization represents best practice due to the likelihood that data on most metrics is not directly comparable (e.g., frequency counts and partial interval recording). However, research will need to be done to clarify which metrics are significantly different from each other and which metrics may be synthesized without standardization.

Similarly, little is known about the fidelity of transformations involved in standardization of data. For example, single-subject data often contain trends. These trends can be misrepresented

by the standardization methods recommended by Van den Noortgate and Onghena (2003a, 2003b, 2008). The use of OLS regression-based techniques, when an overly simplistic model is specified, as well as z-score transformations can vertically or horizontally compress the spread of data points, such that trends are exaggerated or attenuated. The z-score calculations used by Miller (2006) and Wang, Cui, and Parrila (2011) similarly have the potential to bias trended data. Wang, Cui, and Parrila (2011) used a formula that is especially inappropriate for SSED data. Their formula incorporated overall means and standard deviations for individuals' data sets. Since means and standard deviations presumably differ across phases, the use of overall means and standard deviations likely introduced error into their standardized scores.

Extreme data points pose a similar threat as trends to data standardization. Outliers are common in single-subject data (Allison & Gorman, 1994). These data points may confound the use of z-score standardization methods by skewing means and standard deviations, and inducing error in all z-scores. Outliers also threaten to confound the analysis of unstandardized data and data standardization using OLS regression-based techniques. However, outliers have the greatest potential to bias z-scores, due to the impact they have on all scores. In contrast, outliers may not affect all scores when regression-based standardization techniques are used or data is analyzed without standardization. Research will need to be done to clarify which methods are appropriate for single-subject data and what conditions are associated with poor fidelity of transformations.

Treatment of autocorrelation. The levels of residuals' autocorrelation should be assessed during all multilevel analyses of SSED data. Both assessment methods discussed by authors may provide useful information to analysts (Nugent, 1996; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). However, Durbin-Watson statistics are known to contain bias when estimated for small data sets (Huitema & McKean, 1998; Riviello & Beretvas, 2008). Modeling of first-order autocorrelations within subjects using MLM software allows analysts to estimate the

level of autocorrelation in the whole sample (Verbeke & Molenberghs, 1997). This procedure may produce less biased results. The present author's research has not uncovered findings regarding the performance of such estimations.

As stated above, significant levels of autocorrelation may not confound statistical tests. The threshold at which autocorrelation becomes disruptive may be higher than levels identified as statistically significant. Currently, little is known about the impact of various levels of autocorrelation in SSED data on the tests involved in MLM. The work of Jenson et al. (2007) and Ferron et al. (2009) helped to clarify the relationships of type I and II error, confidence interval accuracy, various design variables, and levels of autocorrelation. However, the limitations of their studies (e.g., use of a mean change model only with data that generated to not contain trends) prevent the findings from generalizing to other conditions. Related research on autocorrelation and alternative summary procedures for SSED data (e.g., modified R^2 indices; Beretvas & Chung, 2008) suggests that autocorrelation will likely pose a problem for MLM analyses in certain conditions. Future research should clarify which combinations of conditions are associated with biased standard errors and variance components, and compromised statistical tests.

If researchers have reason to assume that observed levels of autocorrelation will confound statistical tests, an autoregressive covariance structure could be specified for the model (Ferron et al., 2009; Van den Noortgate & Onghena, 2003a, 2003b, 2007; Verbeke & Molenberghs, 1997). This alternative covariance structure has the potential to attenuate or eliminate the impact of autocorrelation on analysis results. However, certain common characteristics of single-subject data (e.g., high variability and tendency toward large residuals) may interfere with the covariance structure's resolution of the problem of autocorrelation. Research is needed to confirm that the practice is efficacious with single-subject data across data conditions.

Analysis process. In both the articles which comment on use of MLM with SSED and those that apply the methods, address of the assumptions of MLM and the associated preliminary analyses is insufficient. A variety of conditions commonly found in SSED data can violate the assumptions and invalidate modeling outcomes and/or statistical tests. Preliminary analyses that involve checking the assumptions and distributions of variables are an important and necessary piece of the analysis process. Admittedly, the commentary articles are not guidebooks and are not responsible for detailing every step of the analysis process. However, all the applied studies should have checked the assumptions of MLM and distributions of variables. Authors' failure to have done so casts doubt on the validity of the studies' modeling outcomes and the appropriateness of models' specifications.

In the applied studies, authors' use of data in selection of model specifications was similarly insufficient. Formulation of an appropriate, good fitting model can lead to valuable insights related to best practice. However, use of an inappropriate, poor fitting model can obscure insights or produce false findings. As stated above, analysts should peruse graphs of data with an eye for general patterns, formulate multiple unconditional models in response to observed patterns, and comparatively evaluate the fit of each, in an effort to determine which model specifications are most appropriate for the data. Evaluation of models' fit should involve both statistical tests and visual analyses.

Both the commentary and applied articles inadequately addressed the formulation and statistical testing of final models. For one, authors of commentary articles did not discuss how to approach discovering that predictor variables contribute to final models insignificantly. Removal of insignificant predictor variables and the associated parameters can result in shifts in values among the retained parameters and greater accuracy of estimates (Raudenbush & Bryk, 2002). However, compelling theoretical justifications may exist for retaining insignificant predictors in

final models. When statistical insignificance is found for a predictor variable, a choice should be made and explained concerning its inclusion or exclusion from the final model. All authors of applied articles who found predictor variables to be insignificant retained the variables in their final models and did not discuss their decision to do so (i.e., Morgan & Sideridis, 2006; Wade, Ortiz, & Gorman, 2007; Wang, Cui, & Parrila, 2011). Features of their data sets provide reasons to assume that modeling outcomes were confounded by not removing the insignificant variables and associated parameters. For example, in Wade, Ortiz, & Gorman (2007), nearly all model slopes were found to be insignificant. Visual analysis of graphed data (all data modeled was presented in graphs) revealed behavior patterns differed across experimental phases in terms of level only. Thus, it is reasonable to assume that modeling slopes confounded estimates of the magnitudes of level changes between phases. On another note, the value of testing random effects was adequately described by authors of commentary articles (Nugent, 1996; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). Arguably, the examination of variability across subjects involved in tests of random effects produces the most interesting results of MLM. Such examination generates new information not obtained in primary studies, provides rationale for including predictor variables at levels 2 and higher, and indicates how much variance models explain. However, several authors of applied studies did not report taking the opportunity to examine variability across subjects (i.e., Morgan & Sideridis, 2006; Terrazas Arellanes, 2009; Wang, Cui, & Parrila, 2011). Consequently, they were unable to determine the need for and/or justify inclusion of predictor variables, and assess the amount of variance explained by their models.

Complimentary use of visual analysis in applications of MLM to primary research was well addressed/executed in both commentary and applied articles. As stated above, all authors of primary research employed visual analysis as a supplement to statistical analysis (Adams, 2009;

Terrazas Arellanes, 2009; Wade, Ortiz, & Gorman, 2006). Also, authors of 2 commentary articles recommended the practice (Nugent, 1996; Van den Noortgate & Onghena, 2007). Use of visual analysis may additionally be imformative in meta-analyses. To assess the degree to which final models explain data phenomena, researchers can visually analyze graphs of data and fitted regression lines for all subjects or a random sample.

Implications of the literature review for this study. This literature review identified a lack of empirical knowledge on MLM and single-subject data regarding data trended data, use of quadratic models at level 1, the impact of discontinuous variances across phases, the relative benefit of different level 1 error specifications, and the accuracy of autocorrelation estimates produced by statistical software used to estimate multilevel models. Also, the review showed little is known about the use of 3 level meta-analytic models with single-subject data and the impact of autocorrelation. In an effort to establish best practices for the meta-analysis of single-subject data, these topics will be addressed in this study.

CHAPTER 3

Methods

This dissertation sought to answer the following research questions:

When MLM is used to meta-analyze single-subject research,

1. What levels of power are achieved for statistical tests of fixed effects?
2. How accurate are estimates of fixed and random effects, and autocorrelation levels in terms of relative parameter bias across conditions examined?
3. What patterns of differences exist in convergence rates, power rates, and relative bias in estimates of fixed effects and random effects' variance components across (a) specifications for model errors at level 1, (b) numbers of data points per experimental phase, (c) numbers of participants per study, (d) numbers of studies meta-analyzed, (e) degrees of autocorrelation in individuals' data, and (f) continuity of level 1 variance across phases?

To answer the questions, a Monte Carlo simulation study was performed. Five factors were manipulated in the simulation of data: (a) number of data points per experimental phase, (b) number of participants per study, (c) number of studies meta-analyzed, (d) degree of autocorrelation in individuals' data, and (e) continuity of level 1 variance across phases. Two or three levels were selected for each factor. All factors were fully crossed to create 48 unique conditions resemblant of circumstances commonly encountered in the single-subject research literature. For each condition, 400 samples of data were generated. The data were generated using models and parameter values that produced realistic appearing single-subject data sets, which document behavior reduction treatment effects. Each sample was then meta-analyzed with 3 model varieties including different level 1 error specifications. All together, the simulation study

comprised 57,600 meta-analyses. The sections below detail the methods by which the levels of factors were selected, how data were generated and analyzed, and how the results were evaluated.

Collection and Analysis of Representative SSED Data

To assure the external validity of the study's results, data were generated to have characteristics typical of SSED data. Determination of these characteristics involved collection and analysis of

representative single-subject data from peer-reviewed journals.

Data collection. A large sample of data was collected from studies of interventions for self-injurious behavior in persons with developmental disabilities. These types of interventions were selected as the focus of the sample due to the extensive body of single-subject research on the topic and their typification of behavior reduction phenomena.

Search procedures. Individual data sets were identified through systematic searches of PsychInfo and ERIC databases. The search was conducted with the single Boolean search phrase "([self injur*] or [self harm] or [self destruct*]) and ([disabilit*] or [autis*] and [retard*])." Asterisks included in the search phrase caused any word beginning with the specified root to produce a search hit. The search was limited to peer-reviewed articles, written in English, published in and between 1960 and 2009. Abstracts of the resulting 1332 articles were reviewed for inclusion.

Selection criteria. To be included in the sample, data sets had to meet 5 criteria. The data sets were required to have (a) a quantitative measure of self-injurious behavior (e.g., head-hitting, skin picking) as the dependent variable, (b) treatment of self-injurious behavior as the independent variable, (c) a person with a developmental disability (e.g., autistic spectrum disorder, intellectual disability) as the subject, (d) a single-subject experimental design (Barlow, Nock, & Herson, 2009) that began with baseline measurements, followed by treatment sessions

and measurements, and (e) presented data for individual measurement sessions graphically or numerically.

Data sets were excluded if (a) data on challenging behaviors other than SIB (e.g., aggression toward others) were collapsed with data on SIB when reporting dependent variable outcomes (e.g., Neidert, Iwata, & Dozier, 2005; Volkert, et al., 2009), (b) data for phases were collapsed into a single number (e.g., Wachtel et al., 2008; Wachtel et al., 2009), (c) phases contained only one treatment and/or measurement session per phase (e.g., Oliver et al., 2006), (d) subjects discontinued multiple medications after baseline measurements (i.e., Lyskowski, Menditto, & Csernansky, 2009), (e) baseline or treatment conditions varied within phases (e.g., multi-element designs [e.g., Roberts, Mace, & Daggett, 1995; Vollmer, et al., 1998], stimulus gradation or response fading [e.g., Blindert, Hartridge, & Gwadry, 1995; Kahng, Abt, & Wilder, 2001]), (f) values along the x-axis of graphs were not discernable (i.e., Arntzen & Werner, 1999), or (g) treatment components were incorporated progressively over several phases until the treatment package of interest was complete (i.e., Mckenzie et al., 2008). Exclusion criteria pertained to data from the first baseline and first treatment sessions only (for rationale, see Extraction procedures below). Rationales for the exclusion criteria are as follows, respective to above: (a) assuring the internal validity of the data characteristics analysis required sampling only data pertaining to SIB, (b) collapsed data omits information required for input in multilevel meta-analysis, (c) the internal validity of studies involving single data points is low (excepting brief experimental designs, none of which were encountered in the literature search), (d) discontinuation of medication after baseline measurements diminishes the internal validity of findings, (e) changes in the independent variable within phases potentially confound dependent measures (f) data coding was not possible when x-values were not discernable, and (g) an intent of the studies which progressively incorporated treatment components was to demonstrate the

relative superiority of treatments or treatment packages introduced in subsequent phases; synthesis of data from such studies with those that intended to demonstrate the effect of a treatment or treatment package in the first treatment phases would have confounded the sample. A total of 199 individual data sets, from 122 studies, were selected for the sample.

Extraction procedures. Measurement outcomes were extracted from only the first baseline and first treatment phases. Data from only the first phases was collected because (a) the studies employed a great diversity of research designs (e.g., ABAB, multiple baseline AB, ABCBA, ABAC, etc), and therefore collection of a large, consistent sample of data required exclusion of data from later phases, and (b) the outcomes observed in the first phases were assumed to be representative of the interventions' effects. When articles presented multiple data sets for a subject (e.g., for generalization probes, or for various experimental conditions, such as those commonly seen in functional analyses), one single data set was chosen (e.g., that which pertained to the primary function of the behavior, or involved the initial treatment application).

Measurement outcomes were extracted from graphs using the Windows-based computer program Ungraph (Biosoft, 2004). Ungraph facilitates identification of X and Y values of data points through a point and click procedure. Digital images of graphs are first imported to an on-screen workspace. Scale calibrations are set by clicking upon three corners of the graph, and typing in known values. Next, each data point is clicked upon, and the program records its position along the X and Y axes. Coordinates for data points are then transferred automatically to Microsoft Excel files (Microsoft, 2007).

Analysis of representative data. The sample was analyzed to determine the ranges and central tendencies of numbers of data points per phase, numbers of participants per study, magnitudes of effects, and error variance and covariance within and between data sets. SAS PROC MEANS was used to compute means, medians, maximums and minimums for the

numbers of data points per phase and subjects per study. The magnitudes of effects were assessed by analyzing each data set using OLS regression procedures and Equation 29 (listed below and in Chapter 2). Coefficients for the baseline intercept, baseline slope, and treatment phase change in level, change in linear slope, and curvilinear slope were included in the regression model (i.e., Equation 29). Since the data derived from multiple dependent variable metrics (e.g., frequency counts, partial 10 second intervals), distributions and central tendencies were estimated separately for each metric. Coefficient estimates were not standardized in an effort to protect data from bias which may result from standardization procedures. After estimates were obtained for all individual data sets, SAS PROC MEANS was used to compute means, medians, maximums, and minimums for each coefficient from each dependent variable metric. To assess whether or not coefficient estimates were correlated, Pearson correlations were estimated and scatterplots made for each pair of coefficients from each metric. Finally, values for the error variances and covariances within and between data sets were estimated for each metric using SAS PROC MIXED and a 3 level meta-analytic model.

Data Simulation

Factors and simulation conditions. As stated above, five factors were manipulated in the study: (a) number of data points per phase, (b) number of participants per study, (c) number of studies meta-analyzed, (d) degree of autocorrelation in individuals' data, and (e) continuity of level 1 variance across phases. Two levels were chosen for each of the factors, except the number of data points per phase, for which 3 levels were selected (see Table 3 for a summary of factor levels). All factors were fully crossed to create 48 simulation conditions. Thus, a 3 x 2 x 2 x 2 x 2 factorial design, with 48 cells, was employed.

Results of analysis of the representative sample of single-subject data, previous simulation studies, other analyses, and theory guided selection of the levels of factors. In the

representative sample, the mean and median numbers of data points were 9.7 and 8, respectively for baseline phases, and 11.8 and 8 respectively for treatment phases. These numbers of data points ranged from 3 to 58 for baseline phases and 2 to 66 for treatment phases. The 3 levels selected for the number of data points per phase represent below average, average, and above average counts. The levels chosen were 5 baseline and 5 treatment data points, 10 baseline and 10 treatment data points, and 20 baseline and 20 treatment data points. In the representative sample, the number of participants per study ranged from 1 to 7 and averaged 1.2. To reflect this range, as well as maintain consistency with previous simulation studies (i.e., Ferron et al., 2010; Van den Noortgate & Onghena, 2011), the levels selected for the number of participants per study were 3 and 6. Selection of the levels of number of studies meta-analyzed was based on a previous meta-analytic simulation study (Van den Noortgate & Onghena, 2011) and the present author's anecdotal observations of typical numbers of studies included in reviews of single-subject research. The levels selected for the number of studies were 10 and 30. Levels for the degree of autocorrelation in individuals' data were also based on previous simulation studies (Bell et al., 2011; Beretvas & Chung, 2008; Ferron et al., 2010; Van den Noortgate & Onghena, 2011), as well as a previous analysis of autocorrelation rates in single subject data (Huitema, 1985). The levels chosen for autocorrelation rates were 0.0 and 0.4. The final factor for which levels were selected was the continuity of level 1 variance across phases. It's theoretically plausible that the variance of single-subject data is often not continuous across phases (Van den Noortgate & Onghena, 2003b). In the representative sample, this proved to be the case. Baseline variances were consistently much larger than treatment phase variances. To examine the performance of multilevel meta-analysis when variance is not continuous across phases, the levels selected were continuous variance and discontinuous variance. The continuous level provided a set of baseline conditions (which were hypothesized to be associated with optimal performance), while the

discontinuous level presented contrast conditions to be judged relative to the baseline. In accord with results of analysis of the representative sample, continuous variances (i.e., σ^2_{single}) were set at 150, while discontinuous variances (i.e., $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$) were set at 300 and 70 for baseline and treatment phases, respectively.

Generating equations and parameter values. Data samples were generated using the SAS computer program (SAS Institute Inc., 2008). Data were generated to fit the following equation in the population at level 1:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}T_{ijk} + \pi_{2jk}(\text{treatment})_{ijk} + \pi_{3jk}(T_{ijk} - [n_{tjk} + 1])(\text{treatment})_{ijk} + \pi_{4jk}(T_{ijk} - [n_{tjk} + 1])^2(\text{treatment})_{ijk} + e_{ijk} \quad (29)$$

where all symbols defined for Equation 29 in chapter 2 have the same meaning; and at level 2:

$$\pi_{0jk} = \beta_{0k} + r_{0jk} \quad (30)$$

$$\pi_{1jk} = \beta_{1k} + r_{1jk} \quad (31)$$

$$\pi_{2jk} = \beta_{2k} + r_{2jk} \quad (32)$$

$$\pi_{3jk} = \beta_{3k} + r_{3jk} \quad (33)$$

$$\pi_{4jk} = \beta_{4k} + r_{4jk} \quad (34)$$

where π parameters are level 1 regression coefficients for subject *j* from study *k*, β parameters are averages of π parameters within study *k*, and *r* parameters are error terms that represent the deviation of subject *j*'s π parameter from study *k*'s average; and, finally, at level 3:

$$\beta_{0k} = \gamma_0 + u_{0k} \quad (35)$$

$$\beta_{1k} = \gamma_1 + u_{1k} \quad (36)$$

$$\beta_{2k} = \gamma_2 + u_{2k} \quad (37)$$

$$\beta_{3k} = \gamma_3 + u_{3k} \quad (38)$$

$$\beta_{4k} = \gamma_4 + u_{4k} \quad (39)$$

where β parameters are level 2 regression coefficients for study *k*, γ parameters are grand means (i.e., averages of β parameters across studies), and *u* parameters are the deviations of study *k*'s averages from the grand means.

Results of analysis of the representative sample of single-subject data guided choices of parameter values for data generation. The regression analyses of individual data sets led to selection of population values for each γ coefficient. All values chosen were selected due to their approximation of the average coefficient value for multiple metrics and their facilitation of production of realistic appearing data sets which possessed many common single-subject data characteristics. The following values were chosen for the γ coefficients:

$$\gamma_0 = 60$$

$$\gamma_1 = 0.25$$

$$\gamma_2 = -25$$

$$\gamma_3 = -4.5$$

$$\gamma_4 = -0.075$$

Figure 2 graphically depicts the population average model. During the first session, a behavior level of 60 is expected, as given by γ_0. Across baseline sessions, behavior is expected to rise at a rate of 0.25 units per session, as reflected in γ_1. When treatment begins, behavior is expected to immediately drop 25 units, as seen in γ_2. Across the treatment phase, behavior decreases with a slope composed of γ_1, γ_3, and γ_4. Together, γ_1 and γ_3 sum to make the linear

component of the slope (0.25 + -4.5 = -4.25). The curvilinear component is determined by γ_4, which is -0.075.

Due to the chosen parameter values, simulated data points for individual subjects could have y-values of less than zero. This artifactual error results from the nature of quadratic functions. Quadratic models do not have the advantage of forming an asymptote at a floor or ceiling value, such as 0 in this case. To prevent such unrealistic values during simulation, generated data points with y-values of less than zero were changed to have a y-value of 0 prior to adding level 1 errors (i.e., e_{ijk}) and lag-1 autoregressive processes.

Selection of values for variances of errors at levels 2 and 3 was guided by results of the 3 level meta-analyses for data on each metric. Similar to above, all values chosen were selected due to their approximation of the error variances for multiple metrics, their consistency with the ICC values obtained from the multilevel meta-analyses, and their facilitation of production of realistic appearing data sets. Selection of values for covariances was guided by the correlation estimates obtained for pairs of coefficients from the regression analyses. Non-zero covariances were chosen for r_{3jk} and r_{4jk}, as well as u_{3k} and u_{4k} due to the discovery of significant and substantial correlations between the regression coefficients π_3 and π_4. A correlation of -0.60 was chosen for the generation of both pairs of errors, due to its approximation of the average correlation across metrics and facilitation of production of realistic appearing data sets. The correlation value was converted into covariances, which are included in the matrices below. Because insignificant and insubstantial correlations were found for all other pairs of regression coefficients, covariances of the remaining errors were generated to be zero.

The correlation of π_3 and π_4/ r_{3jk} and r_{4jk}, and u_{3k} and u_{4k} allows the treatment phase regression lines to bow in concave and convex manners (see Figure 3 for an illustration). When the first and last data points in the treatment phase are fixed, this bowing resembles a

reverberating guitar string that makes a downward arch when π_3 is low (i.e., is negative and has a large magnitude) and π_4 is high (i.e. is greater than zero), and makes an upward arch when π_3 is high (is negative or positive and has a small magnitude) and π_4 is low (i.e., is negative and has a relatively high magnitude). The downward arch pattern is the quadratic model's closest offering of an asymptotic curve that plateaus at the floor value of 0. The upward arch pattern models a slow, accelerating learning curve.

All errors were generated to be normally distributed with a mean of 0 and the following variances and covariances:

$$\text{var } r_{jk} = \begin{bmatrix} 64 & 0 & 0 & 0 & 0 \\ 0 & 0.09 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 \\ 0 & 0 & 0 & 7.29 & -0.1944 \\ 0 & 0 & 0 & -0.1944 & 0.0144 \end{bmatrix}$$

$$\text{var } u_k = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0.144 & -0.0009 \\ 0 & 0 & 0 & -0.0009 & 0.000016 \end{bmatrix}$$

Figures 4 through 6 present three random samples of 10 data sets generated by SAS. In Figure 4, data sets are composed of 5 baseline and 5 treatment data points. Figure 5 contains data sets that have 10 baseline and 10 treatment data points. In Figure 6, data sets have 20 baseline and 20 treatment data points. All data were generated to have a baseline variance of 300, a treatment variance of 70, a lag-1 autocorrelation of 0.0, and random errors for studies, subjects, and time points. The graphs are presented here as confirmation that the simulated data have a realistic character and resemble single-subject data commonly encountered in the research literature. Confounds introduced and study limitations created by the generated data are addressed in Chapters 4 and 5.

Analysis of Simulated Data

Meta-analyses. Simulated data samples were meta-analyzed using SAS PROC MIXED. Each sample was meta-analyzed 3 times with 3 different specifications for level 1 errors: (a) by defaulting to specification of different variance components and 0 covariances for level 1 errors, (b) by specifying lag-1 autoregressive covariance structures for level 1 errors (which allowed non-zero covariances), and (c) by specifying separate error terms for each phase in level 1 models and defaulting to the specification of different variance components and 0 covariances for each error term. Across these 3 analyses, all terms of the level 1 equations were identical, except those for the errors.

When specifying different variance components and 0 covariances, as well as lag-1 autoregressive covariance structures, the following 3 level model was fit to each sample generated:

level 1:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}T_{ijk} + \pi_{2jk}(\text{treatment})_{ijk} + \pi_{3jk}(T_{ijk} - [n_{bjk} + 1])(\text{treatment})_{ijk} + \pi_{4jk}(T_{ijk} - [n_{bjk} + 1])^2(\text{treatment})_{ijk} + e_{ijk} \quad (25)$$

level 2:

$$\pi_{0jk} = \beta_{0k} + r_{0jk} \quad (26)$$

$$\pi_{1jk} = \beta_{1k} + r_{1jk} \quad (27)$$

$$\pi_{2jk} = \beta_{2k} + r_{2jk} \quad (28)$$

$$\pi_{3jk} = \beta_{3k} + r_{3jk} \quad (29)$$

$$\pi_{4jk} = \beta_{4k} + r_{4jk} \quad (30)$$

level 3:

$$\beta_{0k} = \gamma_0 + u_{0k} \quad (31)$$

$$\beta_{1k} = \gamma_1 \quad (32)$$

$$\beta_{2k} = \gamma_2 \quad (33)$$

$$\beta_{3k} = \gamma_3 \quad (34)$$

$$\beta_{4k} = \gamma_4 \quad (35)$$

When specifying separate error terms for each phase in level 1 models, the following level 1 equation replaced Equation 29 in the above 3 level model. Compared to Equation 29, this equation possesses adaptations to the error term.

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{time})_{ijk} + \pi_{2jk}(\text{treatment})_{ijk} + \pi_{3jk}(\text{time} - (n_{bjk} + 1))(\text{treatment})_{ijk} + \pi_{4jk}(\text{time} - (n_{bjk} + 1))^2(\text{treatment})_{ijk} + e_{ijk}(1 - \text{treatment})_{ijk} + e_{ijk}(\text{treatment})_{ijk} \quad (36)$$

These 3 level models can be termed simplification models. The lack of error terms in Equations 36 through 39 forces the study averages of π parameters (i.e., the β parameters) to equal the grand averages. While γ parameters were generated to randomly vary across studies, they were analyzed with the specification that they do not. A simplification model was chosen for analysis due to the fact that fewer units at each level are needed for computation of parameter estimates (Hox, 2010). Simplification models have the potential to be of frequent utility in meta-analyses of single subject research due to the typically small number of data points collected per subject (i.e., level 1 units), the small numbers of subjects included in studies (i.e., level 2 units) and the small numbers of well matched studies (i.e., whose results are fit for aggregation) that typically constitute the samples of published reviews (i.e., level 3 units). Support for the choice of a simplification model was also obtained in analysis of the representative sample. Specification of models for analysis which matched the generating models frequently did not lead to analysis convergence. However, specification of the above simplification model did consistently lead to convergence.

For level 2 errors, analyses involved specification of different variance components and covariances of 0. At level 3, only one variance component was estimated: that for the random effect u_{0k}.

From each analysis, estimates of fixed effects, *p*-values resulting from significance tests of fixed effects, and variances of error terms at levels 1, 2, and 3 were accumulated. For analyses in which lag-1 autoregressive covariance structures were specified, estimates of the lag-1 autocorrelation were also accumulated. After all 400 replications of a single condition were complete, the accumulated figures were analyzed. These analyses yielded values for convergence rates, power of statistical tests of each fixed effect, and relative parameter bias for each fixed effect, each random effect's variance component, and autocorrelation estimates.

The convergence rate is the frequency with which the iterative computations involved in estimating multilevel models successfully identify stable solutions for models' fixed effects. A stable solution in SAS is obtained when the convergence criterion (i.e., a statistic calculated for each iteration of the estimation process) is less than 1E-8. To determine the convergence rate for each condition by level 1 error specification, the numbers of accumulated estimates of fixed effects were counted and divided by 400. When analyses fail to converge in SAS, the software does not output fixed effect estimates. Thus, the convergence rate for a condition is the percent of estimates obtained out of the total possible 400.

The power of statistical tests of fixed effects is the rate at which truly present effects are confirmed to exist in a sample (i.e., the effects are estimated to have a 95% probability of having a non-zero value). To assess the power of statistical tests of fixed effects, *p*-values resulting from statistical tests of the fixed effects were sorted and counted. Because generating values for all fixed effects were non-zero (i.e., effects were generated), tallies were made of falsely insignificant test results ($p > .05$). The tallies were then divided by 400 (i.e., the total number of

p-values for a single fixed effect). Finally, the quotients were subtracted from 1 to compute power for the condition. Power of less than 0.8 was considered inadequate.

Relative parameter bias is the degree to which an analysis technique accurately recovers a true parameter value. Sampling error, the resulting variance within and between subjects and studies, and autocorrelation threaten to obscure true parameter values and bias parameter estimates. To evaluate the robustness of the analysis techniques for each condition, relative bias statistics will be calculated for all fixed effects, random effects' variance components, and autocorrelation estimates. The following equation will be used to calculate relative bias (Hoogland & Boomsma, 1998):

$$B(\widehat{\theta}_i) = \frac{\overline{\widehat{\theta}}_i - \theta_i}{\theta_i} \tag{31}$$

where $\overline{\widehat{\theta}}_i$ is the average of the 400 estimates of parameter *i* for a given condition and θ_i is the true value of parameter *i*. According to Hoogland and Boomsma (1998), relative bias values above .05 in magnitude are considered unacceptable. When the true parameter value is 0, as is the case when autocorrelation = 0.0, relative bias cannot be calculated. In this circumstance, estimates will be considered biased when the mean absolute value of the sample bias exceeds .05 (i.e., when the mean point estimate is outside the range between -.05 and .05).

Tests of the statistical significance of random effects' variance components were not evaluated. The focus of interest in this study is limited to convergence rates, the power of statistical tests of fixed effects and parameter recovery for fixed effects, random effects' variance components, and autocorrelation estimates.

CHAPTER 4

Results

Results of the simulation study are summarized in the sections below. Outcomes are reported for convergence rates, power of statistical tests for fixed effects, and relative parameter bias for fixed effects, random effects' variance components, and autocorrelation estimates. In the tables below, results are first grouped by levels of continuity of variance and autocorrelation, and next organized by levels of number of data points per phase, number of studies meta-analyzed, and number of subjects per study, and finally sorted by level 1 error specifications.

Convergence Rates

For all conditions and error specifications, 100% of analyses met the default convergence criteria of SAS.

Power of Statistical Tests for Fixed Effects

Tables 4 through 7 present the power rates observed for statistical tests by fixed effect and level 1 error specification. In sequential order, the tables present power rates for conditions marked by (a) continuous variance and no autocorrelation, (b) continuous variance and positive autocorrelation, (c) discontinuous variance and no autocorrelation, and (d) discontinuous variance and positive autocorrelation. Across the 4 groupings of levels of continuity of variance and autocorrelation (hereafter referred to as "groupings" and by their order of presentation, i.e., first, second, etc), fairly similar results were obtained.

Below, results are first described by fixed effect. As mentioned in Chapter 3, power was considered inadequate if a rate of less than 0.8 was observed. Then, patterns of association between relative bias levels, factors, and level 1 error specifications are identified.

Results by fixed effect. For the two large effects, γ_0 (i.e., the model intercept) and γ_2 (i.e., the immediate treatment effect), power rates were consistently 1.0 across conditions and level 1 error specifications.

For the relatively moderate effect, γ_3 (i.e., the linear component of the treatment phase slope change), power rates ranged between 0.8 and 1.0 for all conditions and level 1 error specifications, except conditions with the smallest numbers of data points (i.e., 5 baseline and 5 treatment) and studies meta-analyzed (i.e., 10). Some differences existed between groupings and level 1 error specifications in power rates for conditions with 6 participants per study and the smallest numbers of data points and studies (i.e., 5 and 5, and 10, respectively). However, these differences were slight. Specifically, when autocorrelation wasn't generated and either different variance components or lag-1 autoregressive covariance structures were specified, power rates ranged from 0.715 to 0.803. When autocorrelation was generated and either different variance components or lag-1 autoregressive covariance structures were specified, power rates ranged from 0.868 to 0.978.

For the first of the small effects, γ_1 (i.e., the baseline slope), power rates were characterized by fairly consistent patterns across groupings. Power rates were inadequate for all conditions with 5 baseline and 5 treatment or 10 baseline and 10 treatment data points. Some differences existed between groupings in power rates for conditions with 20 baseline and 20 treatment data points. Specifically, when autocorrelation was not generated (i.e., in the first and second groupings), power rates were slightly inadequate only when the number of subjects and studies were smallest (i.e., 3 and 10, respectively; power ranged from 0.763 to 0.790). This pattern had one exception: when the number of subjects and studies were 6 and 10, respectively, and separate level 1 error terms were specified, power was observed at 0.748. In contrast, when

autocorrelation was generated (i.e., in the third and fourth groupings), power rates were only adequate for conditions with the largest numbers of subjects and studies (i.e., 6 and 30, respectively).

For the second small effect, γ_4 (i.e., the curvilinear component of the treatment phase slope), power rates followed a very consistent pattern across groupings. Power rates were inadequate in all conditions with 5 baseline and 5 treatment data points. Power rates were additionally inadequate in all conditions with 10 baseline and 10 treatment data points, when the number of subjects and studies were smallest (i.e., 3 and 6, respectively). The one exception to this pattern was seen when variance was discontinuous, autocorrelation was generated, 10 data points were included in each phase, the number of subjects and studies were 6 and 10, respectively, and a lag-1 autoregressive covariance structure was specified. In this instance, the power for γ_4 was 0.753.

Identification of patterns in power rates. Overall, power rates appeared to be most closely associated with the magnitude of effects. High power was consistently observed for the two large effects (i.e., 1.0). As the magnitude of the effects decreased, increases were observed in the number of conditions in which power was found to be inadequate. It should be noted that while the generating population value for γ_4 (i.e., -0.075) was smaller than that for γ_1 (i.e., 0.25), the values were not on the same scale. In a sense, the magnitude of γ_1 was smaller. The parameter γ_4 was multiplied by a squared term in the combined model, and consequently influenced the outcomes of data generation at a different gradient than did γ_1. It thus makes sense that γ_1 is associated with the largest number of inadequate power rates across conditions.

Factors which dictated the unit counts at level 1, 2, and 3 also appeared to be closely associated with power rates. As the numbers of data points per phase, subjects per study, and studies meta-analyzed increased, power rates also increased. Given the generating parameter values, it appears that when 10 data points are collected for each phase and the number of level 2 units is 60 or greater (i.e., number of subjects per study x number of studies ≥ 60), power rates are adequate for all but very small effects (e.g., γ_1).

In conditions with discontinuous variance (i.e., the second and fourth groupings), it appears specification of separate error terms is associated with slightly improved power relative to the other level 1 error specifications. In 4 instances, when one or more of the other level 1 error specifications are associated with inadequate power, the power level associated with separate error terms is adequate and substantially greater. Also, when power levels are adequate across level 1 error specifications, the power associated with separate error terms tends to be greatest (i.e., greater by a range of 0.005 to 0.092).

Power rates did not appear to be associated with continuity of variance or autocorrelation level. While a few slight differences existed between power rates for levels of each factor, the differences were limited to the comparison of several pairs of conditions and primarily pertained to γ_1.

Relative Parameter Bias of Fixed Effects

Tables 8 through 11 present the relative bias levels observed for models' fixed effects by parameter and level 1 error specification. As in the tables of fixed effects' power rates, Tables 8 through 11 sequentially present results for conditions grouped according to factor levels for continuity of variance and autocorrelation. Across groupings, relative bias was observed to vary systematically for γ_1 and γ_3. Results for γ_0, γ_2, and γ_4 were consistent across groupings.

Below, results are first discussed by fixed effect. As mentioned in Chapter 3, parameters were considered biased if the relative bias statistic was greater in magnitude than 0.05. Then, patterns of association between relative bias levels, factors, and level 1 error specifications are identified.

Results by fixed effect. For γ_1 (i.e., the baseline slope), bias was observed least often when variance was continuous and autocorrelation not generated (i.e., in the first grouping). When variance was discontinuous and/or autocorrelation was generated, bias was observed for greater numbers of conditions and level 1 error specifications. In the first grouping, bias was observed in only one condition: when numbers of data points, subjects, and studies were smallest (i.e., 5 and 5, 3, and 6, respectively; magnitudes of relative bias ranged from 0.068 to 0.073). In the second grouping, when variance was discontinuous and autocorrelation not generated, bias was observed for 5 additional conditions. In these instances, magnitudes of relative bias ranged from 0.054 to 0.094. No associations were evident among bias and numbers of data points, subjects, or studies across this total of 6 conditions. Bias was observed at all levels of the number of data points, number of subjects, and number of studies factors across conditions in the second grouping. Compared to the first grouping, the third grouping (i.e., when variance was continuous and autocorrelation was modeled) was marked by increased bias among analyses employing autoregressive covariance structures. For this third grouping and level 1 error specification, relative bias exceeded the cut-off magnitude of 0.05 in 7 conditions (ranging from 0.056 to 0.170). As in the comparison between the first and second groupings, the fourth grouping (i.e., when variance was discontinuous and autocorrelation was generated), as compared to the third grouping, was marked by increased bias across all conditions, with no apparent association with numbers of data points, subjects, or studies. In the third grouping, bias was observed in 8 conditions, for a total of 11 level 1 error specifications (magnitudes ranged from 0.054 to 0.170).

In the fourth grouping, bias was observed in 10 conditions, for a total of 26 level 1 error specifications (magnitudes ranged from 0.051 to 0.129). Between the second and fourth groupings, which shared discontinuous variance, more frequent instances of bias were observed when autocorrelation was generated (i.e., in the fourth grouping). In the second grouping, when autocorrelation was not generated, 16 instances of bias were observed (magnitudes ranged from 0.054 to 0.094). In the fourth grouping, bias was observed in 26 instances (magnitudes ranged from 0.051 to 0.129).

For γ_3 (i.e., the linear component of the change in treatment phase slope), bias was also observed least often when variance was continuous and autocorrelation was not generated, and more often when variance was discontinuous and/or autocorrelation was generated. In the first grouping, bias was only observed when the numbers of data points per phase was 10 (magnitudes ranged from 0.186 to 0.193). When variance was discontinuous (i.e., in the second and fourth groupings), bias was additionally observed in all conditions with 20 data points per phase, and conditions with 5 data points per phase, when the studies numbered 10 (magnitudes ranged from 0.056 to 0.222 in the second grouping and 0.052 to 0.220 in the fourth grouping). When comparing the first and third groupings, which shared continuous variance, additional bias was observed in conditions with 5 data points per phase and 90 level 2 units or less, when autocorrelation was generated (i.e., in the third grouping; magnitudes ranged from 0.056 to 0.194). No differences were observed in frequency of biased estimates between the second and fourth groupings, which shared discontinuous variance and varied on levels of autocorrelation.

As stated above, results for γ_0, γ_2, and γ_4 were consistent across groupings. No bias was observed in estimates of γ_0 or γ_2. For all conditions and level 1 error specifications, estimates of

γ_4 were greatly biased in the negative direction (magnitudes ranged from 0.165 to 3.91 across conditions and level 1 error specifications).

Identification of patterns in relative bias of fixed effect estimates. Overall, bias in fixed effects estimates appeared to be most closely associated with the factor levels of continuity of variance and autocorrelation. Between levels of continuity of variance, biased estimates were observed more often when variance was discontinuous. Most notably, collection of 20 data points per phase was generally sufficient to produce unbiased estimates when variance was continuous, but insufficient when variance was discontinuous, regardless of level 1 error specification. Between levels of autocorrelation, biased estimates were observed more often when autocorrelation was generated.

The type of fixed effect also appeared closely associated with bias. Estimates for the two intercepts, γ_0 and γ_2, were observed to be unbiased across all conditions. However, estimates for the three slopes/slope components were frequently biased. Generally, estimates were slightly biased for the linear slopes and greatly biased for the curvilinear slope component.

Relative bias generally appeared to improve as the number of data points per phase increased. As noted above, the least bias was observed in conditions with 20 data points per phase. Extrapolation of trends in bias statistics for γ_1 and γ_3 suggests that the effects might be estimated without bias when the number of data points per phase reaches 30. However, extrapolation of trends in bias for γ_4 suggests that the number of data points per phase may have to rise to 40 or 50 before unbiased estimates can be obtained.

The level 1 error specification appeared to have limited associations with relative bias in fixed effects. When autocorrelation was generated, specification of an autoregressive covariance

structure appeared to be associated with more frequent bias in estimates of γ_1, yet also associated with lower levels of bias in estimates of γ_3.

No patterns of association with relative bias were observed for numbers of subjects per study or numbers of studies meta-analyzed. While some slight differences existed across levels of the factors, the differences were not systematic, nor substantial.

Relative Parameter Bias of Random Effects' Variance Components

Tables 12 through 19 present the relative bias observed in variance components of each random effect by factor levels and level 1 error specification. Tables contain results for single random effects, except Table 19, which presents results for both $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$ together. Across random effects, conditions, and level 1 error specifications, relative bias statistics for variance components had acceptable magnitudes (i.e., of less than 0.05) very rarely.

Below, results and patterns of association are discussed for each random effect' variance component separately. As mentioned in Chapter 3, estimates were considered biased if the relative bias statistic was greater in magnitude than 0.05.

Results for $T_{\gamma 0}$. Table 12 presents relative bias statistics for estimates of $T_{\gamma 0}$. The variance component $T_{\gamma 0}$ represents the variability of studies' average baseline intercept. In all but 3 instances, estimates of $T_{\gamma 0}$ were observed to be biased. Magnitudes of the unacceptable levels of relative bias ranged from 0.069 to 5.59. Acceptable levels of bias were observed when variance was discontinuous, autocorrelation was not generated, data points per phase, subjects, and studies numbered 20, 3, and 10, respectively, and either different variance components or separate error terms were specified for level 1 errors. Additionally, an acceptable level of bias was observed when variance was discontinuous, autocorrelation was generated, data points per phase, subjects,

and studies numbered 5, 6, and 30, respectively, and separate error terms were specified for level 1 errors. In these 3 instances, relative bias magnitudes ranged from 0.019 to 0.043.

Identification of patterns in relative bias of $T_{\gamma 0}$. No patterns were apparent among instances of unbiased estimates of $T_{\gamma 0}$. Given their inconsistency with the other relative bias statistics for $T_{\gamma 0}$ estimates, these instances may represent false positives. Despite the lack of patterns among unbiased estimates, two patterns among the biased estimates deserve mention. When 5 or 10 data points were generated for each phase, increases in the numbers of level 2 units (i.e., due to increases in either the number of subjects or studies, or both) was associated with decreases in relative bias. However, extrapolation of this decreasing trend suggests that unbiased estimates cannot be attained in realistic scenarios. Additionally, when autocorrelation was generated, specification of an autoregressive covariance structure was associated with less biased estimates than when different variance components or separate error terms were specified. However, clear patterns were not apparent across the levels of other factors. Thus, extrapolation of the pattern in an effort to theorize factor levels for which estimates were unbiased was not possible.

Results for $T_{\beta 0}$. Table 13 presents relative bias statistics for estimates of $T_{\beta 0}$. The variance component $T_{\beta 0}$ represents the variability in subjects' baseline intercepts within studies. Compared to the other random effects, estimates of $T_{\beta 0}$ were frequently unbiased. However, for the 144 pairs of conditions and level 1 error specifications, acceptable levels of relative bias were observed in only 22 instances. The magnitudes of the remaining 122 relative bias statistics ranged from 0.054 to 1.78.

Most instances of acceptable levels of bias were observed when variance was continuous and autocorrelation not generated. Fourteen instances were found across all level 1 error

specifications and conditions with 5 and 10 data points per phase, and 60 or more level 2 units. Exceptions to this pattern were seen in conditions with 5 data points per phase, 3 subjects per study, and 30 studies, across all level 1 error specifications, as well as 5 data points per phase, 6 subjects per study, and 10 studies, when an autoregressive covariance structure was specified. In these exceptions, relative bias ranged from -0.061 to -0.098.

When variance was discontinuous and autocorrelation was not generated, acceptable levels of bias were observed in 3 instances: in conditions with 5 data points per phase, 6 subjects per study, and either 10 or 30 studies, and when the numbers of data points, subjects, and studies numbered 10, 3, and 10, respectively.

When autocorrelation was generated, acceptable bias was observed in 5 instances when autoregressive covariance structures were specified. When variance was continuous, acceptable levels of bias were found when the numbers of data points, subjects, and studies numbered 5, 3, and either 10 or 30, respectively. When variance was discontinuous, acceptable levels of bias were observed in conditions with 10 data points per phase and more than 60 level 2 units.

Identification of patterns in relative bias of $T_{\beta 0}$. It appears that factors which dictate level 1 and level 2 unit counts are most closely associated with the relative bias of $T_{\beta 0}$. More instances of acceptable levels of bias were found in conditions with 10 data points compared to 5, 6 subjects compared to 3, and 30 studies compared to 10. However, this pattern did not maintain for conditions with 20 data points per phase, for which no acceptable levels of bias, nor systematic decreases in relative bias were observed across levels of numbers of data points per phase, subjects per study and studies meta-analyzed.

When autocorrelation was generated, it appears specification of an autoregressive covariance structure was associated with decreased relative bias. This pattern was consistently

observed across levels of continuity of variance, and numbers of data points per phase, subjects per study, and studies meta-analyzed.

Relative bias appeared to increase across levels of autocorrelation and continuity of variance. Between conditions with continuous variance and discontinuous variance, those with discontinuous variance tended to be associated with higher levels of relative bias. Similarly, conditions in which autocorrelation was generated tended to be associated with higher levels of relative bias than those in which autocorrelation was not generated.

Results for $T_{\beta 1}$. Table 14 presents relative bias statistics for estimates of $T_{\beta 1}$. The variance component $T_{\beta 1}$ represents the variability in subjects' baseline slopes around the grand mean (i.e., γ_1). In all but one instance, estimates of $T_{\beta 1}$ were observed to be biased. When variance was continuous, autocorrelation was not generated, and the numbers of data points per phase, subjects per study, and studies meta-analyzed were 5, 3, and 30, respectively, the estimate of $T_{\beta 1}$ was of an acceptable level of bias. Magnitudes of relative bias for the remaining 143 pairs of conditions and level 1 error specifications ranged from 0.059 to 21.3.

Identification of patterns in relative bias of $T_{\beta 1}$. Because only one instance of acceptable bias was observed for estimates of $T_{\beta 1}$, patterns among acceptable levels of relative bias were not identifiable. Given the inconsistency with the other relative bias statistics for $T_{\beta 1}$ estimates, this instance may represent a false positive.

Several patterns in relative bias among the biased estimates deserve mention. Across levels of autocorrelation, relative bias appeared to increase. When autocorrelation was not generated, relative bias was observed to have magnitudes as large as 1.15. However, when autocorrelation was generated, relative bias was found to rise to magnitudes as high as 21.3. Also, for many conditions, relative bias levels approximated -1. This was especially the case when autocorrelation was not generated and data points per phase numbered 10 or 20, and when

autocorrelation was generated and autoregressive covariance structures were specified. Relative bias statistics of -1.00 indicate that the variance of baseline slopes was estimated to be 0 for all replications of a condition and level 1 error specification pair. When relative bias levels approximate -1, it can be assumed that the majority of variance estimates were 0 for the condition and level 1 error specification pair.

Results for $T_{\beta 2}$. Table 15 presents relative bias statistics for estimates of $T_{\beta 2}$. The variance component $T_{\beta 2}$ represents the variability around the grand mean (i.e., γ_2) of estimates of immediate treatment effects for subjects. Estimates for all pairs of conditions and level 1 error specification were observed to be biased. Magnitudes of relative bias of the estimates ranged from 0.082 to 13.8.

Identification of patterns in relative bias of $T_{\beta 2}$. Levels of the relative bias of estimates of $T_{\beta 2}$ appeared to be associated with factor levels of autocorrelation, and to a limited degree with numbers of data points per phase and level 1 error specification. When autocorrelation was not generated, relative bias levels were lower (i.e., ranging from 0.082 to 3.39) than when autocorrelation was generated (i.e., levels ranged from 1.32 to 13.8). Also, when autocorrelation was generated and autoregressive covariance structures were specified, relative bias levels decreased as the number of data points per phase increased. Extrapolation of this trend suggests that between 40 and 50 data points per phase are required for unbiased estimates of $T_{\beta 2}$, when autocorrelation is present and autoregressive covariance structures are specified. However, an opposite trend was observed across level 1 error specifications when autocorrelation was not generated. In these conditions, as the number of data points increased, relative bias levels also increased.

Results for $T_{\beta 3}$. Table 16 presents relative bias statistics for estimates of $T_{\beta 3}$. The variance component $T_{\beta 3}$ represents the variability around the grand mean (i.e., γ_3) of estimates of

changes in linear components of treatment phase slopes for subjects. In all but one instance, estimates of $T_{\beta 3}$ were found to be biased. When variance was discontinuous, autocorrelation generated, and the numbers of data points per phase, subjects per study, and studies meta-analyzed were 5, 3, and 10, respectively, the relative bias observed was less than 0.05. In the other 143 instances, the magnitude of relative bias was found to range from 0.053 to 1.00.

Identification of patterns in relative bias of $T_{\beta 3}$. The one instance of an acceptable level of bias appeared to be part of a complex association between relative bias, numbers of data points per phase, level 1 error specification, and levels of continuity of variance and autocorrelation. When variance was continuous and autocorrelation was generated, or variance was discontinuous, and separate level 1 error terms were specified for each phase, relative bias was less than when different variance components or autoregressive covariance structures were specified. This difference in relative bias statistics between level 1 error specifications increased as the numbers of data points per phase decreased. This pattern seems to indicate that bias in $T_{\beta 3}$ was attenuated by specification of separate error terms for conditions marked by continuous variance and autocorrelation, or discontinuous variance. However, the pattern may also be explained, in part, by the relationship between numbers of data points per phase and floor effects. As floor effects increase and distributions of y-values are more greatly restricted, the slopes of treatment phase regression lines are also restricted. The restriction of treatment phase slopes reduces the variability in estimates of π_3 and increases the frequency at which $T_{\beta 3}$ is estimated to be 0 or some very small number. The impact of floor effects on variability in estimates of π_3 is likely responsible for the manner in which relative bias approaches –1 in the most rows of Table 16.

Results for $T_{\beta 4}$. Table 17 presents relative bias statistics for estimates of $T_{\beta 4}$. The variance component $T_{\beta 4}$ represents the variability around the grand mean (i.e., γ_4) of estimates of

the curvilinear component of treatment phase slopes for subjects. For all conditions and level 1 error specifications, estimates of $T_{\beta 4}$ were found to be biased. Magnitudes of relative bias of estimates ranged from 0.087 to 12.4.

Identification of patterns in relative bias of $T_{\beta 4}$. As for $T_{\beta 3}$, many relative bias statistics either equaled or approximated -1. It appeared that estimates of $T_{\beta 4}$ equaled 0 more often as the number of data points increased. As for $T_{\beta 3}$, this pattern was likely due, in part, to the increase in floor effects that occurred as numbers of data points per phase rose. Whether variance was continuous or discontinuous also appeared to be associated with the frequency at which estimates of $T_{\beta 4}$ equaled 0, when either different variance components or autoregressive covariance structures were specified at level 1. When variance was discontinuous, relative bias approximated or equaled -1 more often than when variance was continuous. In contrast, when variance was discontinuous, autocorrelation was generated, and separate error terms were specified at level 1, relative bias decreased as the number of data points per phase increased. Extrapolation of this pattern suggests that acceptable levels of bias may be obtained when roughly 40 to 50 data points are included in treatment phases.

Results for σ^2_{single}. Table 18 presents relative bias statistics for estimates of σ^2_{single}. The variance component σ^2_{single} represents the variability of subjects' actual data points around their expected values in the model. In other words, σ^2_{single} is the variance of residuals in level 1 models. The variance component σ^2_{single} was only estimated when different variance components or autoregressive covariance structures were specified at level 1. Estimates of σ^2_{single} were found to be of acceptable levels of bias in 16 instances: when variance was continuous and either (a) autocorrelation was not generated, 5 data points were generated per phase, and either different variance components or autoregressive covariance structures were specified, or (b) autocorrelation was generated, 10 or 20 data points were generated per phase, and an

autoregressive covariance structure was specified. Magnitudes of relative bias in other instances ranged from 0.071 to 0.54.

Identification of patterns in relative bias of σ^2_{single}. As would be expected, estimates of σ^2_{single} were only of acceptable levels of bias when data were generated to have a single, continuous variance. When autocorrelation was not generated, relative bias levels increased in the negative direction as the numbers of data points per phase increased. This pattern probably resulted from the manner in which floor levels were imposed on generated data, which likely reduced variance by restricting the range of treatment phase data points' y-values. In contrast, an opposite pattern was found in conditions in which autocorrelation was generated. In these instances, relative bias levels appeared to decrease as the numbers of data points increased.

Results for $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$. Table 19 presents relative bias statistics for estimates of $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$. The variance components $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$ represent the variability of subjects' baseline and treatment phase data points, respectively, around their expected values in the model. In other words, $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$ are the variances of residuals from baseline and treatment phases, respectively, in level 1 models. The variance components $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$ were only estimated when separate error terms were specified at level 1. Estimates of the two variance components were only observed to be of acceptable levels of bias when autocorrelation was not generated. For $\sigma^2_{baseline}$, estimates were of acceptable levels of bias in all instances when both variance was continuous and discontinuous. (One potential exception involved a relative bias statistic of -0.051. For this instance, the relative bias was regarded as acceptable.) For $\sigma^2_{treatment}$, estimates were of acceptable levels of bias only when variance was discontinuous and 5 data points were generated per phase. (Again, one potential exception involved a relative bias statistic of -0.051. As before, for this instance, the relative bias was regarded as acceptable.) Relative bias statistics for estimates of $\sigma^2_{treatment}$ in other instances, when autocorrelation was not

generated, ranged from -0.083 to -0.229. When autocorrelation was generated, relative bias for $\sigma^2_{baseline}$ ranged from -0.274 to -0.453, and ranged from -0.481 to -0.632 for $\sigma^2_{treatment}$.

Identification of patterns in relative bias of $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$. Levels of relative bias in $\sigma^2_{baseline}$ appeared to be associated with the levels of autocorrelation. As stated above, when autocorrelation was not generated, estimates of $\sigma^2_{baseline}$ were of acceptable levels of bias. On the other hand, when autocorrelation was generated, estimates of $\sigma^2_{baseline}$ were biased. Levels of relative bias in $\sigma^2_{treatment}$ appeared to be associated with the levels of autocorrelation, continuity of variance, and numbers of data points per phase. When autocorrelation was generated and/or variance was continuous, estimates of $\sigma^2_{treatment}$ were consistently biased. However, when autocorrelation was not generated and variance was discontinuous, levels of relative bias increased as the number of data points per phase increased. This pattern, and the high levels of bias in estimates of $\sigma^2_{treatment}$, were likely due to imposition of floor levels on generated data, which probably reduced treatment phase variance by restricting the range of treatment phase data points' y-values.

Relative Bias of Autocorrelation Estimates

Results for $\rho_{ar(1)}$. Table 20 presents relative bias statistics for estimates of the level of autocorrelation in simulated samples by condition. Levels of relative bias were only acceptable when autocorrelation was not generated, variance was continuous, and either 5 or 10 data points were generated for each phase. In all other conditions, relative bias ranged from 0.055 to 0.719.

Identification of patterns in relative bias of $\rho_{ar(1)}$. Levels of relative bias in estimates of $\rho_{ar(1)}$ appear to be associated with levels of continuity of variance, autocorrelation, and numbers of data points per phase. As stated above, estimates of $\rho_{ar(1)}$ were only of acceptable levels of bias when autocorrelation was not generated and variance was continuous. Across levels of continuity of variance and autocorrelation, the relative bias of estimates increased as the numbers of data

points per phase increased. This pattern likely resulted from the relationship between floor effects and number of data points per phase. As the number of data points per phase increased and floor effects became more frequently present, data points assumedly possessed more similar y-values (i.e., all near 0). Thus, when data points per phase were more numerous, the level of autocorrelation was probably inflated by floor effects.

CHAPTER 5

Discussion

This study was driven by several research questions pertaining to the use of MLM in meta-analysis of single-subject research data. These questions prompted examination of rates of convergence of analyses, levels of power and relative bias for fixed effects, and levels of relative bias for random effects and autocorrelation estimates, across (a) specifications for model errors at level 1, (b) numbers of data points per experimental phase, (c) numbers of participants per study, (d) numbers of studies meta-analyzed, (e) degrees of autocorrelation in individuals' data, and (f) continuity of level 1 variance across phases.

This study was inspired by and constitutes an effort to build upon previous theoretical and empirical work regarding the use of statistical analyses with single subject data (e.g., Beretvas & Chung, 2003b; Ferron et al., 2010; Jenson et al., 2007; Van den Noortgate, 2003b; Van den Noortgate & Onghena, 2011). The literature search contained in Chapter 2 identified a lack of empirical knowledge on MLM and single-subject data regarding trends within phases, use of quadratic models at level 1, the impact of discontinuous variances across phases, the relative benefit of different level 1 error specifications, and the accuracy of autocorrelation estimates. Also, the search showed little is known about the use of 3 level meta-analytic models with single-subject data and the impact of autocorrelation. The research questions which guided this study were formulated to address these shortcomings of knowledge.

This chapter presents the strengths and limitations of the simulation study, commentary on cut-offs for acceptable levels of relative bias, implications of study findings for meta-analysts, implications of study findings for SSED primary researchers, and directions for future research.

Strengths of the Simulation Study

The main strength of the study is its sourcing of values for generating parameters from a representative sample of single-subject data. By referencing outcomes of analyses of the representative sample when selecting generating parameter values, the simulated data was assured to possess characteristics of actual, published data. Most notably, the simulated data possessed high levels of variance within subjects' data sets, various degrees of trend, level change, and trend plus level change, floor effects, and a larger proportion of variance within studies than between.

Another strength of the study is the large number of factors and level 1 error specifications included in the design. Properties of single-subject data sets can differ along many variables. By manipulating five factors in the generation of data and analyzing samples with various level 1 error specifications, study results map a fairly wide swath of parameter space and generalize to a fairly large number of data scenarios.

Limitations of the Simulation Study

Unfortunately, as with any study, the results are associated with a number of limitations. Collectively, the limitations cast some doubt on the accuracy and stability of certain findings, as well as constrain generalization of the results.

The problem of floor effects. Unexpected patterns in relative bias of fixed effects were observed for estimates of γ_3 and γ_4, as well as for conditions with 5 data points per phase. Across all conditions, relative bias was positive for estimates of γ_3 and negative for estimates of γ_4. This pattern indicates that the magnitudes of both γ_3 and γ_4 were consistently overestimated. For conditions with 5 data points per phase, relative bias statistics for the treatment phase effects (i.e., γ_2, γ_3, and γ_4) were consistently lower than for conditions with either 10 or 20 data points per phase. However, relative bias was also consistently lower for conditions with 20 data points per

phase than those with 10. Given the trends in relative bias in conditions with 10 and 20 data points per phase, relative bias in conditions with 5 data points per phase was expected to be highest and qualify most frequently as unacceptably biased.

These unexpected patterns may result from the manner in which data were generated. As can be seen above in Figures 4 through 6, the y-values of simulated data for treatment phases with 5 data points dropped to 0 less frequently than treatment phase y-values of simulated data with 10 or 20 data points per phase. The short duration of treatment phases with 5 data points likely did not provide enough distance along the x-axis for generating values for treatment phase slopes to cause y-values to reach 0. This apparent relationship between the number of data points and the presence of floor effects (i.e., when generated data is forced to take on the floor level of 0 as opposed to the level initially specified by generating equations) represents a confound. The improved relative bias of conditions with 5 data points per phase, as well as the degree to which γ_3 and γ_4 were overestimated, may result from this systematic variability in the presence of floor effects across levels of numbers of data points per phase. With fewer floor effects present in conditions with 5 data points per phase, the shape of data trajectories in these conditions may have been more purely representative of the population generating equations. Consequently, the estimates of models' fixed effects would be as observed: less intensely biased and less often characterized by unacceptable levels of bias. Also, patterns in the overestimation of γ_3 and γ_4 may be functions of the confounding relationship.

This relationship between number of data points per phase and frequency of floor effects was assessed in a follow-up analysis. Table 21 presents the percentages at which near-zero y-values were observed in treatment phases of data sets with 5, 10, and 20 data points per phase, after truncation of y-values. The sample analyzed contained generated 3000 data sets, with 1000

sets for each number of data points per phase. The samples of 1000 were composed of 200 simulated studies with 5 participants each.

As can be seen in Table 21, the percentages of near-zero y-values after truncation did, in fact, vary substantially across numbers of data points per phase in a near linear fashion. All percentages of near-zero data points were greatest when 20 data points were generated per phase, and lowest when 5 data points were generated for each phase.

As a result of imposing a floor level at 0, treatment phase data generally followed different trajectories at each level of number of data points per phase, despite constant use of a single generating equation. In other words, the floor effects created differences in the actual population averages of treatment effect measures across numbers of data points per phase. These differences represent a confound to study findings and prevent identification of meaningful patterns across levels of number of data points per phase. Additionally, they introduced bias into relative bias statistics for fixed and random effects pertaining to treatment phase data. The presence of floor effects in treatment data explains, in part, patterns of bias in estimates of γ_3, γ_4, $T_{\beta 3}$, $T_{\beta 4}$, $\sigma^2_{treatment}$, and $\rho_{ar(1)}$.

While the rate of near-zero y-values in data sets with 10 measures per phase is as intended (see Figure 2), the rates are deflated and inflated for data sets with 5 and 20 measures per phase, respectively. Thus, findings for conditions with 10 data points per phase are unaffected by the confounding relationship. To make full use of findings for conditions with 5 and 20 data points per phase, the study design would need to be expanded to include the percentage of near-zero y-values in treatment phases as an additional factor. The factor would have 3 levels: those observed and reported in Table 21 for data sets with 5, 10, and 20 data points per phase. Such an expanded design would involve varying generating values for treatment phase slope parameters across the levels of percentage of near-zero y-values in treatment phases. Alternatively, the study

could be re-run with either no truncation of treatment phase y-values or an increase in the generating value for the baseline intercept, such that initially generated treatment phase y-values rarely or never drop below zero.

The problem of high baseline y-values. As mentioned in Chapter 4, when autocorrelation was not generated, as the number of data points increased, relative bias levels also increased. Assumedly, this trend resulted from a confounding relationship between the number of data points per phase and behavior levels generated at the end of baseline. As the number of data points increased, the range of behavior levels generated at the end of baseline phases was likely to be larger, due to the increased distance along the x-axis and opportunities for y-values to rise. An increase in the range of y-values at the end of baseline phases would lead to an increase in the range of y-values of ends of baseline regression lines. This increase would then cause an increase in the range of vertical distance between ends of baseline regression lines and treatment intercepts (i.e., the variance of immediate treatment effects). This relationship was also assessed in a follow-up analysis. Table 22 presents the percentages of extreme, high y-values observed in baseline phases with 5, 10, and 20 data points. The findings were obtained from the same sample described above.

As can be seen in Table 22, the percentages of extremely high y-values in baseline phases did vary across numbers of data point per phase. However, the differences are slight. For the analysis, cut-off values of 80 and 90 were chosen due to the impression taken from graphs of data (i.e., Figures 4 through 6) that the vast majority of baseline y-values were roughly less than 80. The cut-off points do not represent problematically high values, but rather allow a contrast in distributions of y-values across numbers of data points per phase.

While it did appear that relative bias in estimates of $T_{\beta 2}$ covaried with the number of data points per phase, the differences in percentages of extremely high y-values may or may not have

been large enough to introduce systematic bias in estimates of fixed and random effects. To be certain of the import of these differences, the percentages of extremely high baseline y-values could be paired with the percentages of near-zero treatment phase y-values to create a combined factor, which would then be added to the study design. However, the differences seen in Table 22 do not strongly suggest that differences across numbers of data points per phase were great enough to confound results.

The problem of unstable relative bias statistics. On occasion, individual or small groupings of relative bias statistics for fixed and random effects did not conform to overarching patterns in values. The deviance of these statistics occasionally involved acceptable levels of bias when overarching patterns implied all estimates for certain conditions should be biased and vice versa. Curiosity about these deviant statistics prompted a follow-up analysis to assess the stability of estimates of relative bias.

Data for 4 conditions (i.e., those with continuous variance, no autocorrelation, and 5 data points per phase) were generated and analyzed in a second, limited run of the SAS simulation program. As in the first run of the simulation program, 400 samples were generated and analyzed per condition. Relative bias statistics for the fixed effects and $T_{\beta 0}$ obtained from the second analysis were compared to those of the first analysis. The variance component $T_{\beta 0}$ was chosen for comparison due to the relatively high frequency at which its relative bias was less than 0.05. Tables 23 and 24 present relative bias statistics, for $T_{\beta 0}$ and fixed effects, respectively, obtained from each run of the simulation program.

For both $T_{\beta 0}$ and the fixed effects, relative bias statistics were found to deviate substantially. In the first run of the 4 conditions, 5 instances of acceptable bias were found for $T_{\beta 0}$ (for which magnitudes ranged 0.34 to 0.50). In the remaining 7 instances, magnitudes of relative bias ranged from 0.080 to 0.163. However, in the second run, no instances of acceptable bias

were observed. Here, magnitudes of relative bias for the 12 figures ranged from 0.253 to 0.385. Results for the fixed effects were less differentiated, although important differences were found. For 3 instances in which relative biases obtained from the first run were of acceptable levels (magnitudes ranged from 0.027 to 0.028), unacceptable levels of bias were instead found in the second run (magnitudes ranged from 0.248 to 0.249). For γ_0, γ_1, γ_2, γ_3, and γ_4, respectively, relative bias from the second run deviated from the first by an average magnitude of 0.002, 0.105, 0.006, 0.027, and 0.306.

Together, these follow-up findings suggest that 400 replications of each condition were insufficient to obtain stable estimates of relative bias for fixed and random effects. Further, accuracies of the relative bias statistics reported in Chapter 4 are in question.

Other limitations associated with data generation. Several features of the data generated for this study limit generalization of study findings. Results only generalize to data conditions which match those of the study. In particular, results are only relevant when (a) data points per phase number between 5 and 20, (b) subjects per study number between 3 and 6, (c) studies meta-analyzed number between 10 and 30, (d) variance is continuous and equals approximately 150 or variance is discontinuous and equals approximately 300 in baseline and 70 in treatment phases, (e) lag-1 autocorrelation in the sample equals approximately 0.0 or 0.4, (f) samples contain only 1 baseline and treatment phase pair, (g) subjects' data are all on the same metric (e.g., percentage of 10 second intervals) and (h) magnitudes of effects in data (e.g., baseline intercepts and slopes, treatment intercepts and slopes) approximate those used to generate data. With regard to stipulation (h), three points deserve mention. For one, in the population generating model, the effect of treatment was both immediate and gradual. On occasion, data sets were generated in which treatment effects appeared to be either primarily gradual or primarily immediate (see Figures 4 through 6). Therefore, results of this study

generalize to samples which include data sets marked by a mix of immediate and gradual treatment effects, but only when the effects of treatment are on average both immediate and gradual. Second, baseline trends were generated to be linear. Findings of this study do not generalize to samples which contain baseline data marked by more complex patterns, such as curvilinear trends. Third, extinction spikes were not modeled with generating equations. Thus, results do not generalize to meta-analysis of data containing extinction spikes.

Limitations due to models and analyses. Several features of the models used to meta-analyze samples and the analysis conducted create further limitations. Results only generalize to conditions in which (a) the 3 level model expressed in Equations 29 through 39 in Chapter 3 is employed, and (b) different variance components, an autoregressive covariance structure, or separate error terms for each phase are specified for level 1 error. With regard to stipulation (a), two points deserve mention. For one, predictor variables were not included in level 2 or 3 equations. Consequently, study findings do not comment on moderator and mediation analyses. Second, the frequent estimates of 0 for the variance of level 2 random effects indicates certain of these level 2 random effects should not always or, in some instances, never be included in models for data with characteristics similar to those of the generated data. However, study findings do not generalize to use of models with one or more level 2 random effects excluded. Finally, the meta-analyses performed in this study did not involve statistical tests of random effects. Therefore, study findings do not comment on the performance of statistical tests of random effects.

Additionally, use of analysis models that are different than generating models can result in bias in estimates of variance components (Kwok, West, & Green, 2007). Much of the bias observed in variance components could have been due to use of a simplification model in analyses. As mentioned in Chapter 3, the simplification model used did not include several level 3 random effects which were included in the generating model.

A Challenge to the Face Validity of the Standard Cut-off Point for Acceptable Relative Bias

Figure 7 graphically depicts the population average model, two models at the limits of acceptable bias, and an extremely biased model observed in the simulation study. The graph is presented as a challenge to the face validity of the standard cut-off point of 0.05 for acceptable relative parameter bias for fixed effects. The solid black line in the graph represents the population average model. The blue line with diamond dots represents one limit of acceptable relative bias, when the magnitude of relative bias equals 0.05 for all fixed effects and the direction of bias is consistent with the biases observed in the simulation study. The aqua line with diamond dots represents another limit of acceptable bias, when the magnitude of relative bias equals 0.05 for all fixed effects and the direction of bias is opposite the bias observed in the simulation study. Finally, the red line with square dots represents a model observed in the simulation study with extremely biased values for γ_3 and γ_4. Parameter values for this model were determined by solving for averages in relative bias formulas (i.e., $\bar{\hat{\theta}}$) after inputting observed relative bias statistics. In the condition depicted, variance was discontinuous, autocorrelation was generated, the numbers of data points per phase, subjects per study, and studies meta-analyzed were 10, 6, and 30, respectively, and an autoregressive covariance structure was specified. The relative biases of the fixed effects γ_0 through γ_4 were, respectively, -0.002, 0.007, -0.014, 0.195, and -3.74.

As can be seen in the blue line in Figure 7, models composed of acceptably biased parameter values can misrepresent data phenomena greatly. In contrast, the red line shows how models composed of greatly biased parameter values can approximate data phenomena with a moderate degree of fidelity (even when floor effects confound data generation).

Patterns in Figure 7 suggest that use in this study of a magnitude of 0.05 as the cut-off point for acceptable levels of relative bias may not have produced clear information on whether fixed effect estimates represent data phenomena well or poorly. Fixed effect estimates deemed biased, as well as those deemed of acceptable levels of bias, may or may not have depicted average data trajectories in samples accurately.

Implications of Findings for Meta-analysis of Single-subject Data

Study findings suggest a number of guidelines for the meta-analysis of single-subject data, when employing the model expressed in Equations 29 through 39. These guidelines are discussed below as they relate to fixed effects, random effects, and autocorrelation estimates. Following discussion of the guidelines, comments are offered on the potential for poor fit between single-subject designs and statistical analyses. In the appendix to this paper, readers will find several sets of SAS code which can be used to estimate this study's meta-analytic models.

Fixed effects. To ensure adequate power for statistical tests of models' fixed effects, meta-analysis should only be performed on samples in which (a) the number of data points per phase is 10 or greater and (b) the total number of subjects is 60 or greater. With regard to stipulation (a), it is possible that adequate power can be achieved when samples include data sets with fewer than 10 points per phase, but the average number of data points per phase in the sample is 10 or greater. However, the impact on power of different levels of variance in numbers of data points per phase within samples is not known.

Findings indicated that power for tests of small effects (i.e., γ_1 and γ_4) could be low, especially when numbers of level 1 and 2 units were small. When performing a meta-analysis, if effects which are expected to be small are found to be statistically insignificant, researchers should consider retaining parameters for the effects in the model and disregarding statistical test results for the fixed effects. This practice should be followed especially when inclusion of the

parameters could improve accuracy of estimates for other model parameters (e.g., as inclusion of parameters for the baseline slope aids estimation of the immediate treatment effect).

To ensure that the relative bias of estimates for model intercepts (i.e., γ_0 and γ_2) is within the acceptable range, meta-analysts only need to follow the above mentioned guidelines for obtaining adequate power. However, to ensure that the relative bias of estimates for model slopes (i.e., γ_1, γ_2, and γ_4) is within the acceptable range, it appears meta-analysis should be limited to samples in which the number of data points per phase is somewhere between 30 and 50, or greater. This stipulation should be received with caution, given the confound of floor effects may have introduced bias into otherwise unbiased estimates (i.e., unbiased estimates of model slopes may be able to be obtained with fewer than 30 data points per phase).

Should autocorrelation be present in samples, autoregressive covariance structures should be specified for level 1 error to improve the relative bias of estimates of γ_3. However, it should be recognized that this practice may be associated with increased bias in estimates of γ_1. (For guidelines on how to assess the level of autocorrelation in samples, see the section below on autocorrelation estimates.)

When variance is discontinuous across phases, meta-analysts should be aware that estimates of γ_1 and γ_3 may be biased. (However, relative bias rates for estimates of γ_3 may have been confounded by the floor effects present in generated data.) Unfortunately, no levels of study factors were associated with attenuation of these biases and thus, no practice can be recommended to reduce the bias in γ_1 and γ_3.

Random effects. When employing the 3 level model examined in this study, meta-analysts are likely to find that not all level 2 random effects deserve inclusion. In particular, the

random effects r_{1jk}, r_{3jk}, and r_{4jk} are likely to be estimated to be 0 and thus warrant exclusion from the model.

For most conditions examined in this study, estimates of the variance of random effects were biased. Presumably, the large degree of sampling error generated in data and commonly observed in published single subject data, both within and across subjects, is responsible for the inaccuracy of estimates. To attenuate this bias, several efforts can be made. However, the following practices may not reduce bias to acceptable levels (i.e., relative bias ≤ 0.05).

When autocorrelation is found to exist in a sample, specification of autoregressive covariance structures at level 1 can help reduce the bias in estimates of $T_{\gamma 0}$, $T_{\beta 0}$, and $T_{\beta 2}$. Unfortunately though, when autocorrelation is found to exist, no practice can be recommended to reduce bias in $T_{\beta 1}$ or, when separate error terms are specified for each phase at level 1, in $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$.

When variance is found to be discontinuous across phases, specification of separate error terms for each phase at level 1 can reduce bias in estimates of $T_{\beta 3}$ and $T_{\beta 4}$. However, discontinuous variance is associated with bias in $T_{\beta 0}$ and no practice can be recommended to improve the bias of this estimate.

Increasing the numbers of units at levels 1, 2, and 3 additionally appears to help attenuate bias. Limiting samples to data sets with 40 to 50 data points per phase, or greater, may help reduce bias to acceptable levels in $T_{\beta 2}$ (i.e., when autoregressive covariance structures are specified at level 1) and $T_{\beta 4}$ (i.e., when separate error terms for each phase are specified at level 1). Also, collecting larger samples of subjects and studies can lower relative bias rates in $T_{\gamma 0}$ and $T_{\beta 0}$.

Autocorrelation estimates. Findings from this study indicate that estimates of autocorrelation resulting from SAS PROC MIXED and specification of an autoregressive covariance structure are frequently overestimated and biased. Assessing the presence of autocorrelation in samples is necessary prior to conducting a multilevel meta-analysis of single-subject data. As discussed above, when autocorrelation is found to be present, various practices should be undertaken and certain estimates should be regarded as biased. Findings in this study indicate that autocorrelation is likely to be absent from data sets when estimates of autocorrelation are less than 0.1. Should estimates of greater than 0.1 be obtained, meta-analysts can assume autocorrelation does exist and proceed with their analysis accordingly. Given the consistency with which autocorrelation levels were overestimated and the large degree to which autocorrelation was overestimated when it was generated, meta-analysts can be confident that estimates of less than 0.1 indicate autocorrelation is not present in data sets. However, further research with additional levels of autocorrelation is necessary for confirmation of these assumptions. Alternatively, researchers could instead make use of the test developed by Riviello & Beretvas (2008) to check for the presence of autocorrelation in each data set separately.

Potential for poor fit between single-subject designs and statistical analyses. Single-subject research was developed for use with small, special populations (e.g., people with autism or intellectual disability) and difficult to safely study behaviors (e.g., self-injurious behavior; Kennedy, 2005). In a single geographical area, the number of potential research subjects who meet the criteria for inclusion in a study is often low, due to the small size of the overall population of interest. Also, when data collection puts subjects in harm's way, the number of data points that can be collected safely may be very few. As a result, single-subject researchers have been forced to accept small samples of subjects and data points, the likelihood of sampling error, and the potential for study results to not generalize to other members of the population. To

combat these limitations, single-subject researchers have developed a number of methods to convincingly demonstrate causal relationships between independent and dependent variables and make results relevant to other members of a population.

Statistical analyses have been developed for use with group design research. Statistical procedures rely on large sample sizes for minimization of sampling error and maximization of accuracy. Also, the procedures make use of averages and distributions to increase the precision and degree of nuance of claims made for results' generalization.

The standards for convincing evidence differ in single-subject research and statistical analyses. Whereas confirmation of a treatment effect can be accomplished in single-subject research with as few as two or three data points and/or subject, many more data points and subjects are required for confirmation via statistical analyses. Statistics and single-subject research potentially fit together poorly because of their differences in standards for convincing evidence. Single-subject researchers are able to satisfy their goals with just a few data points, but statistical analysts need more data points to do their job. Should the natural limitations of special populations and the behaviors commonly studied in single-subject research prevent the obtaining of data fit for statistical analysis, a number of outcomes are possible. For one, the findings of single-subject research may be left out of discussions of evidence-based practice (given the emphasis on widely generalizable information). Alternatively, the standards for evidence-based practice, as determined by single-subject research, may be set to be lower than those for group design research. In this sense, the increased potential for sampling error in single-subject research findings would be overlooked or accepted. Unfortunately, this outcome would involve occasional acceptance of biased information and, consequently, misguided decisions on educational policy. In contrast to these disappointing outcomes, researchers will hopefully develop methods of statistical analysis that perform adequately despite the small samples collected in single-subject

research. Several methods that have the potential to fit well with single-subject research (and better than the analyses examined in this study) are described below in Future Research Directions.

Implications of Findings for SSED Primary Research

While the focus of this study was meta-analysis, the findings have several implications for SSED primary research. As discussed above and in Chapter 4, levels of power and relative parameter bias are dependent on numbers of data points collected per phase and numbers of subjects included per study.

In order to minimize bias in meta-analysis results, SSED primary researchers are encouraged to make efforts to collect a minimum of 10 data points in baseline and treatment phases, and up to 40 or 50, if feasible. Given that behavior patterns in treatment phases are more complex and meaningful (e.g., apparent trends are true trends, as opposed to just variance which randomly appears as a trend), primary researchers are especially encouraged to emphasize extended data collection in treatment phases. Unfortunately, there are a number of ethical, practical, and internal validity problems that can result from extended data collection. For example, when studying self-injurious behavior, baseline conditions may place subjects in danger of physical harm and thus should be ended as quickly as possible. Also, data collection costs time and money. When the goals of a primary researcher are met, it may not seem worth additional time and money to collect more data points simply for the sake of improved parameter estimation in a meta-analysis that may or may not be conducted years in the future. As a final example, extended data collection may open findings to threats of maturation effects or measurement-induced behavior changes, and thus may compromise a study's internal validity. Should ethical, practical, or internal validity problems stem from extended data collection, primary researchers

are instead encouraged to make efforts to collect as many data points as safely possible in each phase, according to their best judgment.

To aid the securing of adequate levels of power and acceptable levels of bias, and facilitate the meta-analysis of narrow bodies of research, SSED primary researchers are encouraged to include at least 6 subjects per study, and more if feasible. As above, practical and internal validity problems may result from including relatively large numbers of subjects in studies (e.g., study costs, heterogeneity of subjects' characteristics). When faced with the threat of an insurmountable practical or internal validity problem, primary researchers are encouraged to include as many subjects as safely possible, according to their best judgment.

Also, given the need for similarity across all baseline-treatment phase pairs included in a meta-analysis, researchers who employ alternating treatment designs or gradation/fading of independent variables within treatment phases are encouraged, when feasible, to either begin or end a subjects' data set with a pair of baseline and treatment phases, in which the treatment given is the treatment of greatest interest in the study (and hopefully also to a meta-analyst).

Future Research Directions

Future research should address a variety of topics related to the meta-analysis of single-subject data. For one, the performance of additional level 1 models should be examined. Research should explore models with simple linear treatment phase trajectories, those that make use of log link functions (Beretvas & Wang, 2011), logistic models (Beretvas, 2011), and other functions whose inclusion of a denominator allow modeling of horizontal asymptotes. Given the tendency of the quadratic model examined in this study to predict negative y-values for data points, functions which can model horizontal asymptotes may perform better in various ways. Future research should also explore the meta-analysis of data sets composed of more than one baseline and/or treatment phase. For example, research could look at the meta-analysis of multiple reversal

designs, alternating treatment designs, or designs in which treatment components are gradually introduced across several phases. To examine the performance of MLM in the meta-analysis of multiple baseline studies, future research should examine outcomes associated with different degrees of variation in numbers of data points per phase within and across studies. Given the diversity of metrics used in data collection in SSED research, future studies should additionally explore when standardization of data on different metrics is necessary.

The limits of this study provide further directions for future research. Associations with power and relative parameter bias should be examined for additional levels of (a) number of data points per phase, (b) number of subjects per study, (c) number of studies meta-analyzed, (d) variance in each phase, (e) autocorrelation, and (f) magnitudes of effects/shapes and levels of data trajectories. Additionally, the percentages of near-zero y-values in treatment phases should be included as a factor in the designs of future studies. Also, future research should explore type I error, statistical tests of random effects, and moderator and mediation analyses.

Table 1

Summary of articles on the use of MLM with single-subject data

Study	Model specifications	Data extraction/ simulation	Data standardization method	Unit counts at each level	Treatment of autocorrelation	Analysis process
Bell et al. (2011)	Level 1 $Y_{ij} = \pi_{0j} + \pi_{1j}(\text{phase})_{ij} + \pi_{2jk}(\text{phase})_{ijk}(\text{time in treatment})_{ijk} + e_{ijk}$ Level 2 $\pi_{0j} = \beta_{00} + r_{0j}$ $\pi_{1j} = \beta_{10} + r_{1j}$ $\pi_{2j} = \beta_{20} + r_{1j}$ Notes Acknowledges options of using models which model more complex growth trajectories	Using Monte Carlo methods, generated data to fit the equations at left, and simulated one baseline and one treatment phase (i.e., AB) per subject Simulated data for 810 conditions which varied on the following factors: effects sizes of level and slope changes (0.5, 1.0, 1.25, 1.5, 1.75, or 2), total data points per subject (10, 20, or 30), numbers of subjects per study (3, 4, 8, 16, 32), variance among subjects in initial level (.2, 1, 2), and auto-correlation (0, 0.2, 0.4)	Data standardization was not necessary due to study's focus on the analysis of results from a single study	Provides data that shows (a) large effect sizes (≥ 1.75) are necessary for adequate power in tests of changes in level, across sizes of study samples, and (b) adequate power in tests of changes in slope can be achieved when effect sizes are large or study samples are large, or when both are large Data points per subject not mentioned in report of results	Included level of auto-correlation as a factor in the study design, but did not mention autocorrelation in report of study results	Compared statistical power for fixed effects, across 3 methods of approximating degrees of freedom Aggregated results across levels of data points per subject, variance in initial level, and autocorrelation Found power rates were not substantially different across methods of approximating degrees of freedom; also, differences became more slight as study sample sizes increased

Table 1 (continued)

Summary of articles on the use of MLM with single-subject data

Study	Model specifications	Data extraction/ simulation	Data standardization method	Unit counts at each level	Treatment of auto-correlation	Analysis process
Beretvas (2011)	Level 1 (a) $Y_{ij} = (1 - \text{phase})_{ij}(\pi_{0j}) + (\text{phase})_{ij}\left(\alpha_1 + \frac{\alpha_2 - \alpha_1}{\{1+(\pi_{1j})\exp[-(\pi_{2j})(\text{time in treatment}_{ij})]\}} + (\pi_{0j})\right) + e_{ij}$ (b) $Y_{ij} = (1 - \text{phase})_{ij}[\pi_{0j} + \pi_{1j}(\text{time})_{ij}] + (\text{phase})_{ij}\left(\alpha_1 + \frac{\alpha_2 - \alpha_1}{\{1+(\pi_{2j})\exp[-(\pi_{3j})(\text{time in treatment}_{ij})]\}} + [\pi_{0j} + \pi_{1j}(\text{time in treatment})_{ij}]\right) + e_{ij}$ Level 2 $\pi_{pj} = \beta_p + r_{pj}$ Notes Model (a) recommended when baseline data is not trended Model (b) recommended when baseline data is trended	NA; paper only introduces potential uses of models	NA	NA	NA	NA
Beretvas & Wang (2011)	Level 1 $\text{Log}(Y_{ij}) = \pi_{0j} + \pi_{1j}(\text{time})_{ij} + \pi_{2j}(\text{phase})_{ij} + \pi_{3j}(\text{phase})_{ij}(\text{time in treatment})_{ij} + \pi_{4j}(\text{setting2})_j + \pi_{5j}(\text{setting3})_j + \pi_{6j}(\text{setting2})_j(\text{phase})_{ij} + \pi_{7j}(\text{setting3})_j(\text{phase})_{ij} + \pi_{8j}(\text{setting2})_j(\text{time in treatment})_{ij} + \pi_{9j}(\text{setting3})_j(\text{time in treatment})_{ij}$ Level 2 $\pi_{0j} = \beta_0 + r_{0j}$ $\pi_{1j} = \beta_1 + r_{1j}$ for π_{3j} through π_{9j}: $\pi_{pj} = \beta_p$ Notes Level 1 equation recommended when analyzing data from 3 settings for each subject Parameters π_{4j} through π_{9j}, and the associated terms, should be dropped when analyzing data from 1 setting	Using Monte Carlo methods, generated multiple baseline data, to fit the equations at left, for 3 settings and 5 dependent variables per subject Simulated data for 24 conditions which varied in number of subjects (3 or 5), total data points per subject (15 or 30), and data patterns (6 scenarios)	Data standardization was not necessary due to study's focus on the analysis of results from a single study	Relative bias of fixed effect estimates was less when data points numbered 30 than 15; no systematic differences noted between numbers of subjects Standard errors found to be fairly consistently biased in all conditions	Auto-correlation was not generated in the study	Assessed relative bias of estimates of fixed effects and their standard errors

Table 1 (continued)

Summary of articles on the use of MLM with single-subject data

Study	Model specifications	Data extraction/ simulation	Data standardization method	Unit counts at each level	Treatment of auto-correlation	Analysis process
Ferron et al. (2009)	Level 1 $Y_{ij} = \pi_{0j} + \pi_{1j}(phase)_{ij} + e_{ij}$ Level 2 $\pi_{0j} = \beta_{00} + r_{0j}$ $\pi_{1j} = \beta_{10} + r_{1j}$	Using Monte Carlo methods, generated multiple baseline data to fit the equation at left, for each simulated subject Simulated data for 180 conditions which varied in number of subjects per study (4, 6, or 8), total data points per subject (10, 20, or 30), autocorrelation (0, 0.1, 0.2, 0.3, 0.4), variance among subjects in initial level (0.1 or 0.3) and treatment effect (0.1 or 0.3)	Data standardization was not necessary due to study's focus on the analysis of results from a single study	Increases in numbers of subjects found to be associated with (a) decreases in differences in accuracy among methods of approximating degrees of freedom, and (b) decreases in relative bias in variance components when autoregressive covariance structures are specified (however, relative bias remained above acceptable levels)	Found auto-correlation has potential to bias estimates; when autoregressive covariance structures were specified, (a) the Kenward-Rogers and Satterthwaite methods were found to maintain accurate confidence intervals, and (b) relative biases of variance components were still greater than acceptable levels	Assessed accuracy of confidence intervals for estimates of treatment effects and relative bias of variance components, across 5 methods for approximating degrees of freedom and 2 specifications for level 1 error
Ferron, Farmer, & Owens (2010)	Level 1 $Y_{ij} = \pi_{0j} + \pi_{1j}(phase)_{ij} + e_{ij}$ Level 2 $\pi_{0j} = \beta_{00} + r_{0j}$ $\pi_{1j} = \beta_{10} + r_{1j}$	Using Monte Carlo methods, generated multiple baseline data to fit the equation at left, for each simulated subject Simulated data for 180 conditions which varied in number of subjects per study (4, 6, or 8), total data points per subject (10, 20, or 30), autocorrelation (0, 0.1, 0.2, 0.3, 0.4), variance among subjects in initial level (0.1 or 0.5) and treatment effect (0.1 or 0.5)	Data standardization was not necessary due to study's focus on the analysis of results from a single study	Both OLS and Empirical Bayes estimates improved in accuracy as numbers of data points increased Kenward-Rogers method found to maintain accuracy of confidence intervals across number of data points Numbers of subjects per studies not found to relate to accuracy of estimates or confidence intervals	Bias in estimates not found to be associated with autocorrelation levels While some associations were found between autocorrelation levels and confidence interval accuracy, the Kenward-Rogers method maintained accurate intervals across levels of autocorrelation	Assessed bias of treatment effect estimates, for 2 methods of estimation, and bias of confidence intervals, across 3 methods of approximating degrees of freedom Found the OLS method produced more accurate estimates than Empirical Bayes; also, only the Kenward-Rogers method provided accurate confidence intervals

Table 1 (continued)

Summary of articles on the use of MLM with single-subject data

Study	Model specifications	Data extraction/ simulation	Data standardization method	Unit counts at each level	Treatment of autocorrelation	Analysis process
Jenson et al. (2007)	Level 1 $Y_{ij} = \pi_{0j} + \pi_{1j}(\text{phase})_{ij} + e_{ij}$ Level 2 $\pi_{0j} = \beta_{00} + r_{0j}$ $\pi_{1j} = \beta_{10} + r_{1j}$	Using Monte Carlo methods, generated data to fit the equation at left, simulating one baseline and one treatment phase for each subject Simulated data for 30 conditions which varied in number of subjects included in meta-analysis (15, 40, or 80), data points in baseline and treatment phases (5/10 or 10/20), type of autocorrelation (moving average or autoregressive), and level of autocorrelation (0, 0.4 or 0.8); study-level dependency not generated in data	Need for standardization not addressed	Provides data that reveals (a) small subject sample sizes (≤ 40) can interact with high auto-correlation to decrease power, (b) small samples of data points ($n \leq 15$) are associated with low power, except when subject samples are large, and (c) type I error rate equals the nominal rate across unit counts	Provides data that reveals (a) high autocorrelation (.80) can interact with small sample sizes to decrease power to inadequate levels and (b) low to moderate autocorrelation does not pose a threat to power or type I error	Assessed type I and II error rates across conditions
Nugent (1996)	Level 1 $Y_{ij} = \pi_{0j} + \pi_{1j}(\text{time})_{ij} + \pi_{2j}(\text{time})_{ij}^2 + \ldots + \pi_{pj}(\text{time})_{ij}^p + e_{ij}$ Level 2 $\pi_{pj} = \beta_{p0} + \beta_{p1}(X_{1j}) + \ldots + \beta_{pq}(X_{qj}) + r_{pj}$	Extract and model data from treatment phases only; guidelines for how to combine data from multiple treatment phases not specified	Need for standardization not addressed	Unit counts not addressed	Assess level of autocorrelation When significant level exists, specify a more complex model (i.e., quadratic as opposed to mean change)	Recommendations: Assess unconditional model fit; consider alternative models Estimate unconditional model before adding predictor variables Statistically test fixed and random effects Complement statistical analysis with visual analysis

Table 1 (continued)

Summary of articles on the use of MLM with single-subject data

Study	Model specifications	Data extraction	Data standardization method	Unit counts at each level	Treatment of autocorrelation	Analysis process
Van den Noortgate & Onghena (2003a)	Level 1 (a) $d_j = \delta_j + e_j$ (b) $\hat{\pi}_{1j} = \pi_{1j} + e_{1j}$ $\hat{\pi}_{3j} = \pi_{3j} + e_{3j}$ (c) $Y_{ij} = \pi_{0j} + \pi_{1j}(\text{phase})_{ij} + \pi_{2j}(\text{time})_{ij} + \pi_{3j}(\text{phase})_{ij}(\text{time in treatment})_{ij} + e_{ij}$ Level 2 (a) $\delta_j = \beta_0 + \beta_1(X_1)_j + \ldots + \beta_p(X_p)_j + r_j$ (b) $\pi_{pj} = \beta_{p0j} + \beta_{p1j}(X_1)_j + \ldots + \beta_{pqj}(X_q)_j + r_{pj}$ (c) $\pi_{pj} = \beta_{p0j} + \beta_{p1j}(X_1)_j + \ldots + \beta_{pqj}(X_q)_j + r_{pj}$ Notes Model (a) recommended when data contains no trend Model (b) recommended when data involves different metrics for dependent variables; level 1 outcomes derive from OLS regression analyses Model (c) recommended when data involves consistent dependent variable metrics	Extract data from either 1st baseline and 1st treatment phases (i.e., AB) or multiple baseline and treatment phases (e.g., ABAB, ABAC) When extracting data from multiple baseline and treatment phases, expand model to include parameters and dummy variables for each phase pair, for example: $Y_{ij} = \pi_{0j}(\text{pair 1})_{ij} + \pi_{1j}(\text{phase})_{ij}(\text{pair 1})_{ij} + \pi_{2j}(\text{pair 2})_{ij} + \pi_{3j}(\text{phase})_{ij}(\text{pair 2})_{ij} + e_{ij}$	Convert time variables to same scale For dependent variables: calculate OLS regression coefficients for each data set and divide coefficients by RMSE and the corresponding covariance matrices by MSE; use standardized coefficients with model (b)	Unit counts not addressed	Assess level of autocorrelation When significant level exists, specify an autoregressive covariance structure	Recommendations: Estimate unconditional model before adding predictor variables Statistically test fixed and random effects For multiple dependent variables, use multivariate MLM methods

Table 1 (continued)

Summary of articles on the use of MLM with single-subject data

Study	Model specifications	Data extraction	Data standardization method	Unit counts at each level	Treatment of autocorrelation	Analysis process
Van den Noortgate & Onghena (2003b)	Level 1 (a) $Y_{ij} = \pi_{0j} + \pi_{1j}(\text{phase})_{ij} + e_{ij}$ (b) $Y_{ij} = \pi_{0j} + \pi_{1j}(\text{phase})_{ij} + e_{1ij}(\text{phase1})_{ij} + e_{2ij}(\text{phase2})_{ij}$ Level 2 $\pi_{pj} = \beta_{p0j} + \beta_{p1j}(X_1)_j + \ldots + \beta_{pqj}(X_q)_j + r_{pj}$ Notes Model (b) recommended when variability patterns differ across phases When data contains linear or curvilinear trends, authors recommend expansion of models to include linear or polynomial terms	Extract data from either 1st baseline and 1st treatment phases (i.e., AB) or multiple baseline and treatment phases (e.g., ABAB, ABAC) When extracting data from multiple baseline and treatment phases, either (a) aggregate phase data when patterns are similar across baseline or treatment phases, or (b) expand model to include parameters and dummy variables for each phase pair, as illustrated above	Either (a) subtract baseline mean from all scores and then divide by within-phase SE, or (b) calculate OLS regression coefficients for each data set and divide coefficients by RMSE and the corresponding covariance matrices by MSE (authors refer readers to 2003a paper for guidelines on how to model standardized coefficients)	Include 30+ units at each level	Assess level of autocorrelation When significant level exists, specify an autoregressive covariance structure	Recommendations: Check assumptions of normality and independence Estimate unconditional model before adding predictor variables Statistically test fixed and random effects, and/or compare model fit with and without random effects For multiple dependent variables, use multivariate MLM methods
Van den Noortgate & Onghena (2007)	Level 1 $Y_{ij} = \pi_{0j} + \pi_{1j}(\text{phase}\,2)_{ij} + \pi_{2j}(\text{phase}\,3)_{ij} + e_{ij}$ Level 2 $\pi_{pj} = \beta_{p0j} + \beta_{p1j}(X_1)_j + r_{pj}$ Notes Authors briefly mention several model variations and refer readers to their previous papers for details	In an example, data is extracted from baseline phases and 2 treatment phases (i.e., ABC) No guidelines are explicitly stated for extracting data	Need for standardization not addressed	Include 20+ units at each level	Assess level of autocorrelation When significant level exists, specify an autoregressive covariance structure	Recommendations: Estimate unconditional model before adding predictor variables Statistically test fixed and random effects For multiple dependent variables, use multivariate MLM methods Complement statistical analysis with visual analysis

Table 1 (continued)

Summary of articles on the use of MLM with single-subject data

Study	Model specifications	Data extraction/ simulation	Data standardization method	Unit counts at each level	Treatment of autocorrelation	Analysis process
Van den Noortgate & Onghena (2008)	Level 1 (a) $Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{phase})_{ijk} + e_{ijk}$ (b) $Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{phase})_{ijk} + \pi_{2jk}(\text{time})_{ijk} + \pi_{3jk}(\text{phase})_{ijk}(\text{time in treatment})_{ijk} + e_{ijk}$ (c) $\hat{\pi}_{1jk} = \pi_{1jk} + e_{1jk}$ $\hat{\pi}_{3jk} = \pi_{3jk} + e_{3jk}$ Level 2 $\pi_{pjk} = \beta_{p0k} + \beta_{p1k}(X_1)_{jk} + r_{pjk}$ Level 3 $\beta_{pqk} = \gamma_{p0} + \gamma_{p1}(Z_1)_k + u_{pqk}$ Notes Model (b) recommended when linear trends are present in data	In an example, data is extracted from baseline phases and treatment phases (i.e., AB) Authors briefly mention extracting data for additional treatments (e.g., ABC), but do not clarify how to incorporate data into models	When using model (a) or (b), conduct OLS regression analysis on each data set and divide dependent variable scores by RMSE When using model (c), calculate OLS regression coefficients for each data set and divide coefficients by RMSE and the corresponding covariance matrices by MSE	Unit counts not addressed	Assess level of autocorrelation Treatment of significant levels of autocorrelation not addressed	Recommendations: Estimate unconditional model before adding predictor variables Statistically test fixed and random effects
Van den Noortgate & Onghena (2011)	Level 1 (a) $Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{phase})_{ijk} + \pi_{2jk}(\text{time})_{ijk} + \pi_{3jk}(\text{phase})_{ijk}(\text{time in treatment})_{ijk} + e_{ijk}$ (b) $\pi_{2ik} = b_{2ik} + r_{2ik}$ and $\pi_{3jk} = b_{3jk} + r_{3ik}$ Level 2 (a) $\pi_{pjk} = \beta_{p0k} + r_{pjk}$ (b) $b_{pjk} = \beta_{p0k} + r_{pjk}$ Level 3 $\beta_{pqk} = \gamma_{p0} + u_{pqk}$ Notes Model (b) made use of unstandardized OLS regression coefficients as input at level 1	Using Monte Carlo methods, generated data to fit the equation at left, simulating one baseline and one treatment phase for each subject Simulated data for 30 conditions which varied in number of subjects per study (3, 4, or 8), number of measures per subject (10, 30, or 50), studies meta-analyzed (10 or 30), autocorrelation (0, .2, or .4), and magnitudes and variances of effects	Found that (a) standardization via division by RMSE is associated with biased estimates of fixed effects and inaccurate confidence intervals, and (b) use of OLS regression coefficients as input at level 1 is associated with unbiased estimates of fixed effects and accurate confidence intervals	Found that when data was standardized, (a) bias in fixed effects was worst for small numbers of data points, and (b) inaccuracy of confidence intervals increased as numbers of subjects and studies increased	Included autocorrelation as a factor in the study, but did not mention autocorrelation in report of results	Assessed relative bias and precision of estimates of fixed effects and accuracy of their confidence intervals

AB = baseline (A) and treatment (B) phase pair; SD = standard deviation; OLS = ordinary least squares; ABAB = 2 pairs of baseline (A) and treatment (B) phases; ABAC = 2 pairs of baseline (A) and treatment (B and C) phases in which treatments differ; RMSE = root mean square error resulting from an OLS regression analysis; MSE = mean square error resulting from an OLS regression analysis; MLM = multilevel modeling; ABC = baseline phase (A) followed by 2 different treatment phases (B and C)

Notes: Equation notations have been standardized for presentation in this paper. Please see the text for definitions of symbols.

Table 2

Summary of studies which applied MLM methods to single-subject data

Study	Model specifications	Unit counts at each level	Data extraction	Data standardization method	Treatment of autocorrelation	Analysis process
Adams (2009)	Level 1 $Y_{ij} = \pi_{0j} +$ $\pi_{1j}(\text{wear duration})_{ij} +$ $\pi_{2j}(\text{day of week})_{ij} +$ $\pi_{3j}(\text{time})_{ij} +$ $\pi_{4j}(\text{phase})_{ij} + e_{ij}$ Level 2 Not explicitly defined. Implied to be: $\pi_{pj} = \beta_{pj} + r_{pj}$	Level 1 >350 data points per dependent variable Level 2 5 subjects	All data was extracted (i.e., 1st baseline, treatment, and 2nd baseline phases; ABA)	Not necessary due to common dependent variable metric across cases	Examined level of auto-correlation Specified a heterogeneous autoregressive error covariance structure	Assessed fit of mean change, slope change, and mean and slope change models by comparing variance statistics and visually analyzing level 1 regression models Estimated unconditional model first Tested fixed and random effects, as well as deviance statistics Complemented statistical analysis with visual analysis Interpreted level 1 residuals Analyzed data for 2 dependent variables separately, without multivariate procedures
Hurwitz (2008)	Not explicitly defined. Implied to be: Level 1 $\overline{SMD}_j = \beta_k + r_j$ Level 2 $\beta_k = \gamma_0 + \gamma_1(\text{study2})_k +$ $\gamma_5(\text{\# of completed cases})_k$ $+ u_k$ Notes Collapsed a 4 level hierarchy (i.e., students, teachers, consultants, studies) to 2 levels	Level 1 650 subjects Level 2 202 consultants	All data was extracted (i.e., one baseline and one treatment phase; AB) For some subjects, data was extracted from multiple cases (i.e., 650 subjects comprised 987 cases) Extracted data was used to calculate SMD and average SMD, using one of the 3 different formulas for SMD, depending on data conditions	Not necessary due to common dependent variable metric across cases	Autocorrelation not addressed	Estimated unconditional model first Tested fixed and random effects

Table 2 (continued)

Summary of studies which applied MLM methods to single-subject data

Study	Model specifications	Unit counts at each level	Data extraction	Data standardization method	Treatment of autocorrelation	Analysis process
Miller (2006)	Level 1 $Y_{ij} = \pi_j(\text{phase})_{ij} + e_{ij}$ Level 2 $\pi_j = \beta_0 + r_j$	Level 1 1504 data points Level 2 84 subjects	All data was extracted; depending on data available, extraction involved pairs of baseline and treatment phases (i.e., AB), or first pairs plus additional phases (i.e., ABA, ABAB)	Baseline and treatment data converted to z-scores separately; mean baseline z-score subtracted from all z-scores	Autocorrelation not addressed	Estimated unconditional model first Tested fixed and random effect Explored inclusion of treatment variables at level 2, however all were found to be insignificant
Morgan & Sideridis (2006)	Level 1 $Y_{ij} = \pi_{0j} + \pi_{1j}(\text{time})_{ij} + \pi_{2j}(\text{baseline } \bar{Y})_{ij} + e_{ij}$ Level 2 $\pi_{0j} = \beta_{01}(\text{variable 1})_j + \beta_{02}(\text{variable 2})_j + \ldots + \beta_{010}(\text{variable 10})_j + r_{0j}$ $\pi_{1j} = \beta_{11}(\text{variable 1})_j + \beta_{12}(\text{variable 2})_j + \ldots + \beta_{110}(\text{variable 10})_j + r_{1j}$ $\pi_{2j} = \beta_{20} + r_{2j}$	Level 1 Number of data points not reported Level 2 144 cases	Data was extracted from first pairs of baseline and treatment phases only (i.e., AB) For some subjects, data was extracted from multiple cases (i.e., 107 subjects comprised 144 cases); weights were used to balance subjects' multiple contributions	No standardization method used, despite diversity in dependent variable metrics across cases	Autocorrelation not addressed	No mention of estimating an unconditional model Tested fixed effects and differences between fixed effects for treatment types No tests performed for random effects Adjusted alpha level to .025 to account for large number of tests
Terrazas Arellanes (2009)	Level 1 $Y_{ij} = \pi_{0j} + \pi_{1j}(\text{time})_{ij} + \pi_{2j}(\text{phase})_{ij} + \pi_{3j}(\text{phase})_{ij}(\text{time in treatment})_{ij} + e_{ij}$ Level 2 $\pi_{0j} = \beta_0 + r_{0j}$ $\pi_{1j} = \beta_1 + r_{1j}$ $\pi_{2j} = \beta_2 + r_{2j}$ $\pi_{3j} = \beta_3 + r_{3j}$	Level 1 158 data points per dependent variable Level 2 12 subjects	All data was extracted (i.e., one baseline and one treatment phase; AB)	Not necessary due to common dependent variable metric across cases	Autocorrelation not addressed	Only estimated unconditional model Tested fixed effects No tests performed for random effects Complemented statistical analysis with visual analysis Analyzed data for 2 dependent variables separately, without multivariate procedures

Table 2 (continued)

Summary of studies which applied MLM methods to single-subject data

Study	Model specifications	Unit counts at each level	Data extraction	Data standardization method	Treatment of autocorrelation	Analysis process
Wade, Ortiz, & Gorman (2007)	Level 1 $Y_{ij} = \pi_{1j}(\text{baseline})_{ij} + \pi_{2j}(\text{treatment})_{ij} + \pi_{3j}(\text{follow-up})_{ij} + \pi_{4j}(\text{baseline time})_{ij} + \pi_{5j}(\text{treatment time})_{ij} + \pi_{6j}(\text{follow-up time})_{ij} + e_{ij}$ Level 2 Not explicitly defined. Implied to be: $\pi_{pj} = \beta_{pj} + r_{pj}$	Level 1 210 data points per dependent variable Level 2 5 subjects	All data was extracted (i.e., one baseline, one treatment, and one follow-up phase; ABF)	Not necessary due to common dependent variable metric across cases	Autocorrelation not addressed	Only estimated unconditional model Tested fixed effects, random effects, and differences between fixed effects for adjacent phases Complemented statistical analysis with visual analysis Analyzed data for 3 dependent variables separately, without multivariate procedures
Wang, Cui, & Parrila (2011)	Level 1 $Y_{ijk} = \pi_{jk}(\text{phase})_{ijk} + e_{ijk}$ Level 2 $\pi_{jk} = \beta_k + r_{jk}$ Level 3 $\beta_k = \gamma_0 + \gamma_1(\text{treatment})_k + \gamma_2(\text{age})_k + \gamma_3(\text{age*treatment})_k + u_k$	Level 1 1796 data points Level 2 89 measures of effect Level 3 43 participants	Data was extracted from first pairs of baseline and treatment phases only (i.e., AB) For some subjects, data was extracted for multiple dependent variables (i.e., 43 subjects comprised 89 measures of effect); weights were not used to balance subjects' multiple contributions	Baseline and treatment data converted to z-scores using overall means and standard deviations for each subjects' data; mean baseline z-score subtracted from all z-scores	Autocorrelation not addressed	No mention of estimating an unconditional model Tested fixed effects No tests performed for random effect

ABA = baseline phase (A) followed by a treatment (B) and second baseline (A) phases; AB = baseline (A) and treatment (B) phase pair; SMD = standardized mean difference statistic (Busk & Serlin, 1992); ABAB = 2 pairs of baseline (A) and treatment (B) phases; ABF = baseline (A), treatment (B), and follow-up (F) phases

Notes: Equation notations have been standardized for presentation in this paper. Please see the text for definitions of symbols.

Table 3

Summary of factor levels

Factors	Factor levels		
Number of data points per phase	5 baseline 5 treatment	10 baseline 10 treatment	20 baseline 20 treatment
Number of participants per study	3	6	
Number of studies meta-analyzed	10	30	
Degree of autocorrelation in individuals' data sets	0.0	0.4	
Continuity of level 1 variance across phases	Continuous $\sigma^2_{single} = 150$	Discontinuous $\sigma^2_{baseline} = 300$ $\sigma^2_{treatment} = 70$	

Table 4

Power rates by fixed effect and level 1 error specification for conditions in which $\sigma^2_{baseline} = 150$, $\sigma^2_{treatment} = 150$, and $\rho_{ar(1)} = 0.0$

Participants per study	Studies	Baseline data points	Treatment data points	γ_0			γ_1			γ_2			γ_3			γ_4		
				VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†
3	10	5	5	1.0	1.0	1.0	**.028**	**.028**	**.028**	1.0	1.0	1.0	**.425**	**.403**	**.440**	**.000**	**.000**	**.000**
6	10	5	5	1.0	1.0	1.0	**.020**	**.018**	**.018**	1.0	1.0	1.0	.803	**.795**	.813	**.015**	**.015**	**.018**
3	30	5	5	1.0	1.0	1.0	**.053**	**.048**	**.048**	1.0	1.0	1.0	.975	.968	.978	**.003**	**.003**	**.000**
6	30	5	5	1.0	1.0	1.0	**.075**	**.068**	**.070**	1.0	1.0	1.0	1.0	1.0	1.0	**.003**	**.003**	**.003**
3	10	10	10	1.0	1.0	1.0	**.198**	**.180**	**.178**	1.0	1.0	1.0	1.0	1.0	1.0	**.620**	**.598**	**.663**
6	10	10	10	1.0	1.0	1.0	**.300**	**.270**	**.255**	1.0	1.0	1.0	1.0	1.0	1.0	.938	.930	.948
3	30	10	10	1.0	1.0	1.0	**.468**	**.435**	**.423**	1.0	1.0	1.0	1.0	1.0	1.0	.983	.980	.983
6	30	10	10	1.0	1.0	1.0	**.723**	**.708**	**.700**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	10	20	20	1.0	1.0	1.0	**.790**	**.783**	**.763**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	10	20	20	1.0	1.0	1.0	.960	.950	.945	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	30	20	20	1.0	1.0	1.0	.998	.998	.998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	30	20	20	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure; bold print indicates power was inadequate (i.e., <0.80)

†"Double" refers to the inclusion of two error terms – one for each phase

Table 5

Power rates by fixed effect and level 1 error specification for conditions in which $\sigma^2_{baseline} = 300$, $\sigma^2_{treatment} = 70$, and $\rho_{ar(1)} = 0.0$

Participants per study	Studies	Baseline data points	Treatment data points	γ_0			γ_1			γ_2			γ_3			γ_4		
				VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†
3	10	5	5	1.0	1.0	1.0	**.080**	**.055**	**.023**	1.0	1.0	1.0	**.295**	**.265**	**.705**	**.000**	**.000**	**.030**
6	10	5	5	1.0	1.0	1.0	**.085**	**.075**	**.035**	1.0	1.0	1.0	**.743**	**.715**	.975	**.000**	**.000**	**.010**
3	30	5	5	1.0	1.0	1.0	**.093**	**.083**	**.033**	1.0	1.0	1.0	.928	.923	.998	**.000**	**.000**	**.020**
6	30	5	5	1.0	1.0	1.0	**.158**	**.145**	**.065**	1.0	1.0	1.0	1.0	1.0	1.0	**.000**	**.000**	**.010**
3	10	10	10	1.0	1.0	1.0	**.160**	**.128**	**.075**	1.0	1.0	1.0	1.0	1.0	1.0	**.495**	**.438**	.890
6	10	10	10	1.0	1.0	1.0	**.278**	**.243**	**.150**	1.0	1.0	1.0	1.0	1.0	1.0	.903	.880	.995
3	30	10	10	1.0	1.0	1.0	**.410**	**.380**	**.225**	1.0	1.0	1.0	1.0	1.0	1.0	.990	.980	1.0
6	30	10	10	1.0	1.0	1.0	**.578**	**.548**	**.430**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	10	20	20	1.0	1.0	1.0	**.613**	**.573**	**.480**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	10	20	20	1.0	1.0	1.0	.878	.840	**.748**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	30	20	20	1.0	1.0	1.0	.935	.930	.928	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	30	20	20	1.0	1.0	1.0	.990	.990	.983	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure; bold print indicates power was inadequate (i.e., <0.80)

†"Double" refers to the inclusion of two error terms – one for each phase

Table 6

Power rates by fixed effect and level 1 error specification for conditions in which $\sigma^2_{baseline} = 150$, $\sigma^2_{treatment} = 150$, and $\rho_{ar(1)} = 0.4$

Participants per study	Studies	Baseline data points	Treatment data points	γ_0			γ_1			γ_2			γ_3			γ_4		
				VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†
3	10	5	5	1.0	1.0	1.0	**.205**	**.083**	**.183**	1.0	1.0	1.0	**.740**	**.618**	**.795**	**.055**	**.035**	**.073**
6	10	5	5	1.0	1.0	1.0	**.163**	**.060**	**.133**	1.0	1.0	1.0	.970	.938	.978	**.038**	**.033**	**.080**
3	30	5	5	1.0	1.0	1.0	**.300**	**.155**	**.280**	1.0	1.0	1.0	.985	.975	.990	**.045**	**.038**	**.068**
6	30	5	5	1.0	1.0	1.0	**.250**	**.088**	**.210**	1.0	1.0	1.0	1.0	1.0	1.0	**.053**	**.050**	**.085**
3	10	10	10	1.0	1.0	1.0	**.250**	**.123**	**.255**	1.0	1.0	1.0	.998	.998	1.0	**.723**	**.528**	**.733**
6	10	10	10	1.0	1.0	1.0	**.255**	**.123**	**.273**	1.0	1.0	1.0	1.0	1.0	1.0	.963	.875	.968
3	30	10	10	1.0	1.0	1.0	**.360**	**.183**	**.365**	1.0	1.0	1.0	1.0	1.0	1.0	.983	.965	.985
6	30	10	10	1.0	1.0	1.0	**.690**	**.478**	**.698**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.998	1.0
3	10	20	20	1.0	1.0	1.0	**.450**	**.338**	**.458**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	10	20	20	1.0	1.0	1.0	**.718**	**.570**	**.708**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	30	20	20	1.0	1.0	1.0	**.788**	**.698**	**.793**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	30	20	20	1.0	1.0	1.0	.978	.958	.978	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure; bold print indicates power was inadequate (i.e., <0.80)

†"Double" refers to the inclusion of two error terms – one for each phase

Table 7

Power rates by fixed effect and level 1 error specification for conditions in which $\sigma^2_{baseline} = 300$, $\sigma^2_{treatment} = 70$, and $\rho_{ar(1)} = 0.4$

Participants per study	Studies	Baseline data points	Treatment data points	γ_0			γ_1			γ_2			γ_3			γ_4		
				VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†
3	10	5	5	1.0	1.0	1.0	**.280**	**.160**	**.195**	1.0	1.0	1.0	**.673**	**.493**	.875	**.003**	**.000**	**.093**
6	10	5	5	1.0	1.0	1.0	**.278**	**.133**	**.180**	1.0	1.0	1.0	.940	.868	.993	**.003**	**.003**	**.060**
3	30	5	5	1.0	1.0	1.0	**.205**	**.070**	**.115**	1.0	1.0	1.0	.995	.990	1.0	**.003**	**.003**	**.043**
6	30	5	5	1.0	1.0	1.0	**.238**	**.108**	**.135**	1.0	1.0	1.0	1.0	1.0	1.0	**.003**	**.005**	**.085**
3	10	10	10	1.0	1.0	1.0	**.218**	**.123**	**.215**	1.0	1.0	1.0	1.0	.993	1.0	**.690**	**.333**	**.855**
6	10	10	10	1.0	1.0	1.0	**.195**	**.118**	**.238**	1.0	1.0	1.0	1.0	1.0	1.0	.973	**.753**	.990
3	30	10	10	1.0	1.0	1.0	**.228**	**.150**	**.260**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.975	1.0
6	30	10	10	1.0	1.0	1.0	**.448**	**.325**	**.498**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	10	20	20	1.0	1.0	1.0	**.323**	**.268**	**.345**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	10	20	20	1.0	1.0	1.0	**.540**	**.470**	**.558**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	30	20	20	1.0	1.0	1.0	**.628**	**.598**	**.643**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	30	20	20	1.0	1.0	1.0	.860	.803	.883	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure; bold print indicates power was inadequate (i.e., <0.80)

†"Double" refers to the inclusion of two error terms – one for each phase

Table 8

Relative parameter bias of fixed effects by parameter and level 1 error specification for conditions in which $\sigma^2_{baseline} = 150$, $\sigma^2_{treatment} = 150$, *and* $\rho_{ar(1)} = 0.0$

Participants per study	Studies	Baseline data points	Treatment data points	γ_0			γ_1			γ_2			γ_3			γ_4		
				VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†
3	10	5	5	-.003	-.003	-.003	**.068**	**.073**	**.068**	-.002	-.001	-.002	.040	.038	.040	**-1.97**	**-1.95**	**-1.97**
6	10	5	5	.000	.001	.000	-.027	-.028	-.027	-.004	-.004	-.004	.031	.030	.031	**-2.02**	**-2.01**	**-2.02**
3	30	5	5	.000	.000	.000	.014	.016	.014	-.001	-.001	-.001	.045	.045	.045	**-2.09**	**-2.09**	**-2.08**
6	30	5	5	.000	.000	.000	-.008	-.006	-.008	-.005	-.005	-.005	.025	.025	.025	**-1.82**	**-1.83**	**-1.82**
3	10	10	10	-.001	-.001	-.001	.006	.009	.006	-.021	-.021	-.021	**.193**	**.193**	**.193**	**-3.86**	**-3.85**	**-3.86**
6	10	10	10	.001	.001	.001	-.024	-.017	-.024	-.021	-.020	-.021	**.187**	**.186**	**.187**	**-3.85**	**-3.85**	**-3.85**
3	30	10	10	-.002	-.002	-.002	-.023	-.018	-.023	-.023	-.022	-.023	**.191**	**.190**	**.191**	**-3.89**	**-3.88**	**-3.88**
6	30	10	10	-.000	.000	.000	.005	.009	.005	-.020	-.019	-.020	**.187**	**.186**	**.187**	**-3.85**	**-3.84**	**-3.85**
3	10	20	20	-.001	-.001	-.001	-.013	-.016	-.009	.024	.023	.024	.042	.041	.045	**-2.95**	**-2.94**	**-2.95**
6	10	20	20	-.001	.000	-.001	-.021	-.025	-.020	.031	.030	.031	.027	.026	.028	**-2.90**	**-2.90**	**-2.90**
3	30	20	20	.000	.000	.000	-.024	-.028	-.028	.025	.024	.025	.035	.034	.035	**-2.93**	**-2.92**	**-2.92**
6	30	20	20	.001	.001	.000	-.020	-.025	-.019	.025	.023	.024	.035	.034	.036	**-2.92**	**-2.92**	**-2.93**

VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure; bold print indicates bias was unacceptable (> 0.05)

†"Double" refers to the inclusion of two error terms – one for each phase

Table 9

Relative parameter bias of fixed effects by parameter and level 1 error specification for conditions in which $\sigma^2_{baseline} = 300$, $\sigma^2_{treatment} = 70$, and $\rho_{ar(1)} = 0.0$

Participants per study	Studies	Baseline data points	Treatment data points	γ_0			γ_1			γ_2			γ_3			γ_4		
				VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†
3	10	5	5	-.004	-.004	-.004	**.094**	**.088**	**.094**	-.006	-.007	-.006	**.061**	**.065**	**.061**	**-2.01**	**-2.06**	**-2.01**
6	10	5	5	-.003	-.003	-.003	.014	.020	.014	-.009	-.009	-.009	**.057**	**.056**	**.057**	**-1.80**	**-1.79**	**-1.80**
3	30	5	5	-.003	-.003	-.003	-.030	-.030	-.030	-.009	-.009	-.009	.040	.041	.040	**-1.65**	**-1.65**	**-1.65**
6	30	5	5	-.001	-.001	-.001	**-.082**	**-.076**	**-.082**	-.008	-.008	-.008	.048	.048	.048	**-1.75**	**-1.76**	**-1.75**
3	10	10	10	-.004	-.004	-.004	-.045	-.048	-.045	-.031	-.031	-.031	**.203**	**.202**	**.203**	**-3.80**	**-3.80**	**-3.80**
6	10	10	10	-.003	-.003	-.003	.021	.025	.021	-.027	-.026	-.027	**.222**	**.221**	**.222**	**-3.84**	**-3.84**	**-3.84**
3	30	10	10	-.006	-.006	-.006	**.061**	**.065**	**.061**	-.025	-.024	-.025	**.209**	**.207**	**.209**	**-3.81**	**-3.80**	**-3.81**
6	30	10	10	-.001	-.001	-.001	-.015	-.010	-.015	-.027	-.026	-.027	**.215**	**.213**	**.215**	**-3.84**	**-3.83**	**-3.84**
3	10	20	20	-.001	-.001	-.001	**-.062**	**-.067**	-.047	.007	.006	.009	**.080**	**.079**	**.081**	**-3.01**	**-3.00**	**-3.01**
6	10	20	20	-.004	-.004	-.003	**-.054**	**-.058**	**-.065**	.013	.012	.011	**.068**	**.067**	**.066**	**-2.98**	**-2.98**	**-2.98**
3	30	20	20	-.002	-.002	-.004	-.018	-.022	-.006	.016	.015	.019	**.077**	**.076**	**.075**	**-3.00**	**-2.99**	**-2.99**
6	30	20	20	-.001	.000	.000	**-.055**	**-.059**	-.047	.011	.010	.012	**.076**	**.075**	**.078**	**-3.00**	**-2.99**	**-3.00**

VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure; bold print indicates bias was unacceptable (> 0.05)

†"Double" refers to the inclusion of two error terms – one for each phase

Table 10

Relative parameter bias of fixed effects by parameter and level 1 error specification for conditions in which $\sigma^2_{baseline} = 150$, $\sigma^2_{treatment} = 150$, and $\rho_{ar(1)} = 0.4$

Participants per study	Studies	Baseline data points	Treatment data points	γ_0			γ_1			γ_2			γ_3			γ_4		
				VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†
3	10	5	5	-.001	-.001	-.001	.004	.000	.004	-.009	-.007	-.009	**.078**	**.075**	**.078**	**-2.51**	**-2.51**	**-2.51**
6	10	5	5	-.004	-.004	-.004	**.152**	**.170**	**.152**	-.001	.001	-.001	**.069**	**.066**	**.069**	**-2.41**	**-2.39**	**-2.41**
3	30	5	5	-.001	-.001	-.001	-.007	-.001	-.007	-.005	-.003	-.005	**.059**	**.056**	**.059**	**-2.26**	**-2.26**	**-2.26**
6	30	5	5	.000	.000	.000	.010	.017	.010	-.003	-.001	-.003	.047	.045	.047	**-2.07**	**-2.08**	**-2.07**
3	10	10	10	.000	-.001	.000	.012	**.066**	.012	-.010	.005	-.010	**.187**	**.168**	**.187**	**-3.91**	**-3.82**	**-3.91**
6	10	10	10	-.001	-.001	-.001	-.029	-.011	-.029	-.023	-.012	-.023	**.194**	**.173**	**.194**	**-3.91**	**-3.82**	**-3.91**
3	30	10	10	.001	.001	.001	**-.054**	-.015	**-.054**	-.021	-.007	-.021	**.179**	**.157**	**.179**	**-3.81**	**-3.71**	**-3.81**
6	30	10	10	-.001	-.001	-.001	.018	**.056**	.018	-.019	-.005	-.019	**.184**	**.165**	**.184**	**-3.82**	**-3.74**	**-3.82**
3	10	20	20	.001	.002	.000	.006	**-.060**	.012	.033	.010	.034	.034	.027	.034	**-2.92**	**-2.86**	**-2.97**
6	10	20	20	.000	.002	.001	-.032	**-.102**	-.034	.028	.002	.028	.034	.028	.034	**-2.93**	**-2.88**	**-2.93**
3	30	20	20	.000	.001	.000	-.033	**-.084**	-.034	.023	.005	.024	.036	.025	.032	**-2.93**	**-2.86**	**-2.92**
6	30	20	20	-.002	.000	-.001	-.005	**-.078**	-.009	.028	.003	.026	.032	.025	.032	**-2.92**	**-2.86**	**-2.91**

VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure; bold print indicates bias was unacceptable (> 0.05)

†"Double" refers to the inclusion of two error terms – one for each phase

Table 11

Relative parameter bias of fixed effects by parameter and level 1 error specification for conditions in which $\sigma^2_{baseline} = 300$, $\sigma^2_{treatment} = 70$, and $\rho_{ar(1)} = 0.4$

Participants per study	Studies	Baseline data points	Treatment data points	γ_0			γ_1			γ_2			γ_3			γ_4		
				VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†
3	10	5	5	-.004	-.005	-.004	**.070**	**.129**	**.070**	-.007	-.005	-.007	**.065**	**.067**	**.065**	**-1.93**	**-1.93**	**-1.93**
6	10	5	5	-.003	-.004	-.003	**.067**	**.115**	**.067**	-.007	-.004	-.007	**.055**	**.052**	**.055**	**-1.69**	**-1.67**	**-1.69**
3	30	5	5	-.001	-.001	-.001	**-.051**	**-.058**	**-.051**	-.006	-.005	-.006	.046	.043	.046	**-1.74**	**-1.75**	**-1.74**
6	30	5	5	-.002	-.002	-.002	**-.057**	-.045	**-.057**	-.010	-.009	-.010	.048	.048	.048	**-1.73**	**-1.75**	**-1.73**
3	10	10	10	.001	.001	.001	**-.060**	**-.012**	**-.060**	-.023	-.004	-.023	**.199**	**.174**	**.199**	**-3.73**	**-3.63**	**-3.73**
6	10	10	10	-.004	-.004	-.004	-.002	.013	-.003	-.026	-.011	-.026	**.206**	**.180**	**.206**	**-3.79**	**-3.68**	**-3.79**
3	30	10	10	-.001	-.002	-.001	**-.093**	-.035	**-.093**	-.030	-.012	-.030	**.220**	**.196**	**.220**	**-3.90**	**-3.78**	**-3.90**
6	30	10	10	-.001	-.002	-.001	-.029	.007	-.028	-.030	-.014	-.030	**.219**	**.194**	**.219**	**-3.85**	**-3.74**	**-3.85**
3	10	20	20	.002	.002	.002	**-.085**	**-.120**	**-.086**	.001	-.011	.001	**.079**	**.064**	**.079**	**-3.01**	**-2.93**	**-3.01**
6	10	20	20	-.006	-.004	-.006	-.008	**-.074**	-.009	.016	-.006	.015	**.071**	**.060**	**.071**	**-2.98**	**-2.91**	**-2.98**
3	30	20	20	-.002	-.001	-.002	**-.064**	**-.109**	**-.065**	.009	-.009	.009	**.066**	**.055**	**.065**	**-2.97**	**-2.91**	**-2.97**
6	30	20	20	-.001	.001	-.001	**-.056**	**-.116**	**-.052**	.011	-.006	.010	**.073**	**.060**	**.073**	**-2.99**	**-2.92**	**-2.99**

VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure; bold print indicates bias was unacceptable (> 0.05)

†"Double" refers to the inclusion of two error terms – one for each phase

Table 12

Relative parameter bias of $T_{\gamma 0}$ by factor levels and level 1 error specification

	Participants per study	3	6	3	6	3	6	3	6	3	6	3	6
	Studies	10	10	30	30	10	10	30	30	10	10	30	30
	Baseline data points	5	5	5	5	10	10	10	10	20	20	20	20
	Treatment data points	5	5	5	5	10	10	10	10	20	20	20	20
Grouping	Level 1 error spec.	Relative parameter bias of $T_{\gamma 0}$											
$\sigma^2 = 150, 150$	VC	**1.84**	**-.536**	**.776**	**-.746**	**2.74**	**-.638**	**.337**	**.424**	**.302**	**-.267**	**-.231**	**-.621**
$\rho_{ar(1)} = 0.0$	AR(1)	**1.84**	**-.536**	**.780**	**-.746**	**2.70**	**-.641**	**.321**	**.416**	**.308**	**-.268**	**-.244**	**-.625**
	Double†	**1.86**	**-.531**	**.787**	**-.743**	**2.71**	**-.635**	**.339**	**.434**	**.308**	**-.262**	**-.228**	**-.620**
$\sigma^2 = 300, 70$	VC	**5.59**	**1.27**	**4.68**	**.655**	**1.99**	**.513**	**1.95**	**.578**	-.043	**-.377**	**-.454**	**.430**
$\rho_{ar(1)} = 0.0$	AR(1)	**5.51**	**1.25**	**4.56**	**.672**	**1.93**	**.487**	**1.86**	**.552**	**-.086**	**-.377**	**-.473**	**.385**
	Double†	**5.55**	**1.36**	**4.71**	**.710**	**2.04**	**.628**	**1.79**	**.679**	.019	**-.308**	**-.413**	**.518**
$\sigma^2 = 150, 150$	VC	**3.56**	**-.644**	**1.87**	**-.783**	**4.80**	**-.293**	**-.689**	**-.606**	**1.52**	**.790**	**5.32**	**2.63**
$\rho_{ar(1)} = 0.4$	AR(1)	**3.39**	**-.617**	**1.58**	**-.796**	**2.45**	**-.541**	**-.697**	**-.694**	**0.15**	**.141**	**2.48**	**1.40**
	Double†	**3.47**	**-.636**	**1.68**	**-.796**	**4.41**	**-.330**	**-.671**	**-.601**	**1.41**	**.733**	**5.15**	**2.57**
$\sigma^2 = 300, 70$	VC	**4.99**	**1.66**	**1.32**	**.091**	**3.17**	**-.658**	**-.804**	**-0.81**	**1.00**	**1.59**	**6.04**	**2.83**
$\rho_{ar(1)} = 0.4$	AR(1)	**3.99**	**1.17**	**1.24**	**-.069**	**1.28**	**-.735**	**-.801**	**-.904**	**-.321**	**0.28**	**.426**	**-.135**
	Double†	**4.69**	**1.51**	**1.50**	-.021	**2.60**	**-.670**	**-.700**	**-.839**	**.714**	**1.23**	**5.05**	**2.28**

Spec. = specification; VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure

†"Double" refers to the inclusion of two error terms – one for each phase

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 13

Relative parameter bias of $T_{\beta 0}$ by factor levels and level 1 error specification

	Participants per study	3	6	3	6	3	6	3	6	3	6	3	6
	Studies	10	10	30	30	10	10	30	30	10	10	30	30
	Baseline data points	5	5	5	5	10	10	10	10	20	20	20	20
	Treatment data points	5	5	5	5	10	10	10	10	20	20	20	20
Grouping	Level 1 error spec.	Relative parameter bias of $T_{\beta 0}$											
$\sigma^2 = 150, 150$	VC	**-.143**	-.041	**-.080**	-.034	**-.122**	.044	-.009	.007	**.111**	**.142**	**.183**	**.184**
$\rho_{ar(1)} = 0.0$	AR(1)	**-.163**	**-.061**	**-.098**	-.050	**-.149**	.017	-.033	-.019	**.088**	**.126**	**.160**	**.160**
	Double†	**-.154**	-.050	**-.086**	-.042	**-.132**	.035	-.018	-.003	**.103**	**.138**	**.179**	**.179**
$\sigma^2 = 300, 70$	VC	**-.272**	-.044	**-.185**	-.038	**.157**	**.159**	**.134**	**.169**	**.155**	**.174**	**.191**	**.138**
$\rho_{ar(1)} = 0.0$	AR(1)	**-.321**	**-.095**	**-.237**	**-.093**	**.080**	**.097**	**.075**	**.111**	**.090**	**.118**	**.139**	**.087**
	Double†	**-.401**	**-.169**	**-.297**	**-.147**	.045	**.076**	**.054**	**.094**	**.103**	**.124**	**.154**	**.111**
$\sigma^2 = 150, 150$	VC	**.767**	**.842**	**0.76**	**.814**	**.704**	**.897**	**.936**	**.928**	**.771**	**.798**	**.683**	**.784**
$\rho_{ar(1)} = 0.4$	AR(1)	-.025	**.113**	.048	**.135**	**-.300**	**-.137**	**-.098**	**-.103**	**-.243**	**-.244**	**-.347**	**-.288**
	Double†	**.690**	**.773**	**.699**	**.754**	**.586**	**.771**	**.805**	**.793**	**.662**	**.686**	**.571**	**.658**
$\sigma^2 = 300, 70$	VC	**1.78**	**1.72**	**1.77**	**1.65**	**1.93**	**2.08**	**2.04**	**2.02**	**1.67**	**1.66**	**1.05**	**1.41**
$\rho_{ar(1)} = 0.4$	AR(1)	**.065**	**.275**	**.300**	**.365**	**-.079**	-.011	-.002	-.009	**-.250**	**-.259**	**-.383**	**-.316**
	Double†	**1.10**	**1.15**	**1.18**	**1.15**	**1.05**	**1.12**	**1.08**	**1.06**	**1.15**	**1.12**	**.624**	**.934**

Spec. = specification; VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure

†"Double" refers to the inclusion of two error terms – one for each phase

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 14

Relative parameter bias of T_{B1} by factor levels and level 1 error specification

	Participants per study	3	6	3	6	3	6	3	6	3	6	3	6
	Studies	10	10	30	30	10	10	30	30	10	10	30	30
	Baseline data points	5	5	5	5	10	10	10	10	20	20	20	20
	Treatment data points	5	5	5	5	10	10	10	10	20	20	20	20
Grouping	Level 1 error spec.	Relative parameter bias of $T_{\beta 1}$											
$\sigma^2 = 150, 150$	VC	**-.182**	**-.641**	**-.811**	**-.969**	**-.747**	**-.859**	**-.927**	**-.933**	**-.730**	**-.775**	**-.891**	**-.896**
$\rho_{ar(1)} = 0.0$	AR(1)	**-.131**	**-.621**	**-.780**	**-.965**	**-.764**	**-.875**	**-.936**	**-.944**	**-.861**	**-.924**	**-.978**	**-.987**
	Double†	**-.191**	**-.647**	**-.825**	**-.971**	**-.814**	**-.902**	**-.959**	**-.974**	**-.797**	**-.850**	**-.950**	**-.964**
$\sigma^2 = 300, 70$	VC	**.884**	**-.336**	**.122**	**-.589**	**.084**	**-.209**	**-.085**	**-.412**	**-.820**	**-.859**	**-.911**	**-.961**
$\rho_{ar(1)} = 0.0$	AR(1)	**.798**	**-.314**	**.166**	**-.526**	**-.078**	**-.321**	**-.227**	**-.538**	**-.915**	**-.945**	**-.972**	**-.997**
	Double†	**1.15**	**-.270**	-.015	**-.819**	**-.715**	**-.923**	**-.937**	**-.992**	**-.629**	**-.676**	**-.757**	**-.847**
$\sigma^2 = 150, 150$	VC	**9.36**	**5.29**	**5.57**	**3.51**	**10.1**	**10.3**	**10.2**	**9.94**	**3.43**	**3.44**	**3.47**	**3.41**
$\rho_{ar(1)} = 0.4$	AR(1)	**.253**	**-.694**	**-.784**	**-.975**	**-.535**	**-.852**	**-.932**	**-.982**	**-.944**	**-.978**	**-.992**	**-1.00**
	Double†	**4.09**	**.977**	**1.04**	**-.337**	**7.81**	**7.91**	**7.83**	**7.46**	**3.02**	**3.05**	**3.08**	**3.00**
$\sigma^2 = 300, 70$	VC	**15.3**	**7.61**	**7.73**	**4.01**	**20.8**	**21.3**	**20.0**	**20.4**	**6.25**	**5.89**	**6.27**	**6.18**
$\rho_{ar(1)} = 0.4$	AR(1)	**.059**	**-.922**	**-.974**	**-1.00**	**-.682**	**-.878**	**-.971**	**-.997**	**-.955**	**-.987**	**-.973**	**-.995**
	Double†	**.240**	**-.976**	**-.975**	**-1.00**	**5.88**	**5.30**	**4.78**	**4.26**	**4.13**	**3.81**	**4.32**	**4.23**

Spec. = specification; VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure

†"Double" refers to the inclusion of two error terms – one for each phase

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 15

Relative parameter bias of $T_{\beta 2}$ by factor levels and level 1 error specification

	Participants per study	3	6	3	6	3	6	3	6	3	6	3	6
	Studies	10	10	30	30	10	10	30	30	10	10	30	30
	Baseline data points	5	5	5	5	10	10	10	10	20	20	20	20
	Treatment data points	5	5	5	5	10	10	10	10	20	20	20	20
Grouping	Level 1 error spec.	Relative parameter bias of $T_{\beta 2}$											
$\sigma^2 = 150, 150$	VC	**.610**	**.859**	**.731**	**.820**	**1.33**	**1.41**	**1.37**	**1.42**	**2.85**	**2.97**	**2.94**	**3.04**
$\rho_{ar(1)} = 0.0$	AR(1)	**.529**	**.733**	**.608**	**.712**	**1.14**	**1.20**	**1.19**	**1.24**	**2.42**	**2.54**	**2.50**	**2.58**
	Double†	**.640**	**.897**	**.752**	**.843**	**1.44**	**1.51**	**1.46**	**1.51**	**3.16**	**3.29**	**3.28**	**3.39**
$\sigma^2 = 300, 70$	VC	**.384**	**.558**	**.482**	**.669**	**1.32**	**1.33**	**1.30**	**1.36**	**2.27**	**2.44**	**2.30**	**2.39**
$\rho_{ar(1)} = 0.0$	AR(1)	**.146**	**.181**	**.082**	**.207**	**1.12**	**1.15**	**1.14**	**1.19**	**1.76**	**1.93**	**1.81**	**1.85**
	Double†	**.682**	**.846**	**.716**	**.856**	**1.85**	**1.85**	**1.92**	**1.88**	**3.06**	**3.21**	**3.16**	**3.36**
$\sigma^2 = 150, 150$	VC	**9.64**	**9.23**	**9.13**	**8.97**	**10.3**	**11.1**	**11.0**	**10.9**	**9.56**	**9.70**	**9.46**	**9.59**
$\rho_{ar(1)} = 0.4$	AR(1)	**6.12**	**5.92**	**5.88**	**5.94**	**3.66**	**3.79**	**3.64**	**3.63**	**2.22**	**2.15**	**2.08**	**2.11**
	Double†	**9.44**	**9.25**	**9.14**	**9.11**	**9.75**	**10.4**	**10.4**	**10.3**	**9.62**	**9.74**	**9.58**	**9.67**
$\sigma^2 = 300, 70$	VC	**10.0**	**9.05**	**9.45**	**9.26**	**13.3**	**13.8**	**13.2**	**13.5**	**12.1**	**12.0**	**10.8**	**11.0**
$\rho_{ar(1)} = 0.4$	AR(1)	**4.93**	**4.76**	**5.04**	**5.14**	**2.99**	**3.01**	**2.85**	**2.89**	**1.53**	**1.23**	**1.34**	**1.33**
	Double†	**8.70**	**8.43**	**8.60**	**8.73**	**8.91**	**8.89**	**8.60**	**8.64**	**9.82**	**9.44**	**9.07**	**9.06**

Spec. = specification; VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure

†"Double" refers to the inclusion of two error terms – one for each phase

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 16

Relative parameter bias of $T_{\beta 3}$ by factor levels and level 1 error specification

	Participants per study	3	6	3	6	3	6	3	6	3	6	3	6
	Studies	10	10	30	30	10	10	30	30	10	10	30	30
	Baseline data points	5	5	5	5	10	10	10	10	20	20	20	20
	Treatment data points	5	5	5	5	10	10	10	10	20	20	20	20
Grouping	Level 1 error spec.	Relative parameter bias of $T_{\beta 3}$											
$\sigma^2 = 150, 150$	VC	**-.585**	**-.587**	**-.579**	**-.563**	**-.852**	**-.841**	**-.842**	**-.833**	**-.946**	**-.947**	**-.951**	**-.952**
$\rho_{ar(1)} = 0.0$	AR(1)	**-.622**	**-.617**	**-.614**	**-.601**	**-.862**	**-.853**	**-.853**	**-.845**	**-.957**	**-.958**	**-.959**	**-.959**
	Double†	**-.570**	**-.574**	**-.562**	**-.539**	**-.843**	**-.832**	**-.832**	**-.822**	**-.932**	**-.939**	**-.946**	**-.947**
$\sigma^2 = 300, 70$	VC	**-.870**	**-.914**	**-.924**	**-.954**	**-.878**	**-.884**	**-.868**	**-.878**	**-.966**	**-.963**	**-.964**	**-.964**
$\rho_{ar(1)} = 0.0$	AR(1)	**-.912**	**-.951**	**-.958**	**-.981**	**-.891**	**-.897**	**-.879**	**-.890**	**-.972**	**-.969**	**-.969**	**-.968**
	Double†	**-.534**	**-.502**	**-.453**	**-.420**	**-.795**	**-.797**	**-.774**	**-.786**	**-.787**	**-.797**	**-.821**	**-.834**
$\sigma^2 = 150, 150$	VC	**.130**	**.104**	**.148**	**.143**	**-.457**	**-.477**	**-.473**	**-.482**	**-.549**	**-.598**	**-.604**	**-.602**
$\rho_{ar(1)} = 0.4$	AR(1)	**-.569**	**-.623**	**-.561**	**-.574**	**-.925**	**-.932**	**-.935**	**-.941**	**-.986**	**-.988**	**-.988**	**-.989**
	Double†	**.248**	**.242**	**.284**	**.267**	**-.339**	**-.436**	**-.457**	**-.482**	**-.334**	**-.363**	**-.361**	**-.342**
$\sigma^2 = 300, 70$	VC	**-.600**	**-.671**	**-.712**	**-.761**	**-.472**	**-.439**	**-.454**	**-.436**	**-.795**	**-.814**	**-.792**	**-.805**
$\rho_{ar(1)} = 0.4$	AR(1)	**-.997**	**-1.00**	**-1.00**	**-1.00**	**-.996**	**-.998**	**-.999**	**-1.00**	**-.997**	**-.999**	**-1.00**	**-1.00**
	Double†	**.010**	**.077**	**.053**	**.055**	**-.176**	**-.222**	**-.249**	**-.345**	**-.305**	**-.310**	**-.303**	**-.326**

Spec. = specification; VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure

†"Double" refers to the inclusion of two error terms – one for each phase

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 17

Relative parameter bias of $T_{\beta 4}$ by factor levels and level 1 error specification

	Participants per study	3	6	3	6	3	6	3	6	3	6	3	6
	Studies	10	10	30	30	10	10	30	30	10	10	30	30
	Baseline data points	5	5	5	5	10	10	10	10	20	20	20	20
	Treatment data points	5	5	5	5	10	10	10	10	20	20	20	20
Grouping	Level 1 error spec.	Relative parameter bias of $T_{\beta 4}$											
$\sigma^2 = 150, 150$	VC	**2.81**	**1.76**	**.981**	**-.344**	**-.931**	**-.969**	**-.984**	**-.998**	**-.960**	**-.962**	**-.968**	**-.967**
$\rho_{ar(1)} = 0.0$	AR(1)	**2.73**	**1.40**	**.827**	**-.404**	**-.950**	**-.984**	**-.993**	**-.999**	**-.971**	**-.973**	**-.977**	**-.977**
	Double†	**3.53**	**2.64**	**1.32**	**-.087**	**-.899**	**-.943**	**-.969**	**-.992**	**-.944**	**-.951**	**-.959**	**-.958**
$\sigma^2 = 300, 70$	VC	**-.964**	**-.973**	**-1.00**	**-1.00**	**-.999**	**-1.00**	**-1.00**	**-1.00**	**-.976**	**-.979**	**-.982**	**-.985**
$\rho_{ar(1)} = 0.0$	AR(1)	**-.975**	**-.997**	**-1.00**	**-1.00**	**-.999**	**-1.00**	**-1.00**	**-1.00**	**-.985**	**-.988**	**-.990**	**-.993**
	Double†	**2.53**	**2.00**	**.378**	**-.308**	**-.853**	**-.912**	**-.944**	**-.958**	**-.759**	**-.770**	**-.801**	**-.814**
$\sigma^2 = 150, 150$	VC	**10.1**	**7.31**	**7.51**	**6.07**	**.441**	**.219**	**.179**	**.109**	**-.578**	**-.635**	**-.637**	**-.636**
$\rho_{ar(1)} = 0.4$	AR(1)	**2.69**	**1.26**	**.735**	**-.160**	**-.964**	**-.996**	**-.999**	**-1.00**	**-.998**	**-1.00**	**-1.00**	**-1.00**
	Double†	**12.4**	**10.6**	**10.9**	**10.5**	**1.30**	**.762**	**.628**	**.489**	**-.335**	**-.374**	**-.368**	**-.349**
$\sigma^2 = 300, 70$	VC	**-.699**	**-.819**	**-.957**	**-1.00**	**-.846**	**-.944**	**-.953**	**-.987**	**-.911**	**-.923**	**-.887**	**-.894**
$\rho_{ar(1)} = 0.4$	AR(1)	**-.993**	**-1.00**	**-1.00**	**-1.00**	**-1.00**	**-1.00**	**-1.00**	**-1.00**	**-1.00**	**-1.00**	**-1.00**	**-1.00**
	Double†	**5.97**	**5.53**	**5.33**	**4.27**	**2.27**	**2.10**	**2.01**	**1.53**	**-.322**	**-.307**	**-.289**	**-.306**

Spec. = specification; VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure

†"Double" refers to the inclusion of two error terms – one for each phase

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 18

Relative parameter bias of σ^2_{single} by factor levels level 1 error specification

	Participants per study	3	6	3	6	3	6	3	6	3	6	3	6
	Studies	10	10	30	30	10	10	30	30	10	10	30	30
	Baseline data points	5	5	5	5	10	10	10	10	20	20	20	20
	Treatment data points	5	5	5	5	10	10	10	10	20	20	20	20
Grouping	Level 1 error spec.	Relative parameter bias of $T_{\sigma single}$											
$\sigma^2 = 150, 150$	VC	-.036	-.030	-.033	-.033	**-.086**	**-.089**	**-.091**	**-.087**	**-.148**	**-.147**	**-.147**	**-.145**
$\rho_{ar(1)} = 0.0$	AR(1)	-.023	-.015	-.019	-.019	**-.071**	**-.074**	**-.078**	**-.074**	**-.129**	**-.127**	**-.129**	**-.127**
$\sigma^2 = 300, 70$	VC	**.244**	**.256**	**.250**	**.254**	**.199**	**.199**	**.202**	**.203**	**.179**	**.183**	**.178**	**.185**
$\rho_{ar(1)} = 0.0$	AR(1)	**.277**	**.294**	**.287**	**.293**	**.224**	**.221**	**.223**	**.225**	**.205**	**.208**	**.201**	**.208**
$\sigma^2 = 150, 150$	VC	**-.540**	**-.526**	**-.529**	**-.524**	**-.465**	**-.468**	**-.468**	**-.467**	**-.407**	**-.404**	**-.404**	**-.404**
$\rho_{ar(1)} = 0.4$	AR(1)	**-.173**	**-.174**	**-.179**	**-.186**	-.015	-.014	-.012	-.012	-.004	.003	.013	.013
$\sigma^2 = 300, 70$	VC	**-.354**	**-.320**	**-.321**	**-.304**	**-.284**	**-.282**	**-.283**	**-.282**	**-.146**	**-.141**	**-.149**	**-.149**
$\rho_{ar(1)} = 0.4$	AR(1)	**.222**	**.199**	**.194**	**.165**	**.317**	**.335**	**.322**	**.331**	**.355**	**.355**	**.366**	**.369**

Spec. = specification; VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 19

Relative parameter bias of $\sigma^2_{baseline}$ and $\sigma^2_{treatment}$ by factor levels

	Participants per study	3	6	3	6	3	6	3	6	3	6	3	6
	Studies	10	10	30	30	10	10	30	30	10	10	30	30
	Baseline data points	5	5	5	5	10	10	10	10	20	20	20	20
	Treatment data points	5	5	5	5	10	10	10	10	20	20	20	20
Grouping	Variance parameter	Relative parameter bias statistics											
$\sigma^2 = 150, 150$	$\sigma^2_{baseline}$	.002	.005	-.005	.000	.000	-.004	-.006	-.001	.002	.002	.000	.003
$\rho_{ar(1)} = 0.0$	$\sigma^2_{treatment}$	**-.082**	**-.074**	**-.067**	**-.073**	**-.181**	**-.183**	**-.185**	**-.182**	**-.311**	**-.307**	**-.306**	**-.305**
$\sigma^2 = 300, 70$	$\sigma^2_{baseline}$	-.028	-.024	-.027	-.029	-.039	-.040	-.039	-.040	-.049	-.047	**-.051**	-.046
$\rho_{ar(1)} = 0.0$	$\sigma^2_{treatment}$	**-.051**	-.031	-.042	-.031	**-.093**	**-.097**	**-.085**	**-.083**	**-.229**	**-.222**	**-.222**	**-.214**
$\sigma^2 = 150, 150$	$\sigma^2_{baseline}$	**-.453**	**-.433**	**-.437**	**-.432**	**-.377**	**-.383**	**-.384**	**-.380**	**-.277**	**-.275**	**-.277**	**-.274**
$\rho_{ar(1)} = 0.4$	$\sigma^2_{treatment}$	**-.632**	**-.630**	**-.630**	**-.631**	**-.550**	**-.549**	**-.548**	**-.546**	**-.558**	**-.557**	**-.557**	**-.559**
$\sigma^2 = 300, 70$	$\sigma^2_{baseline}$	**-.438**	**-.424**	**-.424**	**-.418**	**-.354**	**-.343**	**-.346**	**-.339**	**-.286**	**-.282**	**-.290**	**-.290**
$\rho_{ar(1)} = 0.4$	$\sigma^2_{treatment}$	**-.626**	**-.621**	**-.624**	**-.622**	**-.520**	**-.518**	**-.520**	**-.514**	**-.481**	**-.484**	**-.487**	**-.486**

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 20

Relative bias of autocorrelation parameter estimates by condition and generating value for $\rho_{ar(1)}$

$\sigma^2_{baseline}$	$\sigma^2_{treatment}$	Subjects per study	Studies	Baseline data points	Treatment data points	Relative bias	
						$\rho_{ar(1)} = 0.0$	$\rho_{ar(1)} = 0.4$
150	150	3	10	5	5	0.023	**0.327**
150	150	6	10	5	5	0.026	**0.330**
150	150	3	30	5	5	0.028	**0.329**
150	150	6	30	5	5	0.028	**0.332**
150	150	3	10	10	10	0.045	**0.594**
150	150	6	10	10	10	0.048	**0.607**
150	150	3	30	10	10	0.044	**0.610**
150	150	6	30	10	10	0.046	**0.613**
150	150	3	10	20	20	**0.083**	**0.705**
150	150	6	10	20	20	**0.086**	**0.713**
150	150	3	30	20	20	**0.083**	**0.725**
150	150	6	30	20	20	**0.086**	**0.724**
300	70	3	10	5	5	**0.055**	**0.462**
300	70	6	10	5	5	**0.065**	**0.464**
300	70	3	30	5	5	**0.065**	**0.467**
300	70	6	30	5	5	**0.069**	**0.439**
300	70	3	10	10	10	**0.059**	**0.621**
300	70	6	10	10	10	**0.055**	**0.635**
300	70	3	30	10	10	**0.053**	**0.632**
300	70	6	30	10	10	**0.057**	**0.640**
300	70	3	10	20	20	**0.087**	**0.705**
300	70	6	10	20	20	**0.086**	**0.706**
300	70	3	30	20	20	**0.084**	**0.715**
300	70	6	30	20	20	**0.087**	**0.719**

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 21

Percentages of near-zero y-values in treatment phases with 5, 10, and 20 data points, after truncation of y-values

Cut-offs used in calculating percentages	Number of data points per phase		
	5	10	20
$y \leq 10$	12.46%	32.95%	54.32%
$y \leq 5$	7.48%	24.05%	42.57%
$y = 0$	3.82%	15.15%	28.46%

Table 22

Percentages of extremely high y-values in baseline phases with 5, 10, and 20 data points

Cut-offs used in calculating percentages	Number of data points per phase		
	5	10	20
$y \geq 90$	6.42%	7.17%	8.47%
$y \geq 80$	16.28%	16.89%	19.50%

Table 23

Comparison of relative bias of estimates for $T_{\beta 0}$ from two simulation runs

		Participants per study	3	6	3	6
		Studies	10	10	30	30
		Baseline data points	5	5	5	5
		Treatment data points	5	5	5	5
Simulation run	Grouping	Level 1 error spec.	Relative parameter bias of $T_{\beta 0}$			
First	$\sigma = 150, 150$ $\rho_{ar(1)} = 0.0$	VC	**-.143**	-.041	**-.080**	-.034
		AR(1)	**-.163**	**-.061**	**-.098**	-.050
		Double†	**-.154**	-.050	**-.086**	-.042
Second	$\sigma = 150, 150$ $\rho_{ar(1)} = 0.0$	VC	**-.313**	**-.253**	**-.383**	**-.288**
		AR(1)	**-.321**	**-.253**	**-.370**	**-.288**
		Double†	**-.320**	**-.258**	**-.385**	**-.293**

Spec. = specification; VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure

†"Double" refers to the inclusion of two error terms – one for each phase

Bold print indicates relative bias was unacceptable (i.e., > 0.05)

Table 24

Comparison of relative bias of estimates for fixed effects from two simulation runs in which $\sigma^2_{baseline} = 150$, $\sigma^2_{treatment} = 150$, *and* $\rho_{ar(1)} = 0.0$

Participants per study	Studies	Baseline data points	Treatment data points	γ_0			γ_1			γ_2			γ_3			γ_4		
				VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†	VC	AR(1)	Double†
3	10	5	5	-.003	-.003	-.003	**.068**	**.073**	**.068**	-.002	-.001	-.002	.040	.038	.040	**-1.97**	**-1.95**	**-1.97**
6	10	5	5	.000	.001	.000	-.027	-.028	-.027	-.004	-.004	-.004	.031	.030	.031	**-2.02**	**-2.01**	**-2.02**
3	30	5	5	.000	.000	.000	.014	.016	.014	-.001	-.001	-.001	.045	.045	.045	**-2.09**	**-2.09**	**-2.08**
6	30	5	5	.000	.000	.000	-.008	-.006	-.008	-.005	-.005	-.005	.025	.025	.025	**-1.82**	**-1.83**	**-1.82**
3	10	5	5	.000	.000	.000	**.209**	**.193**	**.209**	.009	.009	.009	-.005	-.007	-.005	**-1.35**	**-1.34**	**-1.35**
6	10	5	5	.003	.003	.003	**-.248**	**-.249**	**-.248**	-.013	-.013	-.013	.040	.040	.040	**-2.12**	**-2.12**	**-2.12**
3	30	5	5	-.002	-.002	-.002	.009	.005	.009	-.002	-.002	-.002	.015	.015	.015	**-1.72**	**-1.72**	**-1.72**
6	30	5	5	-.001	-.001	-.001	.048	.048	.048	-.008	-.008	-.008	.050	.050	.050	**-2.05**	**-2.05**	**-2.05**

VC = SAS default of different variance components and 0 covariances; AR(1) = lag-1 autoregressive covariance structure; bold print indicates bias was unacceptable (> 0.05)

†"Double" refers to the inclusion of two error terms – one for each phase

FIGURES

Figure 1

Visual illustration of parameters in Equation 29

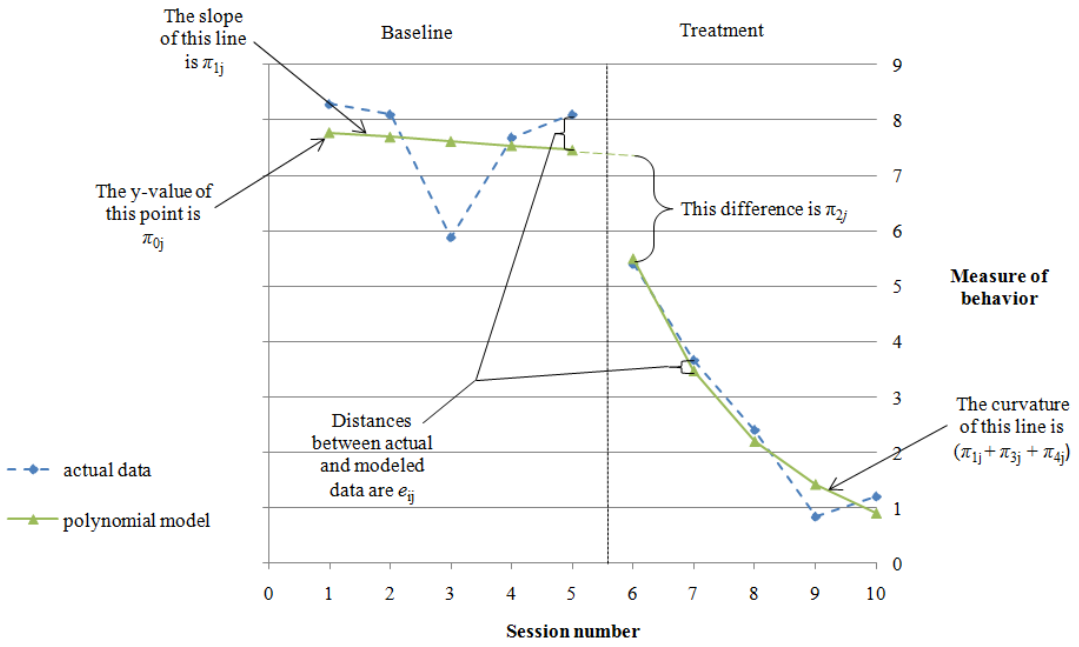


Figure 2

Population average model for data generation

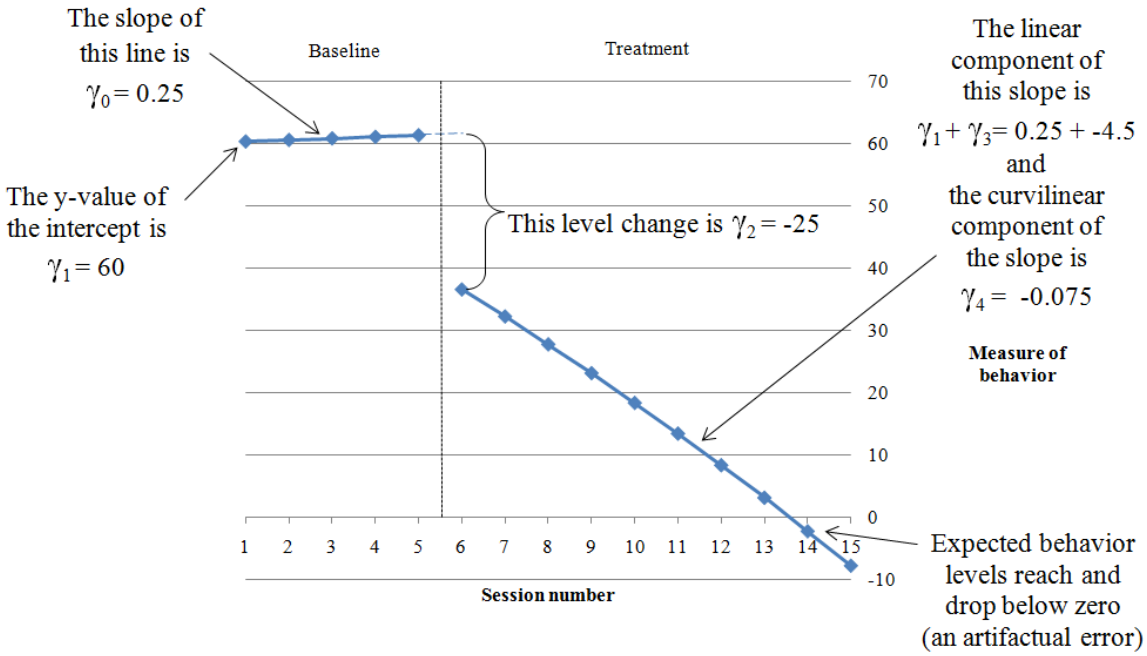


Figure 3

Graphic illustration of potential treatment phase trajectories due to the correlation of π_3 and π_4/ covariance of r_{3jk} and r_{4jk}, and u_{3k} and u_{4k}

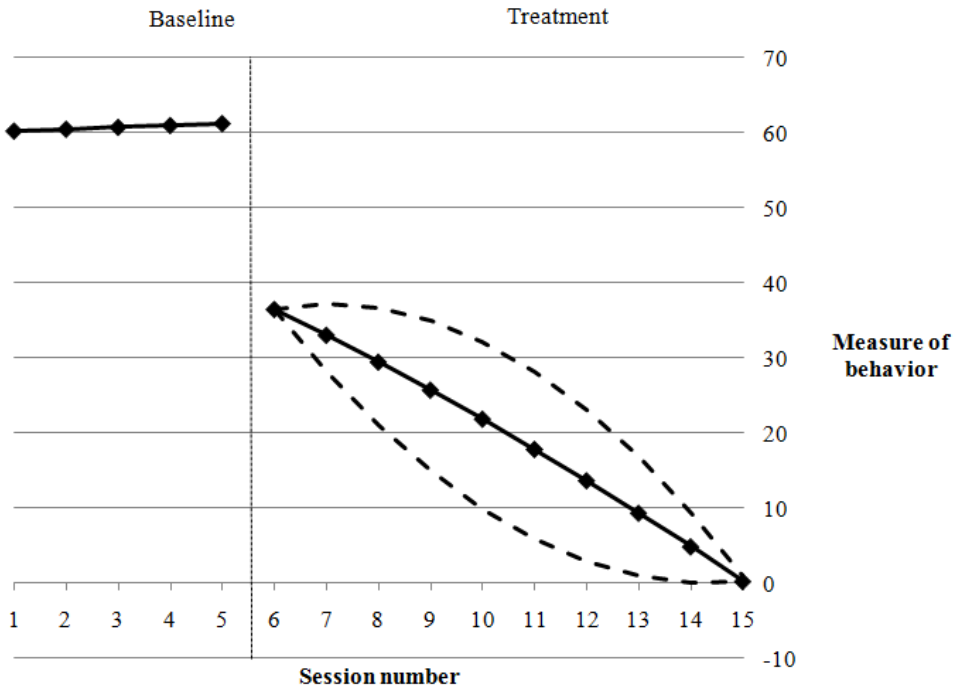


Figure 4

Random sample of 10 simulated data sets with 5 baseline and 5 treatment data points from 2 studies, when $\sigma^2_{baseline} = 300$, $\sigma^2_{treatment} = 70$, *and* $\rho_{ar(1)} = 0.0$

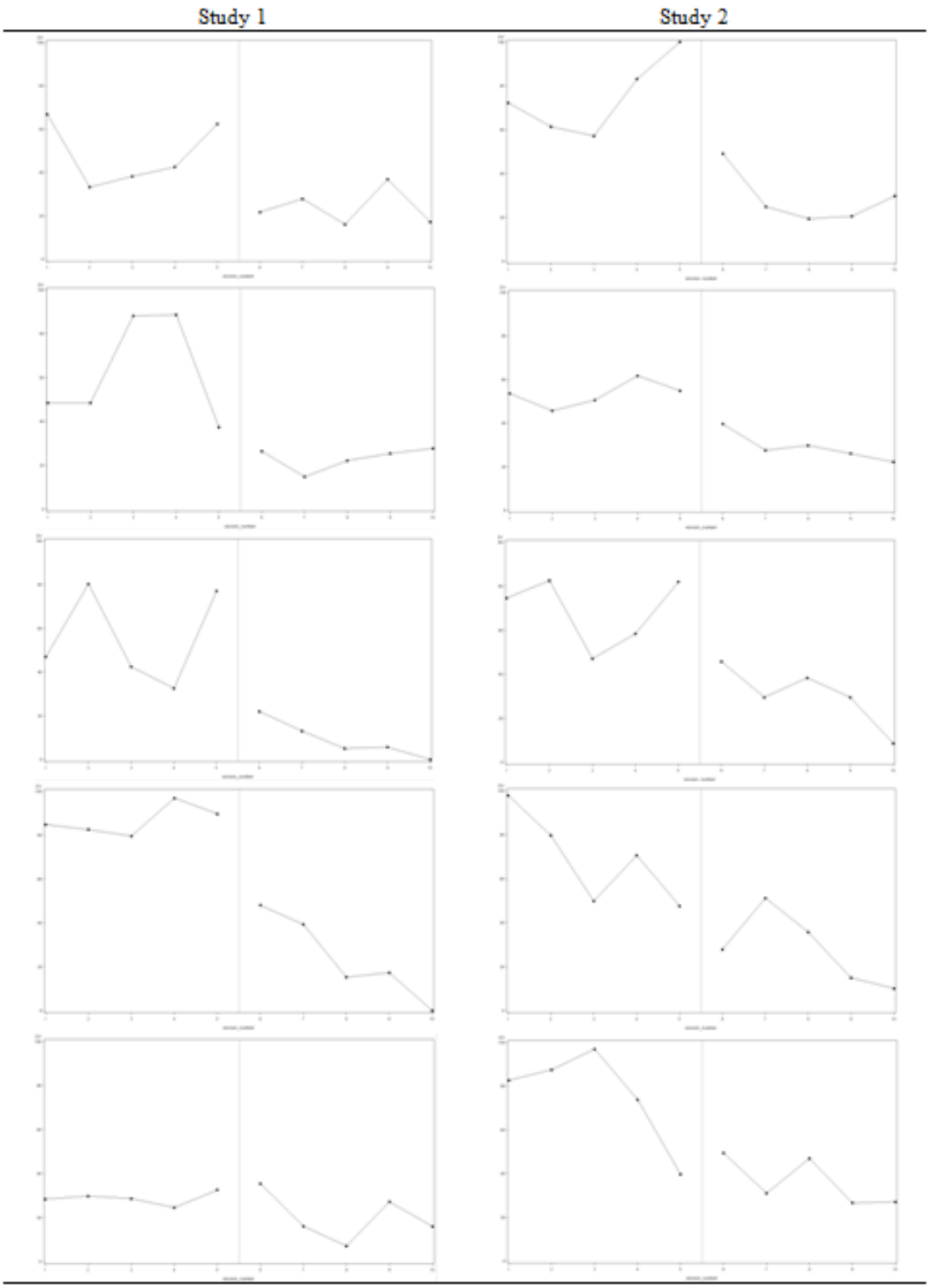


Figure 5

Random sample of 10 simulated data sets with 10 baseline and 10 treatment data points from 2 studies, when $\sigma^2_{baseline} = 300$, $\sigma^2_{treatment} = 70$, *and* $\rho_{ar(1)} = 0.0$

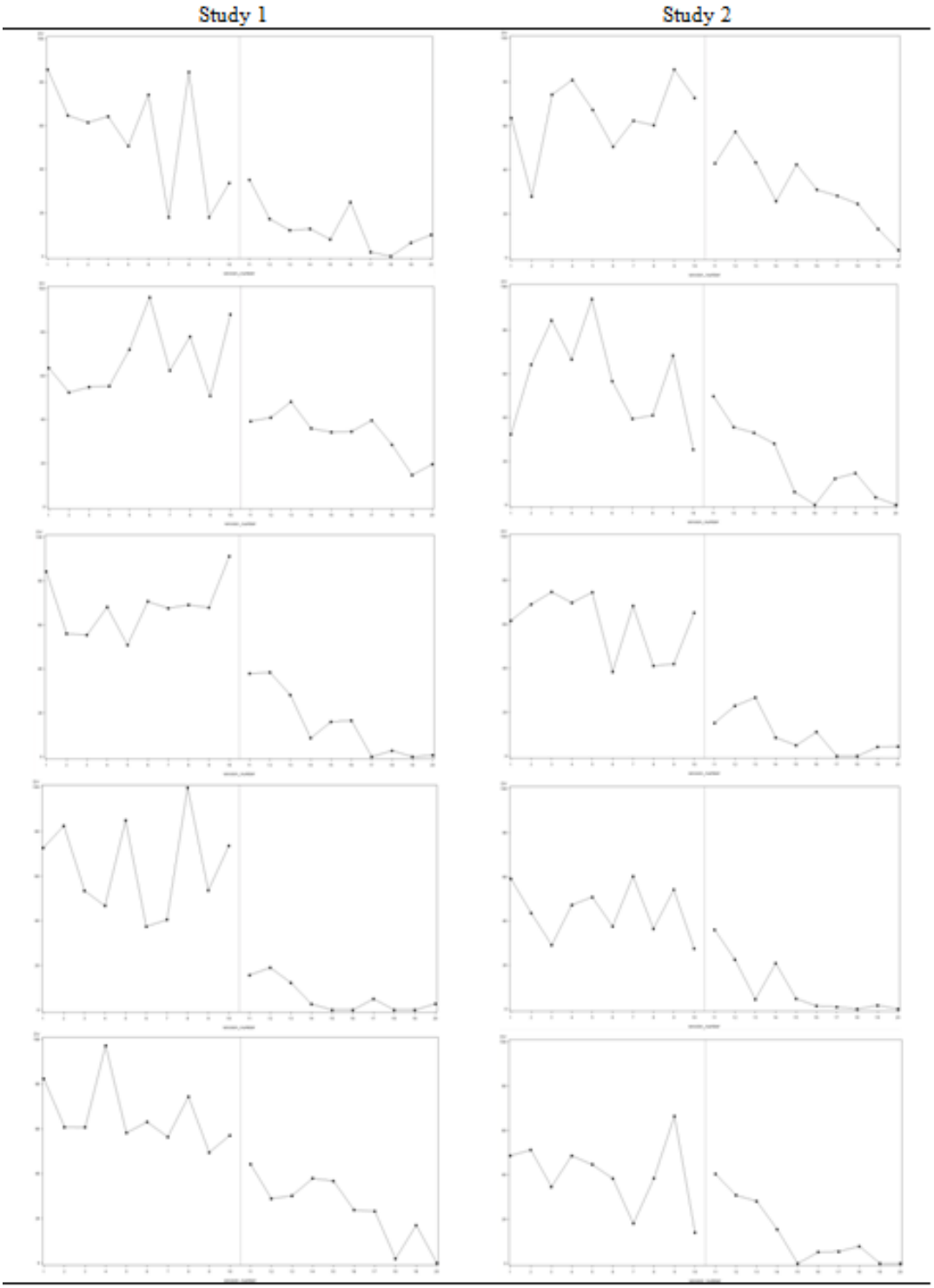


Figure 6

Random sample of 10 simulated data sets with 20 baseline and 20 treatment data points from 2 studies, when $\sigma^2_{baseline} = 300$, $\sigma^2_{treatment} = 70$, *and* $\rho_{ar(1)} = 0.0$

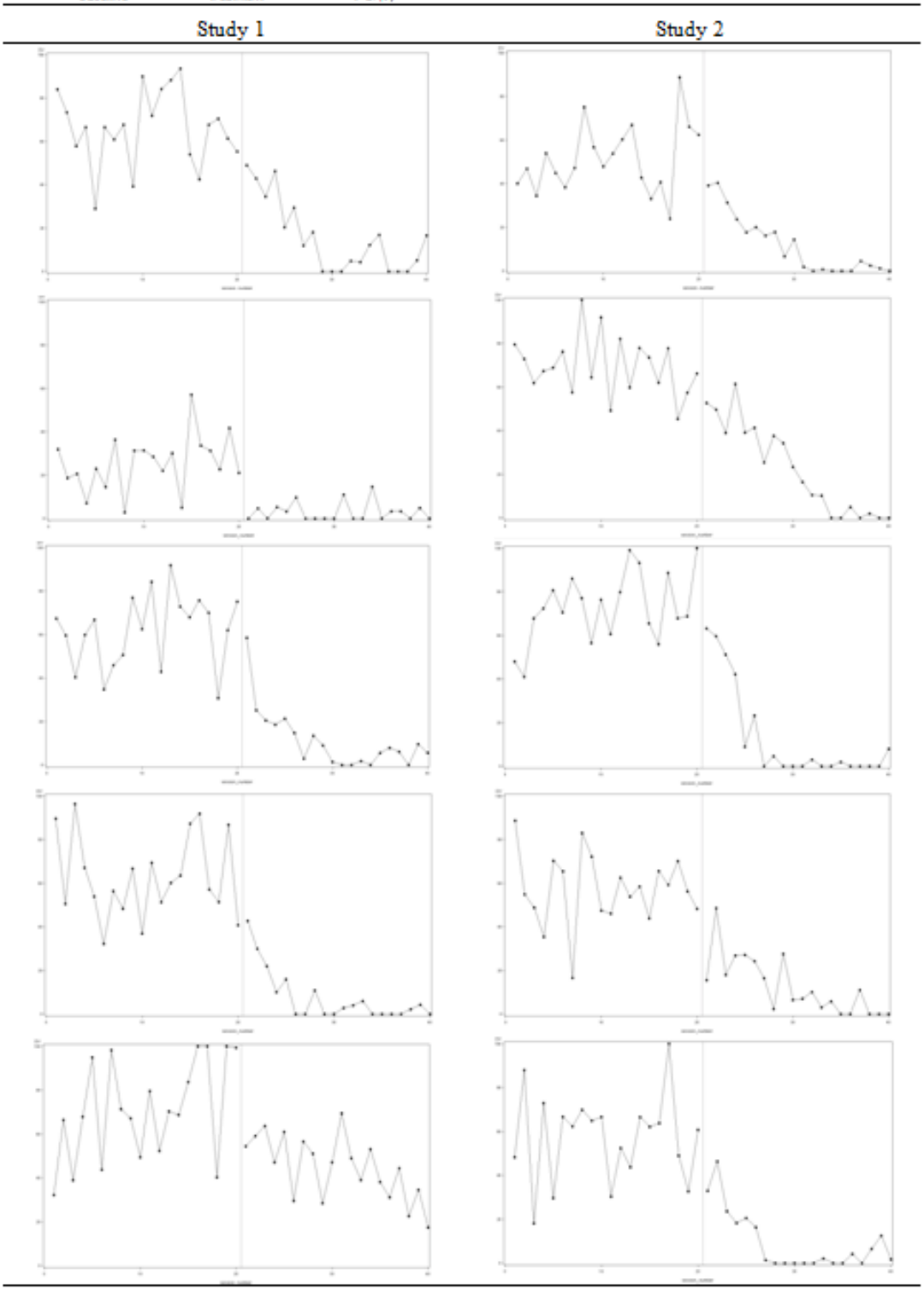
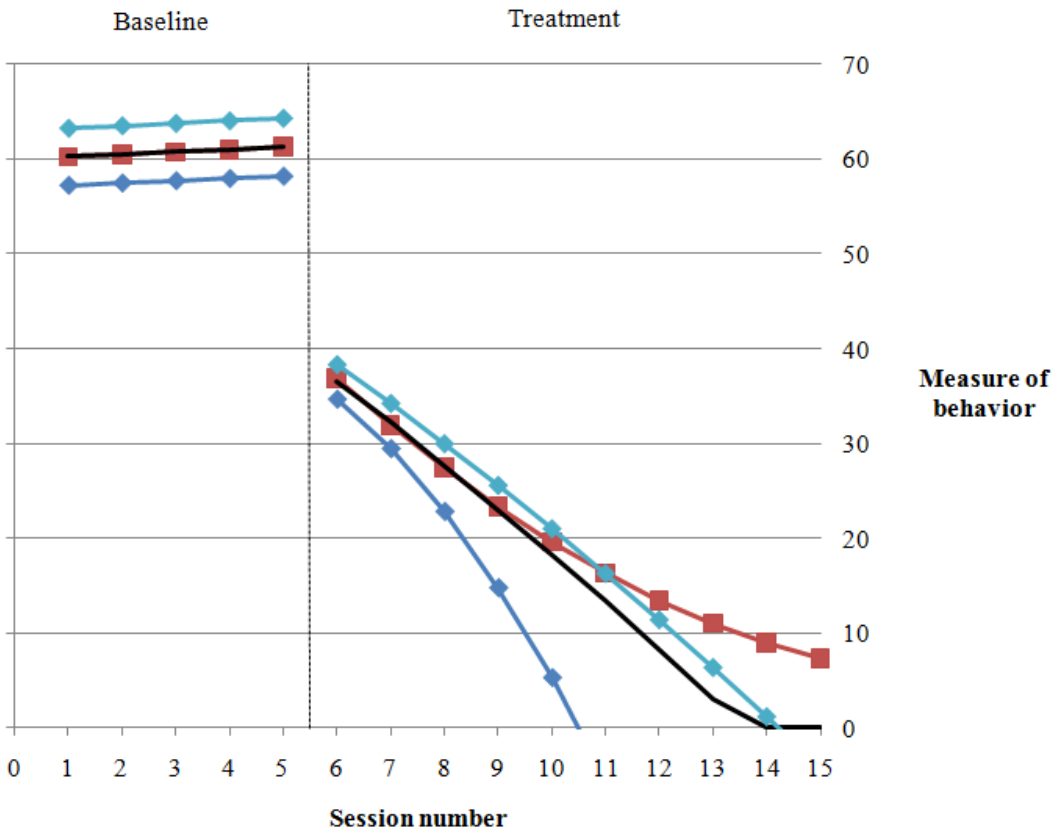


Figure 7

Graphic depiction of the population average model, examples of the limits of acceptable bias, and an extremely biased model observed in the simulation study



Black line represents the population average model; blue line with diamond dots represents one limit of acceptable bias (when the direction of bias is consistent with the bias observed in the simulation study); aqua line with diamond dots represents another limit of acceptable bias (when the direction of bias is opposite the bias observed in the simulation study; and the red line with square dots represents an extremely biased model observed in the simulation study

APPENDIX

The lines of SAS code below can be used to estimate the 3 level models examined in this study.

When specifying different variance components and covariances of 0 for level 1 error, the following code can be used:

```
proc mixed covtest;
class subject study;
model DV=session_number trmt term term2/solution ddfm=kr notest;
random intercept/subject=study;
random intercept session_number trmt term term2/subject=subject(study);
run;
```

where "subject" is the subject identifier; "study" is the study identifier; "DV" is Y_{ijk}; "session_number" is T_{ijk}; "trmt" is $(\text{treatment})_{ijk}$; "term" is $(T_{ijk} - [n_{bjk} + 1])(\text{treatment})_{ijk}$; and term2 is $(T_{ijk} - [n_{bjk} + 1])^2(\text{treatment})_{ijk}$.

When specifying autoregressive covariance structures at level 1, the following code can be used:

```
proc mixed covtest;
class subject study;
model DV=session_number trmt term term2/solution ddfm=kr notest;
random intercept/subject=study;
random intercept session_number trmt term term2/subject=subject(study);
repeated/ type=ar(1) subject=subject(study);
run;
```

When specifying separate error terms for each phase at level 1, the following code can be used:

```
proc mixed covtest;
class subject study trmtcl;
model DV=session_number trmt term term2/solution ddfm=kr notest;
random intercept/subject=study;
random intercept session_number trmt term term2/subject=subject(study);
repeated trmtcl/type=un;
run;
```

where "trmtcl" is identical to "trmt."

To exclude a level 2 random effect, simply delete the variable name from the 5th line of code (e.g., to exclude r_{1jk}, delete "session_number.")

REFERENCES

Adams, M. (2009). *A pedometer-based intervention to increase physical activity: Applying frequent, adaptive goals and a percentile schedule of reinforcement*. Unpublished dissertation, University of California, San Diego.

Allison, D., & Gorman, B. (1994). 'Make things as simple as possible, but no simpler': A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy, 32*(8), 885-890.

Allison, D., &Gorman, B. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 6*, 621-631.

Bell, B., Morgan, G., Zhu, M., & Schoeneberger, J. (2011). *Statistical power and multiple-baseline data: A Monte Carlo examination of alternative multilevel modeling approaches.* Poster presented at the annual meeting of the American Educational Research Association, New Orléans, LA.

Beretvas, N. (2011). *A multilevel model for nonlinear trajectories with smooth phase transitions for multiple baseline design data*. Poster presented at the annual meeting of the American Educational Research Association, New Orléans, LA.

Beretvas, 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*(3), 129-141.

Beretvas, N., & Chung, H. (2008b). An evaluation of modified R^2-change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*(3), 120-128.

Beretvas, N., & Wang, D. (2011). *Synthesis of count data outcome trajectories within studies in single-subjects' experimental design studies*. Paper presented at the annual meeting of the American Educational Research Association, New Orléans, LA.

Biosoft (2004). *UnGraph for Windows* (Version 5.0) [computer software]. Cambridge, U.K.: Author.

Borckardt, J., Nash, M., Murphy, M., Moore, M., Shaw, D., & Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research. *American Psychologist, 63*(2), 77-95.

Brown, W. H., Odom, H. L., & Conroy, M. A. (2001). An intervention hierarchy for promoting young children's peer interactions in natural environments. *Topics in Early Childhood Special Education, 21*(3), 162-175.

Burns, M. K., & Ysseldyke, J. E. (2009). Reported prevalence of evidence-based instructional practices in special education. *The Journal of Special Education, 43*(1), 3-11.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10*(3), 229-242.

Busk, P., & Serlin, R. (1992). Meta-analysis for single-case research. In T. Kratochwill & J. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187 -212). Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Center, B., Skiba, R., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19*, 387-400.

Chan, J. M., Lang, R., Rispoli, M., O'Reilly, M., Sigafoos, J., & Cole, H. (2009). Use of peer-mediated interventions in the treatment of autism spectrum disorders: A systematic review. *Research in Autism Spectrum Disorders, 3*(4), 876-889.

Clark-Carter, D. (2004). *Quantitative psychological research: A student's handbook.* London: Psychology Press.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cooper, H., & Hedges, L. (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Diamond, M., & Hopson, J. (1999). *Magic trees of the mind: How to nurture your child's intelligence, creativity, and healthy emotions from birth through adolescence.* New York: Plume.

Dunst, C. J., Trivette, C. M., & Cutspec, P. A. (2002). Toward an operational definition of evidence-based practice. *Centerscope, 1*, 1-10.

Engelshcall, R. S. (1997). Module mod_rewrite: URL Rewriting Engine. In *Apache HTTP Server Version 1.3 Documentation* (Apache modules). Retrieved from http://httpd.apache.org/docs/1.3/mod/mod_rewrite.html

Eren, R., & Brucker, P. (2011). Practicing evidence-based practices. In B. Reichow, P. Doehring, D. V. Cicchetti, F. R. Volkmar, B. Reichow, P. Doehring, et al. (Eds.), *Evidence-based practices and treatments for children with autism* (pp. 309-341). New York, NY US: Springer Science and Business Media.

Faith, M., Allison, D., & Gorman, B. (1996). Meta-analysis of single-case research. In R. Franklin, D. Allison, B. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments & Computers, 34*, 324-331.

Ferron, J., Bell, B., Hess, M., Rendina-Gobioff, G., & Hibbard, S. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41,* 372-384.

Ferron, J., Farmer, J., & Owens, C. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilevel-modeling approaches. *Behavior Research Methods, 42,* 930-943.

Forness, S. R., Kavale, K. A., Blum, I. M. & Lloyd, J. W. (1997). Mega-analysis of meta-analysis: What works in special education and related services? *Teaching Exceptional Children, 29*(6), 4–9.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality Indicators for Group Experimental and Quasi-Experimental Research in Special Education. *Exceptional Children,* 71(2), 149-164.

Gersten, R., Schiller, E. P., & Vaughn, S. (2000). *Contemporary special education research: Synthesis of the knowledge base on critical instructional issues.* Mahwah, NJ: Lawrence Erlbaum Associates.

Green, V. A. (2007). Parental experience with treatments for autism. *Journal of Developmental and Physical Disabilities, 19*(2), 91-101.

Green, V. A., Pituch, K. A., Itchon, J., Choi, A., O'Reilly, M., & Sigafoos, J. (2006). Internet survey of treatments used by parents of children with autism. *Research in Developmental Disabilities, 27*(1), 70-84.

Greene, W. (1990). *Econometric analysis* (2nd ed). New York: Macmillan.

Greenwood, C. R. (2001). Bridging the gap between research and practice in special education: issues and implications for teacher preparation. *Teacher Education and Special Education, 24*(4), 273-275.

Guralnick, M. J. (2004). Effectiveness of early intervention for vulnerable children: A developmental perspective. In M.A. Feldman (Ed.), *Early intervention: The essential readings* (pp. 9-50). Oxford, United Kingdom: Blackwell Publishing.

Guyatt,G., Cairns, J., Churchill, D., Cook, D., Haynes, B., Hirsh, B., et al. (1992). Evidence-based medicine: A new approach to teaching the practice of medicine. *The Journal of the American Medical Association, 268*(17), 2420-2425.

Harvey, A. (1990). *The econometric analysis of time series* (2nd ed.). Cambridge: MIT Press.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education. *Exceptional Children*, 71(2), 165-179.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

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*, 38–58.

Hurwitz, J. (2008). *Magnitude and variability of problem-solving consultation outcomes.* Unpublished dissertation, University of Wisconsin, Madison.

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108-446, 118 Stat. 2647

Jacobson, J. W., Foxx, R. M., & Mulick, J. A. (2005). *Controversial therapies for developmental disabilities: Fad, fashion, and science in professional practice*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston: Allyn & Bacon.

Kratochwill, T. R., & Stoiber, K. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the Procedural and Coding Manual of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly, 17*(4), 341-389.

Kutash, K., Duchnowski, A. J., & Lynn, N. (2009). The use of evidence-based instructional strategies in special education settings in secondary schools: Development, implementation and outcomes. *Teaching and Teacher Education,* 25(6), 917-923.

Kwok, O., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research, 42*(3), 557-592.

Lundervold, D., & Bourland, G. (1988). Quantitative analysis of treatment of aggression, self-injury, and property destruction. *Behavior Modification, 12*(4), 590-617.

Lundervold, D., & Bourland, G. (1988). Quantitative analysis of treatment of aggression, self-injury, and property destruction. *Behavior Modification, 12*(4), 590-617.

Marquis, J., Horner, R., & Carr, E. (2000). A meta-analysis of positive behavior support. In R. Gersten, E. Schiller, & S. Vaughn (Eds.), *Contemporary special education research: Syntheses of the knowledge base on critical instructional issues* (pp. 137 -178). Mahwah, NJ, US: Lawrence Erlbaum Associates.

Mayton, M. R., Wheeler, J. J., Menendez, A. L., & Zhang, J. (2010). An analysis of evidence-based practices in the education and treatment of learners with autism spectrum disorders. *Education and Training in Autism and Developmental Disabilities,* 45(4), 539-551.

McCain, M. N. & Mustard, J. F. (1999) *Reversing the real brain drain: Early years study – Final report.* Toronto: Publications Ontario.

Microsoft (2007). *Microsoft Excel* [computer software]. Redmond, Washington: Author.

Miller, L. (2006). *Interventions targeting reciprocal social interaction in children and young adults with autism spectrum disorders: A meta-analysis.* Unpublished dissertation, University of Utah.

Morgan, P. &Sideridis, G. (2006). Contrasting the effectiveness of fluency interventions for students with or at risk for learning disabilities: A multilevel random coefficient modeling meta-analysis. *Learning Disabilities Research & Practice*, *21*, 191-210.

National Autism Center (2009). *National Standards Report.* Randolph, MA: Author.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Odom, S. L. & Worley, M. (2003). A unified theory of practice in early intervention/early childhood special education: Evidence-based practices. *Journal of Special Education, 37*, 164-173.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in Special Education: Scientific Methods and Evidence-Based Practices. *Exceptional Children*, 71(2), 137-148.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357-367.

Pucketts Institute. (2009). *Goals of the Center for Evidence-Based Practice,* Retrieved February 3, 2011 from http://www.evidencebasedpractices.org/goals.php

Raudenbush, S. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185.

Raudenbush, S., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213.

Raudenbush, S., Bryk, A., Cheong, Y., & Congdon, R. (2004). *HLM 6: Hierarchical Linear and Nonlinear Modeling.* Lincolnwood, IL: Scientific Software International.

Raudenbush, S.,& Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods.* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Richman, D. M., Reese, R., & Daniels, D. (1999). Use of evidence-based practice as a method for evaluating the effects of secretin on a child with autism. *Focus on Autism and Other Developmental Disabilities*, *14*(4), 204-211.

Riviello, C. & Beretvas, S. (2009). Detecting lag-1 autocorrelation in interrupted time series designs with small sample sizes. *Journal of Modern Applied Statistical Methods, 8, 469-477.*

Salzberg, C., Strain, P., & Baer, D. (1987). Meta-analysis for single-subject research: When does it clarify, when does it obscure? *RASE: Remedial & Special Education, 8*(2), 43-48.

SAS Institute Inc. (2008). *SAS* (Version 9.2) [computer software]. Cary, NC: Author.

Savin, N.E., & White, K. J. (1977). The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors. *Econometrica, 45*, 1989-1996.

Schreiber, C. (2011). Social Skills Interventions for Children with High-Functioning Autism Spectrum Disorders. *Journal of Positive Behavior Interventions, 13*(1), 49-62.

Scotti, J., Evans, I., & Meyer, L. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation*, *96*(3), 233-256.

Scruggs, T., Mastropieri, M., & Casto, G. (1987). The quantitative synthesis of single-subject research: methodology and validation. *RASE: Remedial & Special Education, 8*(2), 24-33.

Shadish, W. R., Rindskopf, D. M. & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*(3), 188-196.

Shogren, K., Faggella-Luby, M., Bae, S., & Wehmeyer, M. (2004). The effect of choice-making as an intervention for problem behavior: A meta-analysis. *Journal of Positive Behavior Interventions*, *6*(4), 228-237.

Shonkoff, J. P. & Phillips, D. A. (2000). *From neurons to neighborhoods: The science of early childhood development.* Washington, DC: National Academies Press.

Singer, J. D. & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. New York, NY: Oxford University Press.

Singh, S., & Ernst, E. (2008). *Trick or treatment: The undeniable facts about alternative medicine.* New York, NY: W. W. Norton.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21.

Smith, B. J., Strain, P. S., Snyder, P, Sandall, S. R., McLean, M. E., Boudy-Ramsey, A., et al. (2002). DEC recommended practices: A review of 9 years of EI/ECSE research literature. *Journal of Early Intervention,25,* 108-119.

Spybrook, J., Raudenbush, S.W., Congdon, R., & Martinez, A. (2009). *Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software V.2.0*. Available at www.wtgrantfoundation.org

Terrazas Arellanes, F. (2009). *The effects of the 'templates' for direct and explicit Spanish instruction on English language learners reading outcomes.* Unpublished dissertation, University of Oregon.

Turner, W., Realon, R., Irvin, D., & Robinson, E. (1996). The effects of implementing program consequences with a group of individuals who engaged in sensory maintained hand mouthing. *Research in Developmental Disabilities, 17*(4), 311-330.

Van den Noortgate, W. & Onghena, P. (2003a). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments & Computers, 35*, 1-10.

Van den Noortgate, W. & Onghena, P. (2003b). Combining Single-Case Experimental Data Using Hierarchical Linear Models. *School Psychology Quarterly, 18*, 325-346.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2*(3), 142-151.

Van den Noortgate, W., & Onghena, P. (2011). *A multivariate meta-analysis of regression coefficients from single-subject experimental designs*. Poster presented at the annual meeting of the American Educational Research Association, New Orléans, LA.

Verbeke, G. & Molenberghs, G. (1997). *Linear mixed models in practice: a SAS-oriented approach.* New York: Springer.

Wade, C. M., Ortiz, C., & Gorman, B. S. (2007). Two-session group parent training for bedtime noncompliance in head start preschoolers. *Child & Family Behavior Therapy, 29*(3), 23-55.

Wang, S., Cui, Y., & Parrila, R. (2011). Examining the effectiveness of peer-mediated and video-modeling social skills interventions for children with autism spectrum disorders: A meta-

analysis in single-case research using HLM. *Research in Autism Spectrum Disorders, 5*(1), 562-569.

White, O. (1987). ‘The quantitative synthesis of single-subject research: Methodology and validation': Comment. *RASE: Remedial & Special Education, 8*(2), 34-39.