1-1-2015

Consequences of Non-Modeled and Modeled Between Case Variation in the Level-1 Error Structure in Multilevel Models for Single-Case Data: A Monte Carlo Study

Eun Kyeng Baek
University of South Florida, ebaek@usf.edu

Consequences of Non-Modeled and Modeled Between Case Variation in the Level-1 Error Structure in Multilevel Models for Single-Case Data: A Monte Carlo Study

by

Eun Kyeng Baek

A dissertation submitted in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
in Curriculum and Instruction with an emphasis in
Measurement and Evaluation
Department of Educational & Psychological Studies
College of Education
University of South Florida

Major Professor: John Ferron, Ph.D.
Jeffrey Kromrey, Ph.D.
Eun Sook Kim, Ph.D.
Danielle Dennis, Ph.D.

Date of Approval:
March 27, 2015

Keywords: Hierarchical, Multilevel, Single-case study, Bayesian estimation, Level-1 error structure

DEDICATION

This dissertation is dedicated in unconditional love and support from my parents, Namsoon Lee and Yonggon Baek and my brother, Hyoin Baek who lives in my missed home country, Korea, and my husband, Sungyub Han, and my two-year-old son, Chaeyun (Jayden) Han who lives in here, Tampa, with me. This dissertation could not have accomplished without their belief in me and love. Thank you all for your love, support, and belief in me.

사랑하는 엄마, 아빠, 효인이,
언제나 믿어주고 응원해줘서 고마워요.
그 믿음 덕분에 무사히 박사학위 받게 됐네요.
앞으로 조금이나마 그 믿음과 사랑에 보답할수 있기를 바랍니다.

사랑하는 남편 ,우리 윤이,
늘 옆에서 힘이 되어줘서 고마워요.
우리 윤이, 엄마 아빠가 바빠서 늘 함께해주지 못했는데도 잘커줘서 너무 고마워.
우리가족 앞으로 더 사랑하며 살자.

마지막으로 어머님 아버님,
늘 이해해주시고 기다려주셔서 감사합니다.

ACKNOWLEDGMENTS

I could not have completed this fun journey of doctoral studies without supports from my professors and colleagues. I would like to thank my major professor, Dr. John Ferron. He is greatest mentor and professor that I have ever met. He had well prepared me to accomplish this dissertation by constantly guiding and teaching me throughout entire my doctoral studies. He always encouraged me to going forward and always believed in me. His belief in me heartened me a lot that allows me to believe myself and my own potential. I could not imagine anyone else who is better than him as a major professor.

I also want to thank you for other members of my doctoral committee. Dr. Jeffrey Kromrey, who challenged me to learn constantly, Dr. Eun Sook Kim, who provided me a great insight of this dissertation, Dr. Danielle Dennis, who allowed me think from a different angle of applied researchers.

In addition, I would like to thank you the rest of professors in our department, Dr. Liliana Rodriguez-Campos, Dr. Robert Dedrick, Dr. Yi-Hsin Chen, and Dr. Jeniffer Wolgemuth who always inspires and cares students in our program, our office manager, Jody Duke who is always so helpful and kind to students, and my fellow graduate students, Merlande Petit-Bois, Aarti Bellara, Diep Nguyen, Thanh Pham, Chunhua Cao, Connie Walker, and Tyler Hicks who have been going through this journey with me and who will be continued to accomplish their journey.

Lastly, I have truly enjoyed this journey because of all of people that I mentioned. I believe that it was my luck to meet all these great people in my doctoral program, so I thank you for my luck.

Do not dwell in the past, do not dream of the future, concentrate the mind on the present moment.

- Buddha -

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The Multilevel modeling (MLM) approach has a great flexibility in that can handle various methodological issues that may arise with single-case studies, such as the need to model possible dependency in the errors, linear or nonlinear trends, and count outcomes (e.g.,Van den Noortgate & Onghena, 2003a). By using the MLM framework, researchers can not only model dependency in the errors but also model a variety of level-1error structures.

The effect of misspecification in the level-1 error structure has been well studied for MLM analyses. Generally, it was found that the estimates of the fixed effects were unbiased but the estimates of variance parameters were substantially biased when level-1 error structure was misspecified. However, in previous misspecification studies as well as applied studies of multilevel models with single-case data, a critical assumption has been made. Researchers generally assumed that the level-1 error structure is constant across all participants.

It is possible that the level-1 error structure may not be same across participants. Previous studies show that there is a possibility that the level-1 error structure may not be same across participants (Baek & Ferron, 2011; Baek & Ferron, 2013; Maggin et al., 2011). If much variation in level-1 error structure exists, this can possibly impact estimation of the fixed effects and random effects. Despite the importance of this issue, the effects of modeling between-case variation in the level-1 error structure had not yet been systematically studied. The purpose of this simulation study was to extend the MLM modeling in growth curve models to allow the level-1 error structure to vary across cases, and to identify the consequences of modeling and not modeling between-case variation in the level-1 error structure for single-case studies.

A Monte Carlo simulation was conducted that examined conditions that varied in series length per case (10 or 20), the number of cases (4 or 8), the true level-1 errors structure (homogenous, moderately heterogeneous, severely heterogeneous), the level-2 error variance in baseline slope and shift in slope (.05 or .2 times the level-1 variance), and the method to analyze the data (allow level-1 error variance and autocorrelation to vary across cases (Model 2) or not allow level-1 error variance and autocorrelation to vary across cases (Model 1)). All simulated data sets were analyzed using Bayesian estimation. For each condition, 1000 data were simulated, and bias, RMSE and credible interval (CI) coverage and width were examined for the fixed treatment effects and the variance components.

The results of this study found that the different modeling methods in level-1 error structure had little to no impact on the estimates of the fixed treatment effects, but substantial impacts on the estimates of the variance components, especially the level-1 error standard deviation and the autocorrelation parameters. Modeling between case variation in the level-1 error structure (Model 2) performs relatively better than not modeling between case variation in the level-1 error structure (Model 1) for the estimates of the level-1 error standard deviation and the autocorrelation parameters. It was found that as degree of the heterogeneity in the data (i.e., homogeneous, moderately heterogeneous, severely heterogeneous) increased, the effectiveness of Model 2 increased.

The results also indicated that whether the level-1 error structure was under-specified, over-specified, or correctly-specified had little to no impact on the estimates of the fixed treatment effects, but a substantial impact on the level-1 error standard deviation and the autocorrelation. While the correctly-specified and the over-specified models perform fairly well, the under-specified model performs poorly.

Moreover, it was revealed that the form of heterogeneity in the data (i.e., one extreme case versus a more even spread of the level-1 variances) might have some impact on relative effectiveness of the two models, but the degree of the autocorrelation had little to no impact on the relative performance of the two models.

CHAPTER ONE: INTRODUCTION

Single-case research measures an outcome repeatedly for a single case or small samples which allow researchers to fully explore treatment effects (Kazdin, 2011). There is growing interest in single-case designs due to many advantages that these designs offer. For example, single-case designs provide information about not only the treatment effect for each individual, but also individual variations in the treatment effect (Barlow, Nock, & Hersen, 2009), and they also allow researchers to study population groups that have a low prevalence rate (Van den Noortgate & Onghena, 2003a). In addition, using single-case designs allows practitioners to implement research in their own setting which reduces the gap between research and practice (Morgan & Morgan, 2001). There are a variety of single-case designs that are commonly used (Kazdin, 2011; Shadish & Sullivan, 2011). In single-case designs, data are obtained before implementing intervention (baseline phase) and after implementing intervention (treatment phase). AB design is the most basic design that has a baseline phase and a treatment phase. The additional designs include an extension of this design, such as an ABAB design that has more phases for removal of the treatment and reintroduction of the treatment. There are other alternative designs that are commonly used, such as the multiple baseline design that can be used to study several cases at the same time.

Many methods have been developed to analyze single-case data. Traditionally, several non-parametric and statistical methods have been proposed to analyze single-case data (e.g., visual analysis, nonoverlap statistics, and randomization tests); and more recently, regression based methods have been developed. Regression based analyses include single-level analyses, such as ordinary least squares (OLS) and generalized least squares (GLS) regression, and multi-level analysis, such as multilevel modeling (MLM).

Generally, in single-case studies, the errors are considered to be autocorrelated as opposed to independent. It has been found that misspecification issues could arise if the possible dependency of errors is not taken into account in the statistical model. It was found that positive autocorrelation inflates Type I error rates in significance tests of the treatment effect when autocorrelation is not taken into account (Matyas & Greenwood, 1990; Toothaker, Banz, Noble, Camp, & Davis, 1983). For example, in the regression based models, the regression coefficients are unbiased, but the standard errors of the regression coefficients would be underestimated, which leads to confidence intervals that are too small (Neter, Wasserman, & Kutner, 1990). Specifically, for the multilevel models, many researchers found that when level-1 errors are assumed to be independent, it may bias the estimation of the standard errors of the fixed effects and estimation of the random effects (Ferron, Dailey, & Yi, 2002; Kwok, West, & Green, 2007; Sivo, Fan & Witta, 2005; Sivo & Willson, 2000).

There are several methods available which take autocorrelation into consideration. Particularly, the GLS regression method and multilevel modeling can take autocorrelation into consideration (Mcknight & Huitema, 2000; Maggin et al., 2011). However, studies have demonstrated that GLS methods still produce high Type I error rates when applied to small samples (e.g., Johnston, 1984; Huitema & Mckean, 1991; Solanas, Manolov, & Sierra, 2010). Multilevel modeling is flexible for handling dependency of errors in that researchers are able to model various dependent error structures and complex models (e.g., heterogeneity of variance, and the nesting of cases within studies).

There are several estimation methods available to run multilevel analysis of single-case data, including restricted maximum likelihood (REML) and Bayesian methods. The REML method is the most commonly used method to analyze multilevel models, and has been implemented by several software procedures that allow easy access. However, the REML has inferential and technical issues associated with analyzing complex multilevel models of single-case data such as non-convergence with more complex models (Baek, Petit-Bois, & Ferron, 2012). The Bayesian method has the potential to resolve the issue with REML. It was found that a complex multilevel model that fails to converge using REML can be run by using the Bayesian approach (Baek, Petit-Bois, & Ferron, 2013). Studies in multilevel research have also found that Bayesian methods have potential benefits over likelihood methods in that the Bayesian approach could perform as well or better regarding bias, efficiency, and coverage (Browne, 2008; Baldwin & Fellingham, 2013), and provide more accurate results in cases using small samples or unequal sample sizes per subject (Shadish, Kyse, & Rindskopf, 2013).

Problem Statement

Although single-case researchers have recognized the misspecification effect of level-1 error structures on statistical inferences of multilevel models, researchers have overlooked how they have made a critical assumption in their studies. They have generally assumed that the level-1 error structure is constant across all cases. Past applications of multilevel modeling to single-case data (e.g., Van Noortgate & Onghena, 2003a, 2003b) as well as methodological studies of multilevel models with single-case data (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens, 2010) have assumed the level-1 error structure is the same for all cases. It is possible that the error structure may not be same across cases (Baek &

Ferron, 2011; Baek & Ferron, 2013). If great variation exists in the level-1 error structure, and it is not taken into account, this can possibly impact the inferences of a study. Thus, it is important to examine the consequences of not modeling and modeling between case variation in the level-1 error structure. Despite the importance of this issue, neither the effects of non-modeled between case variation nor the performance of modeled between case variation in the level-1 error structure have been systematically examined.

Purpose of the Study

The purpose of this simulation study is to extend the MLM modeling in single-case design to allow between case variation in the level-1 error structure which allows the level-1 error and autocorrelation to vary across cases, and to identify the consequences of not modeling and modeling between case variation in the level-1 error structure for single-case studies using Bayesian estimation. Specifically, two level models where the level-1 error structures are modeled different ways (i.e., not modeling between case variation vs. modeling between case variation) will be examined in terms of the accuracy of estimates of parameters. More specifically, credible interval coverage rates, credible interval widths, the bias of the point estimates, and the root mean squared error (RMSE) will be investigated as functions of specific design (number of cases and series length per case), and data factors (true level-1 error structure, average level of autocorrelation, and variance of level-2 error). The following research questions are of interest:

Research Questions

1. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **fixed treatment effect** in single-case design?
    1) to what extent are the ***bias and RMSE for the fixed treatment effects*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation at the level-2 error)?
    2) to what extent are the ***credible interval coverage and width for the fixed treatment effects*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation at the level-2 error)?
2. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **variance components** in single-case design?
    1) to what extent are the ***bias and RMSE for the variance components*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation at the level-2 error)?
    2) to what extent are the ***credible interval coverage and width for the variance components*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation at the level-2 error)?

Overview of the Study

Monte Carlo simulation methods will be used to address the impact of modeling and not modeling between case variation in the level-1 error structure on inferences of two-level multilevel single-case study using the Bayesian estimation approach. In the study, multiple data, design and analysis factors will be manipulated. The data factors include three factors. These are (a) true level-1 error structure (homogeneous, moderately heterogeneous, severely heterogeneous); (b) variation in the level-2 errors (most of the variance at level-1 and most of the variance at level-2). More specifically for the true level-1 error structure, the data set will be generated in two ways where the level-1 error structure is constant across cases, referred to as the Homogeneous error structure, and where the level-1 error structure is varying across cases, referred to as the Moderate or the Severe heterogeneous error structure, depending on the degree of severity in the generated data sets. There are two factors included in the design factors. These factors are (a) number of cases (4 and 8); (b) series length per case (10 and 20). The analysis factor addresses how to model the level-1 error structure (not modeling between case variation (Model 1), and modeling between case variation (Model 2)) to analyze the Homogeneous, the Moderate or the Severe heterogeneous error structures. Crossing all the factors creates a total of 48 conditions (see Table 1). The impact of the inferences will be made from the 95% credible interval coverage, width, and the RMSE as well as the bias of point estimates.

Significance of the Study

This dissertation provides insights about how non- modeled and modeled between case variation in level-1 error structure, a misspecification issue of the level-1 error structure, impacts statistical inferences, an issue which has not been systematically explored. It could possibly influence the precision of estimation and the efficiency of inferences on single-case data. This study also provides a way to model between case variation in level-1 error structure using WinBUGS, making these created codes accessible to applied researchers for use in their own research.

Table 1
Study design

			True level-1 error structure					
			Homogeneous		Moderately heterogeneous		Severely heterogeneous	
			Method to modeling the level-1 error structure					
Number of cases	Series length per case	Error variance (Most of variance at)	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
4	10	Level-1						
		Level-2						
	20	Level-1						
		Level-2						
8	10	Level-1						
		Level-2						
	20	Level-1						
		Level-2						

Limitations

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

Definitions of Terms

Autocorrelation. The degree to which errors from repeated measured data are correlated with each other (dependency of the errors).

Bayesian estimation. A practical method for analyzing multilevel modeling that is known to take into account the uncertainty of estimating both fixed effect and variance components by using constructed prior distributions. Bayesian inference is the process of fitting a probability model, given the observed data, and summarizing uncertainty of parameters by a probability distribution (Gelman, Carlin, Stern, & Rubin, 2004).

Bias. The difference between a known parameter value (true value) and an estimated parameter value.

Credible interval. Known as Bayesian confidence interval that is corresponding to the confidence interval in general statistics.

Credible interval coverage. The proportion of 95% credible intervals that contain a true value for the estimated parameter.

Credible interval width. The difference between the upper and lower limits of the 95% credible intervals for the estimated parameter.

Fixed effects. Parameters that estimate average effects (e.g., average intercept, average treatment effect) that are represented by regression coefficients in the multilevel model.

Hyperparameters. Parameters of prior distributions, not the direct parameters of the model.

Level-1 error. The difference between the observed values and predicted values of an outcome in a case in multilevel single-case designs.

Level-1 error structure. A variance and covariance structure among the level-1errors..

Multilevel modeling (MLM). A statistical model that accounts for nested data (e.g., students in classrooms, repeated observations of students) or more than one level of the parameters. It is also known as hierarchical linear modeling or random effects modeling.

Prior distribution. A probability distribution represents the approximation about an unknown parameter that is believed prior to observing the specific data.

Restricted maximum likelihood estimation (REML). A traditional estimation method to analyze multilevel modeling. The rationale behind likelihood estimation is that the best way to estimate a parameter is to find the value that allows the observed data most likely to have occurred (Fienberg &Linden, 1997).

Root Mean Squared Error (RMSE). The square root of the average squares of the errors.

Series length. The level-1 sample size in the multilevel model, or the number of observations of a case in a single-case study.

Single-case research. The intensive study that repeatedly measures a single case or small samples to determine the effectiveness of one or more treatments.

Treatment effect. The change in a dependent variable attributed to a specific treatment.

Variance components. Parameters that estimate variation within cases and between cases.

CHAPTER TWO: LITERATURE REVIEW

This literature review will be divided into four parts. First, single-case studies are introduced, and a brief overview of the design and analysis is given. Next, level-1 error structures and the effects of misspecification in level-1 error structures are described. Third, a typical assumption that the level-1 error structure is constant across cases is addressed. Finally, a method to model between case variation in level-1 error structures is suggested.

Single-Case Studies

Single-case research focuses on studying changes in an outcome over time. By measuring an outcome repeatedly through time, single-case studies allow the direct study of changes within individuals and the factors that influence changes. However, unlike other forms of longitudinal research that gathers information from relatively large samples (> 30; Hox, 1998), single-case research focuses on the study of a single case or small samples and its growth over time. Thus single-case research can be defined as a study that repeatedly measures a single case over time to examine the effectiveness of treatments (Kazdin, 2011).

In single-case designs, observations are obtained during at least two phases, one baseline phase and one treatment phase. Phase is an important feature of the single-case design. When the observations of the outcome occur before a treatment, it refers to a baseline phase. When the observations of the outcome occur after a treatment, it refers to a treatment phase. By comparing outcome scores from both phases, single-case researchers can evaluate changes in the outcome scores after introducing the treatment (Onghena & Edgington, 2005).

Interest in single-case designs has been growing in many areas of research, including psychology, education, social science, counseling, and other disciplines (Barlow, Nock, & Hersen, 2009; Franklin, Allison, & Gorman, 1997; Ittenbach & Lawhead, 1997; Kazdin, 2011; Kratochwill, 1985; Wacker, Steege, & Berg, 1988) because they have several advantages over other designs. Single-case design allows researchers to investigate the effect of intervention for each individual by providing information about individual treatment effects and variation of the treatment effects among cases. This type of information is difficult to capture using group comparison designs where the focus is the average treatment effect (Barlow, Nock, & Hersen, 2009). In addition, because only a small sample size is needed, single-case studies allow researchers to study populations of people that have a low prevalence rate (e.g., children with autism) that are difficult to study with large sample based designs (Van den Noortgate & Onghena, 2003a). There are more benefits to using single-case designs. By using single-case designs, researchers can reduce the gap between research and practice because practitioners can implement research in their current setting (Morgan & Morgan, 2001). Finally, this type of design also allows researchers to design an experimental condition within a case by measuring outcome variables prior to the treatment and after the treatment. This feature makes it feasible for the case to provide its own control for the comparison.

Type of Single-Case Design

There are several commonly used single-case designs, such as an AB design, an ABAB design, and a multiple-baseline design (Kazdin, 2011; Shadish & Sullivan, 2011). The AB design is one of the basic designs that has a baseline phase (A) and a treatment phase (B). By comparing outcome scores from the baseline phase and the treatment phase, the treatment effect (changes in the outcome scores between the baseline and the treatment phase) can be evaluated. Figure 1 illustrates a visual display of the basic AB design.

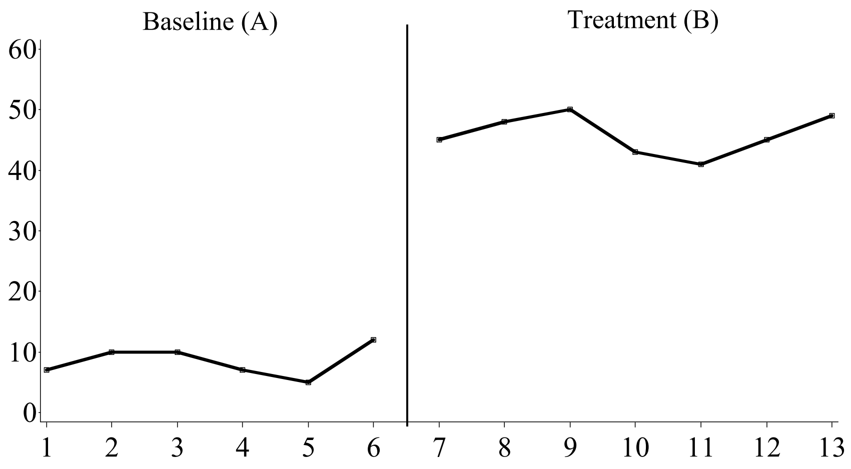


Figure 1. AB design

There is a criticism to using the basic AB design. When using the AB design, it is difficult to conclude that a change of outcome between a baseline and a treatment phase is solely due to a treatment and not due to some external factors which could have occurred at the same time (Ferron & Rendina-Gobioff, 2005). For example, in a case that a researcher finds that the reading score of a child increases after implementing a new reading treatment using an AB design, the researcher may conclude that the new reading treatment is effective in improving reading performance. However, the improvement of the reading performance may be due to natural growth of learning, or due to academic assistance at home from the parent of the child that occurs at the same time that the treatment occurs. Thus, there is a limitation in examining the true effect of the treatment by using the basic AB design. This limitation can be partially overcome by applying more complex designs, such as an ABAB design or a multiple baseline design.

The ABAB design is an extension of the AB design. The ABAB design consists of four phases, two baseline phases and two treatment phases. It has observations of an initial baseline phase (A), followed by observations of an initial treatment phase (B), then observations of a second baseline phase (A), followed by observations of a second treatment phase (B). A treatment is introduced in the initial treatment phase like the AB design, and then the treatment is withdrawn in the second baseline phase and reintroduced in the second treatment phase. A second treatment phase provides the opportunity to demonstrate the performance of the initial treatment phase in that the observed performance pattern of the second treatment phase should replicate the performance change shown in the initial treatment phase. Figure 2 shows a visual display of the ABAB design.

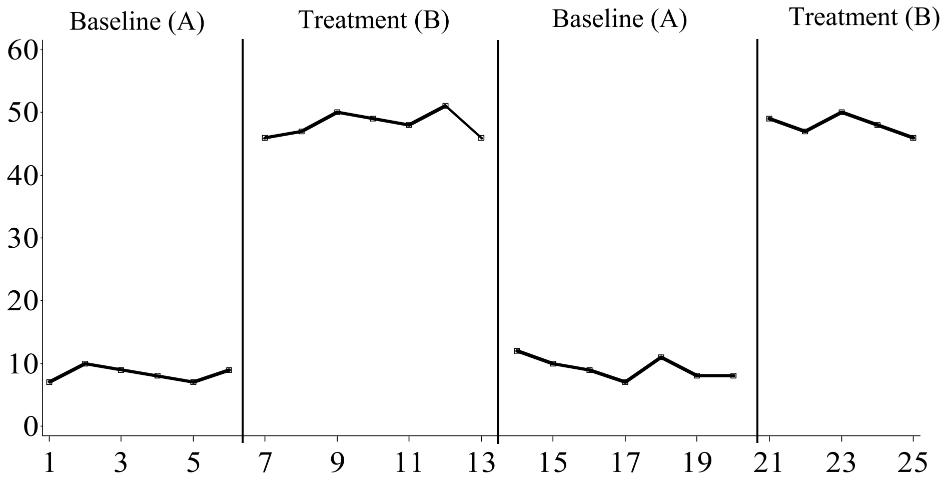


Figure 2. ABAB design

There is an ethical or practical concern for using the ABAB design due to the fact that the treatment should be withdrawn (Kazdin, 2011). Researchers may expect that the behavior will revert toward baseline levels when the treatment is withdrawn which is required to demonstrate the treatment effect. However, in some cases, the treatment effect might be permanent, or maintained after treatment is withdrawn. For example, in an educational setting, once learning occurs after introducing a treatment, it is hard to remove and might be maintained even after withdrawing the treatment.

Another type of extension of the AB design is a multiple-baseline design. Multiple-baseline designs have a baseline phase and a treatment phase that is established for multiple cases. The treatment is introduced to different cases at different points in time so that the initiation of the treatment phase can be staggered across time for the different cases. If changes occur for each baseline when the treatment is introduced, then the treatment effects can be more likely to be attributed to the treatment, not to extraneous events (e.g., history or maturation) (Ferron & Rendina-Gobioff, 2005; Kazdin, 2011). Another benefit of the multiple-baseline design is that the treatment does not need to be removed once the treatment is introduced. This benefit allows researchers to avoid the practical or ethical issues commonly encountered when removing the treatment in the ABAB design. Figure 3 illustrates a visual display of the multiple-baseline design.

Although multiple-baseline designs have some advantages over other designs, there is a limitation due to the potential dependence among cases. In multiple- baseline designs, baselines can be interconnected in that change in a behavior for one case carries over to other cases where the treatment has not been introduced (e.g.,Whalen, Schreibman, &Ingersoll, 2006; Watson , Meeks, Dufrene, & Lindsay, 2002). For example, in the multiple-baseline design across individuals, it is plausible that changes in the behavior of an individual who has received a treatment could impact the behavior of another individual who has not received the treatment.

This can occur more likely in school or home settings where a child or sibling can usually observe the behavior changes of other children or siblings.

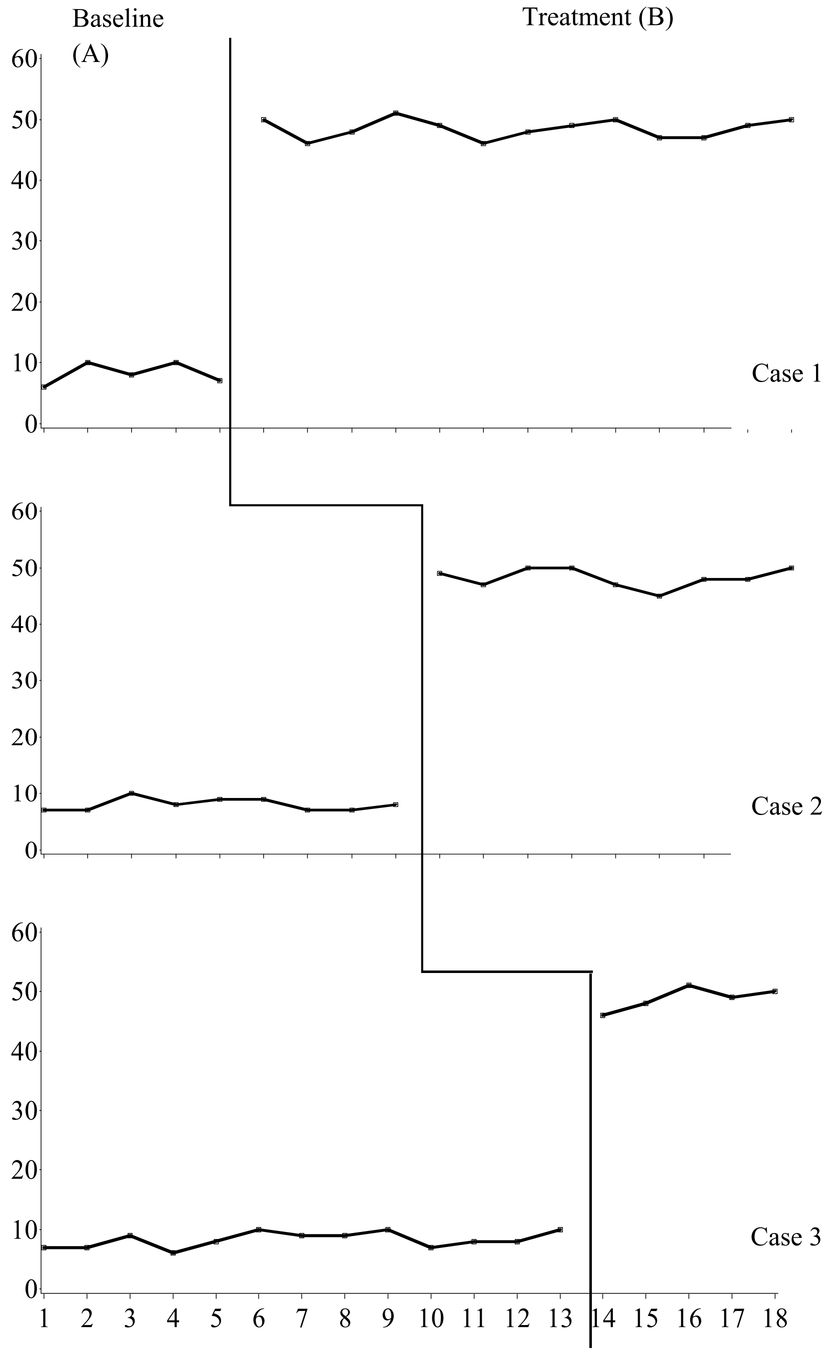


Figure 3. Multiple baseline design

Analysis of Single-Case Design

Several methods to analyze single-case design data have long been developed. These methods can be categorized with four groups: (1) visual analysis, (2) overlap statistics, (3) randomization tests, and (4) regression based analyses.

Visual analysis. Visual analysis has been historically the most commonly used analysis method (Kazdin, 2011; Parsonson & Baer, 1992). Visual analysis is conducted to examine treatment effects by visually inspecting graphed data (Kazdin, 2011). This analysis is intended to focus on a potent treatment effect that can be obviously observed by graphed data. Therefore, it has been argued that researchers who typically use visual analysis tend to be more conservative when evaluating a treatment effect. This can lead the researchers to commit fewer Type I errors but more Type II errors than those who primarily use statistical analyses (Parsonson & Baer, 1986; Kazdin, 2011).

However, several studies have found that using visual analysis is not as conservative as previously thought, and several factors can influence a judgment of treatment effects examined by visual analysis (DeProspero & Cohen, 1979; Fisch, 2001; Jones, Weinrott, & Vaught,1978; Matyas & Greenwood, 1990; Wampold & Furlong, 1981). For example, Matyas and Greenwood (1990) found that visual analysts tend to make high Type I error rates, and relatively low Type II error rates. Fisch (2001) also found that trained behavior analysts often misreport treatment effects when a visual graph of data displayed no treatment effects (Type I error). In order to handle the issue of accuracy raised in visual analysis, several methods such as training, structured criteria and response-guided modification have been suggested (Hogaopian et al., 1997; Ferron, & Jones; 2006; Fisher, Kelley & Lomas; 2003; Parsonson & Baer, 1992). By using these methods, it was demonstrated that the accuracy of visual analysis as well as agreement among visual analysts can be improved (Ferron, & Jones; 2006; Fisher et al., 2003; Hagopian et al., 1997).

However, many researchers have still suggested that it is more valuable to use visual analysis along with other statistical models when evaluating more complex data that have variability in baselines, trends, and complex error structures (Barlow, Nock, & Hersen, 2009; Ferron, & Jones; 2006; Kazdin, 2011).

Nonoverlap statistics. A number of nonoverlap statistics can be utilized in order to describe an overall size of a treatment effect. The underlying rationale for these statistics is to consider nonoverlapping data as an indicator of performance differences between baseline and treatment phases (Sidman, 1960). The extent to which data overlap between baseline and treatment phases can be quantified as the percentage of non-overlapping data (PND; Scruggs, Mastropieri, & Castro, 1987), percentage of all non-overlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007), and percent exceeding the median (PEM; Ma, 2006). Nonoverlap methods have some strengths in that they don't require an assumed parametric model (Armitage, Berry, & Matthews, 2002).

Despite these strengths, several weaknesses are more often addressed. Parker and Vennest (2009) indicate these weaknesses for the previously listed three nonoverlap indices. They claim that (a) PND has a lack of a known underlying distribution that limits building confidence intervals, (b) PEM has issues of a weak relationship with other effect sizes, (c) PEM and PND are hardly able to discriminate among published studies, and (d) all three indices have also an issue of human error from hand calculations of the graphed data. Recently, new indices have been developed to overcome these weaknesses. Nonoverlap of all pairs (NAP; Parker &Vannest, 2009), and Tau-U (Parker, Vannest, Davis, & Sauber, 2011) have been suggested as alternative nonoverlap indices that potentially overcome some of the weaknesses of the traditional nonoverlap indices.

Randomization tests. Randomization tests can also be used to test the effectiveness of a treatment for single-case studies. This method allows single-case studies to be experimental designs by randomly assigning measurement occasions to the baseline or treatment phase (Onghena & Edgington, 2005). The logic behind these tests is that if there are no treatment effects on an outcome, the observations should not be influenced by random assignment of measurement occasions to the baseline or treatment, and therefore, the same scores of the outcome will be found regardless of the treatment assignment (Barlow, Nock, & Hersen, 2009). Based on assuming this null hypothesis is true, a randomization distribution is formed in a randomization test. Randomization tests are not driven by theoretical distributions. They only utilize available sample data to create a randomization distribution. This distribution is formed by rearranging the data to consider all permutations –one rearrangement for each of the possible random assignments. By comparing an obtained test statistic to the randomization distribution, the null hypothesis can be tested (Barlow, Nock, & Hersen, 2009).

There are several benefits to using randomization tests to analyze single-case data. The use of an experimental design with randomization tests can improve both internal validity and statistical conclusion validity of the study by controlling extraneous variables related with natural growth or history. In addition, several studies show that the presence of a treatment effect can be examined while controlling Type I error rates by incorporating a randomized component in single-case design (Edgington, 1980; Ferron & Jones, 2006).

However, there are several drawbacks of this method. A limitation of this method is that it only provides inferences about the presence of a treatment effect. It does not provide inferences about the form of the effect (i.e., change in level and change in trend) or the size of the effect (Morgan & Morgan, 2001; Onghena & Edgington, 2005). Another concern relates to statistical power. It was found that power for randomization tests can be influenced by many factors, such as design types, effect sizes, series lengths, and forms of randomization which in turn, make it difficult to estimate the power of randomization tests (Ferron & Ware, 1995; Ferron & Onghena, 1996; Onghena & Edgington, 2005).

Regression based analyses. Regression analyses have been proposed as methods that are able to capture both changes in level and changes in trend in single-case data. Regression methods can be categorized based on the number of levels allowed in the analysis: (1) single-level analysis for one case, and (2) multilevel analysis for multiple cases.

Single-level analysis. Single-level analyses are simple regression types of analyses including ordinary least squares regression and generalized least squares regression. Ordinary least squares (OLS) regression was first suggested (Center, Skiba, & Casey, 1985-1986; Huitema & McKean, 1998) as a single-level regression method to analyze a single-case. This OLS regression can be illustrated by the following regression model:

$$y_i = \beta_0 + \beta_1 \ phase \ + e_i \qquad (1)$$

where y_i is the observed value at the i^{th} point in time, β_0 is an average of the baseline phase, *phase* is a dummy variable with 0 for the baseline phase and 1 for the treatment phase, β_1 is the mean difference between the baseline and the treatment phase which indicates the treatment effect, and e_i is the error at the i^{th} point in time. This simple regression model can be expanded to include more variables to capture trends in phases (e.g., Center, Skiba, & Casey, 1985; Huitema & McKean, 2000). The use of OLS regression methods has raised concern that errors in the statistical model are considered to be independent as opposed to dependent (autocorrelated) (e.g., Kratochwill et al., 1974; McKnight, McKean, & Huitema, 2000).

Some alternative approaches have been suggested to resolve the dependency of the errors, autocorrelation, in single-case data. Generalized least squares (GLS) regression is one of the alternative single-level analyses that can handle the autocorrelated errors (Cochrane & Orcutt, 1949; McKnight, McKean, & Huitema, 2000; Simonton, 1977; Solanas, Manolov, & Sierra, 2010). The GLS regression shares a similar statistical framework with the OLS regression, but unlike OLS regression, the autocorrelation among the errors can be estimated and taken into account for the analyses (Maggin et al., 2011). More explicit explanation about autocorrelation has been provided in a later section (see the Level-1 Error Structure section).

Multi-level analysis. In recent years, multilevel modeling (MLM) has been suggested as an alternative method to the single-level model to analyze single-case data (e.g., Nugent,1996; Shadish & Rindskopf, 2007; Shadish, Rindskopf, & Hedges, 2008; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008; Baek et al, 2013).

Multilevel modeling provides great flexibility which is considered as a potential advantage of using multilevel modeling over single-level analyses. Multilevel modeling can provide more detailed information regarding the treatment effects than single-level models because in addition to individual treatment effect estimates, they also provide an estimate of the average treatment effect, and the variability of treatment effects across cases. In addition, since multilevel analyses can provide empirical Bayes estimates, person specific estimates of short series from multilevel analyses can be more reliable than the estimates from single-level analyses (Raudenbush & Bryk, 2002). Moreover, multilevel models can handle a variety of modeling issues that may arise in single-case studies (e.g., the modeling of possible dependency, linear or nonlinear trends, and count outcomes) (Van den Noortgate & Onghena, 2003a). Thus, this flexible modeling approach can provide more in-depth information regarding inferences of the study.

A basic two-level multilevel model for single-case studies (e.g. an AB design), assuming no time trends during the baseline and treatment phase, is shown in equations 2 and 3. Equation 2 is for the first level of the multilevel model, which is comparable to the OLS regression model.

$$y_{ij} = \beta_{0j} + \beta_{1j}\ Phase_{ij} + e_{ij} \qquad (2)$$

$$\beta_{0j} = \theta_{00} + u_{0j} \qquad (3)$$

$$\beta_{1j} = \theta_{10} + u_{1j}$$

y_{ij} is the observed value (outcome) at the i^{th} observation for the j^{th} case. β_{0j} is the baseline intercept for the j^{th} case, and $Phase_{ij}$ is a dichotomous variable that indicates the phase in which the observation occurred, being 0 indicates the baseline phase and 1 indicates the treatment phase. β_{1j} is the difference between the baseline level and the treatment level (shift in level) for the j^{th} case which indicates a treatment effect. e_{ij} is residual that indicates within case variation (level-1 error). Equation 3 is for the second level of the multilevel model which can allow variation in the baseline intercept and the shift in level across cases. θ_{00} is the average baseline intercept, θ_{10} is the average shift in level, and u_{0j} and u_{1j} are errors for the average baseline intercept and the average shift in level across cases. u_{0j} and u_{1j} are assumed to be multivariate normally distributed N(0,Σ_u).

This basic model can be extended to include slopes in the baseline and the treatment phase. Equation 4 is the first level of the extended model that includes the $Time_{ij}$ variable as an indicator of the slope. β_{0j} is the baseline intercept for the j^{th} case and β_{1j} is the difference between the baseline level and the treatment level (shift in level) for the j^{th} case when $Time_{ij}$

equal to 0. β_{2j} as the baseline slope for the j^{th} case, and β_{3j} as the change in slopes between the baseline and the treatment phase (shift in slope).

$$y_{ij} = \beta_{0j} + \beta_{1j}\ Phase_{ij} + \beta_{2j}\ Time_{ij} + \beta_{3j}\ Time_{ij} {*} Phase_{ij} + e_{ij} \qquad (4)$$

$$\begin{aligned} \beta_{0j} &= \theta_{00} + u_{0j} \qquad (5) \\ \beta_{1j} &= \theta_{10} + u_{1j} \\ \beta_{2j} &= \theta_{20} + u_{2j} \\ \beta_{3j} &= \theta_{30} + u_{3j} \end{aligned}$$

Equation 5 is the second level of the extended model that allows variation across cases in the baseline intercept, the baseline slope, the shift in level, and the shift in slope. θ_{00} is the average baseline intercept and θ_{10} is the average shift in level at $Time_{ij}$ equal to 0*,* θ_{20} is the average baseline slope, and θ_{30} is the average shift in slope. u_{0j} , u_{1j} , u_{2j} and u_{3j} are errors in the second level equation.

Although several advantages exist, some concerns involving the use of multilevel models also exist regarding assumptions. In order to make valid inferences of multilevel models, several assumptions need to be met. For example, the variance in the baseline phase and in the treatment phase is assumed to be equal, and the level-1 variance is also assumed to be equal for all the cases. However, it is difficult to test the violation of these assumptions prior to conducting the analyses, particularly with the single-case data that have typically small sample sizes.

Level-1 Error Structures

As mentioned previously, the errors in the first-level model (eij) in equations 1, 2, and 4 are within case errors that indicate the discrepancy between the observed values and predicted values of outcome from an individual's growth trajectory. Several assumptions regarding the within case errors (or level-1 errors) have to be taken into account when we use regression based methods to analyze the data. Errors are assumed to have covariance Σ_e, and they are both identically and normally distributed. Various error structures can be assumed for the covariance Σ_e. It can be assumed as either having an independent error structure or having an autocorrelated structure. In the following sections, autocorrelation in single-case design is introduced and then it is explained how the covariance Σ_e can be modeled in single-level and multilevel models using the autocorrelation. Some issues which arise when misspecifying the level-1 error structures are also discussed.

Autocorrelation in Single-Case Design

Many researchers have argued that the observations from single-case design may yield positive autocorrelations (Busk & Marascuilo, 1988; Huitema, 1985; Huitema & McKean, 1998; Matyas & Greenwood, 1996). Since an outcome is measured repeatedly across time in a single-case study, it is possible level-1 errors produced by these repeated measurements may be more similar when they are close in time which leads to dependency in the errors, or autocorrelation. A number of non-modeled factors (e.g., illness, moving to a new school) could affect the level-1 errors that indicate discrepancy between actual observed outcome values and predicted outcome values from an individual's growth trajectory. If the non-modeled factors affect the sequential errors that are close in time, then the errors may be more similar at close points in time. For example, a growth trajectory of reading achievement for a child may show a constant increasing trend. Actual observations of the child, however, may deviate from this trajectory due to a non-modeled factor such as illness of the child. She might feel tired and sick; that could affect an observation of reading achievement. The sickness of the child is more likely to affect the next couple or more sequential observations. In this case, the errors that were closer in time would be more similar, which leads to positive autocorrelation.

Level-1 Error Structures in Single-Case Design

There are a number of possible level-1 error structures Σ_e that can be modeled in single-case design. Level-1 error structures can be modeled as being autocorrelated or as independent in single-case data analysis. The independent error structure is a fairly simple structure compared to autocorrelated error structures. Variance components (VC) or Identity structure (ID) is the simplest error structure and assumes the errors are independent of each other. There are various error structures that assume the errors to be autocorrelated. These error structures include unstructured, compound symmetry, banded toeplitz or moving average, first-order autoregressive [AR(1)], AR(1) plus a diagonal, AR(1) plus a common covariance, and an AR(1) generalization for unequally-spaced observations (Goldstein, 1995; Goldstein, Healy, & Rasbash, 1994; Heitjan & Sharma, 1997; Jennrich & Schluchter, 1986; Ware, 1985; Wolfinger, 1993; Yang & Goldstein, 1996). The recognition that autocorrelation may exist among the level-1 errors leads autocorrelated error structures to be utilized more often in single-case data analysis. Figure 4 illustrates examples of the level-1 error structures generally used for single-case data analysis. Identity structure (ID) contains a single parameter (σ^2) on the main diagonal of a diagonal matrix that assumes no correlation between any pair of random errors (Raudenbush & Bryk, 2002). This oversimplified structure is very unlikely to be true in repeated measures data (Goldstein, Healy, &, Rasbash, 1994). First-order autoregression [AR(1)] structures are composed of two parameters, σ^2 and ρ, and ρ represents the autocorrelation coefficient. The correlations between two errors that are separated by one, two, three, and *n* points in time are represented by ρ, ρ^2, ρ^3, and ρ^n, respectively. First-order autoregression and first-order moving average [ARMA(1,1)] has the same two parameters (σ^2 , ρ) as AR(1) has and the moving average coefficient (r). This structure contains σ^2 on the main diagonal to represent error variance, and the correlations between two errors that are separated by one, two, three, and n points in time represent by r, $r\,\rho$, $r\,\rho^2$, and $r\,\rho^n$, respectively. Second-banded Toeplitz [TOEP(2)] contains two parameters, σ^2 and σ_1, and σ_1 represents constant covariance between two errors that are separated by one point in time. This error structure assumes the errors that are separated by more than one point in time are not correlated, which means zero correlation.

$$\begin{array}{cccc}
\text{ID} & \text{AR(1)} & \text{ARMA(1,1)} & \text{TOEP(2)} \\
\begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} &
\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} &
\sigma^2 \begin{bmatrix} 1 & r & r\rho & r\rho^2 \\ r & 1 & r & r\rho \\ r\rho & r & 1 & r \\ r\rho^2 & r\rho & r & 1 \end{bmatrix} &
\begin{bmatrix} \sigma^2 & \sigma_1 & 0 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 & 0 \\ 0 & \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & 0 & \sigma_1 & \sigma^2 \end{bmatrix}
\end{array}$$

Figure 4. Examples of the level-1 error structures used in single-case data analysis

Misspecification Issues of Level-1 Error Structures in Single-Case Design

When the existing autocorrelation is not modeled in the analysis, it can lead level-1 error structures to be misspecified. Research has shown significant impacts of misspecifying level-1 error structure on statistical inferences. These misspecification issues of level-1 error structure arise for both single-level and multilevel analyses.

Single-level model. In single-level model analyses, much research shows that positive autocorrelation inflates Type I error rates in significance tests of the treatment effect when the autocorrelation is not taken into account (Matyas & Greenwood, 1990; Toothaker, Banz, Noble, Camp, & Davis, 1983). More specifically, under a general linear model like OLS regression, the positive autocorrelation can lead the regression coefficients to be unbiased, but the standard errors of the regression coefficients to be underestimated which implies inflated t-values. As a result, 95% confidence intervals tend to be too small and significance tests of the treatment effect tend to be liberal (Neter, Wasserman, & Kutner, 1990). Typically, as the level of autocorrelation increases, the degree to which confidence intervals and significance tests are impacted increases. The impact of positive autocorrelation has been also demonstrated with various series lengths and patterns of autocorrelation (Beretvas & Chung, 2008; Greenwood & Matyas, 1990; Huitema, McKean, & McKnight, 1999; Scheffé, 1959; Toothaker, Banz, Noble, Camp, & Davis,1983).

Some efforts have recently been made to resolve this issue by using the GLS regression method. GLS regression requires two steps to account for autocorrelation in the analyses. The autocorrelation can be first estimated from the errors of the initial fit of the linear model, and then can be included in the analyses to refit the linear model (Mcknight, Meckean & Huitema, 2000; Maggin et al., 2011). There are several methods that are available to estimate the autocorrelations under the GLS regression approaches, such as Simonton (Simonton,1977), Cochrane-Orcutt (Cochrane-Orcutt, 1949), and Paris-Winsten (Paris-Winsten,1954) versions of GLS (McKnight, McKean, & Huitema, 2000; Solanas, Manolov, & Sierra, 2010). However, studies have demonstrated that GLS methods still produce high Type I error rates when applied to small samples (e.g., Johnston, 1984; Huitema & Mckean, 1991; Solanas, Manolov, & Sierra, 2010). McKnight, McKean, and Huitema (2000) found that a double-bootstrapping procedure under the GLS regression can improve the accuracy of the parameter estimates as well as autocorrelation estimates and control Type I error rates. Their Monte Carlo simulation study shows that the bootstrap bias-adjusted method estimates of the autocorrelation are substantially less biased than initial estimates of the autocorrelation obtained by other traditional GLS methods (i.e., Cochrane-Orcutt, and Paris-Winsten). Type I error rates for all parameter estimates using the bootstrap bias-adjusted method are close to the nominal level, less than .05. In addition, Maggin et al. (2011) proposed applying the Bayesian estimation approach under the GLS regression method to compute effect sizes for single-case data. This method is particularly applicable to small-sample time-series data with autoregressive errors. They recommend the use of the GLS method as a support for visual analysis. However, sufficient empirical evidence has not yet been gathered for this method.

Multi-level model. Multilevel modeling (MLM) is another method that allows the possible dependency of the observations to be taken into account, and has been used as an alternative method for analyzing single-case data (Nugent, 1996; Shadish & Rindskopf, 2007; Shadish, Rindskopf, & Hedges, 2008; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). The flexibility of the multilevel approach makes it possible not only to allow for dependent error structures, but also to allow the covariance parameter values to differ across cases. By using this approach, researchers can model the variety of error structures described in the previous section.

Misspecifying the level-1 error structure in MLM analyses has also been found to bias estimates of the parameters (Ferron et al., 2009; Ferron, Dailey, & Yi, 2002; Guerin & Stroup, 2000; Kwok et al., 2007; Sivo, Fan, & Witta, 2005; Sivo & Willson, 2000). For example, Ferron et al. (2009) conducted a Monte Carlo simulation study to examine the utility of multilevel models for multiple baseline design of single-case data. They found that the fixed effect estimate of the average treatment effect was relatively unbiased, regardless of whether the level-1 error structure was correctly specified or not. However, they indicated that the confidence interval coverage of the treatment effect was less accurate and estimates of the variance components tended to be more biased when level-1 error structure was misspecified. Ferron, Dailey and Yi (2002) also studied the effect of the misspecified level-1 error structure using a parsimonious covariance structure (ID) rather than the true structure [AR(1)] in MLM analyses. In their simulation study, they found that the estimates of the fixed effects were unbiased but the estimates of variance parameters were substantially biased when the level-1 error structure was misspecified for all conditions (i.e.,variety of series lengths, sample sizes, and levels of autocorrelation). Specifically, both variance in the intercept and the slope (level-2 variance) were overestimated; the level-1 error variance was underestimated. Kwok and his colleagues (2007) studied the impact of broader types of misspecifying the level-1 error structure in repeated measured data analysis under the multilevel model framework. Their simulation results implied the impact of misspecification of the Σ_e matrices were more likely to result in overestimation in random effects, when parsimonious covariance structures were used, and underestimation in random effect variances when other types of misspecification occurred. Furthermore, using parsimonious covariance structure resulted in overestimation of the standard errors in the fixed effect, which resulted in lower statistical power relative to the correct specification. Recently, Petit-Bois (in press) investigated the effects of various types of misspecifications of the level-1 error structure when using a three-level meta-analytic single-case model. She found consistent results from the previous studies. Her simulation results indicate that misspecification of the level-1 error structure has little or no impact on the treatment effects, but, it has significant impact on the variance components. Specifically, the estimates of error variances and autocorrelation were more biased; confidence interval coverage for the level-2 and level-1 error variance, and autocorrelation tended to be small, and confidence interval width tended to be large for some cases.

Overall, previous research for both single-level and multilevel models implies that misspecification of level-1 error structure has little to no impact on the point estimates of the fixed effects, but it has a significant impact on the corresponding standard errors of the fixed effects. These impacts can lead to lower statistical power of the inferences. Moreover, misspecification of level-1 error structure leads to significant bias on random effect estimation. Depending on the types of the misspecification, it was more likely to be either overestimated or underestimated. Thus, single-case researchers should inspect for the presence of the autocorrelation in their data, and consider modeling autocorrelation if it presents in their data. By doing this, they can avoid possible misspecification on the level-1 error structure (Barlow, Nock, & Hersen, 2009; Kazdin, 2011). If there is uncertainty about the level-1 error structure, it is generally recommended to avoid an overly parsimonious error structure (i.e., ID) (Ferron, Dailey, & Yi, 2002), and to consider using a slightly over-specified model (e.g., TOEP(2) or AR(1)) (Kwok et al., 2007).

Assumption of Between Case Homogeneity in Level-1 Error Structures

Although multilevel modeling allows autocorrelation among level-1 errors to be taken into consideration in single-case data analyses, this approach still holds a critical assumption that the level-1 error structure is the same for all cases. Specifically, it is assumed that (a) the degree of autocorrelation is the same for all cases and (b) the level-1 error variance is the same for all cases. Previous single-case research using multilevel modeling application as well as misspecification research of level-1 error structures has often assumed the autocorrelation and level-1 error variance to be equal for all cases (Ferron et al.,2009; Ferron, Farmer, & Owens, 2010; Van Noortgate & Onghena, 2003; Kwok et al., 2007).

However, it is possible that this assumption may not be true all the time. The autocorrelation and level-1 error variance may vary across cases. Level-1 errors could be attributed by measurement errors, and the differences of measurement errors across cases can lead the level-1 error variances to vary. For example, differences in mood, motivation, and fatigue among cases are some of the sources causing measurement error. The measurement error caused by these personal related factors is likely to be different across cases, and this could lead the level-1 error variances to vary. The findings from previous studies of level-1 error structures in single-case data support that variations in level-1 error structures could exist (Baek & Ferron, 2011; Baek & Ferron, 2013). Baek and Ferron (2013) discovered relatively large differences found in terms of estimates of autocorrelation and level-1 error variances, after estimating level-1 errors separately for each case. In the study, five single-case data sets from published papers were selected and reanalyzed separately using a two-level multilevel model with varying error structures across cases. The results of the analyses found substantial differences in terms of the autocorrelation [AR(1)] estimates among the cases in all five studies. For example, in one study, the autocorrelation ranged from -.04 to .46 when estimated separately for each case, while it was estimated to be .22 when estimated to be constant across cases. The study also found that level-1 error variance estimates were substantially different across cases in all five studies. For example, in one study the error variance ranged from 164.41 to 795.62 when estimated separately for each case, while it was estimated to be 269.54 when estimated to be constant across cases.

If the variation which exists in a level-1 error structure is not taken into consideration, it can conceivably impact the inferences of the study for both fixed effects and random effects.

Thus, it is critical to examine the consequences of different modeling approaches (modeling and not modeling between case variation) in the level-1 error structure. Despite the importance of this issue, the effects of the different approaches to modeling the level-1 error structure has not been systematically studied.

Modeling Between Case Variation in Level-1 Error Structures

The two level model that allows between case variation in the level-1 error structure in single-case design can still be represented by the Equations (4) and (5). In Equation (4), eij represents level-1 errors, and the covariance structure $\sum_e$ of the errors can be assumed as any of the error structures being autocorrelated or being independent that have been introduced previously.

When we model between case variation in the level-1 error structure, the covariance structure $\sum_e$ is assumed to be one of the autocorrelated covariance structures, and is allowed to vary across cases. More specifically, autocorrelation and level-1 error variance are estimated separately for each case; therefore, every case is allowed to have a unique autocorrelation and level-1 error variance value. The following example illustrates three different ways of modeling level-1 covariance structure $\sum_{e\,.}$ Assume that there are single-case data with three cases. The simplest way to model the covariance structure $\sum_e$ is to assume it to have an independent structure (ID). Assume the level-1 error variance is estimated as 35, and held constant across cases. This is illustrated in Figure 5.

$$\begin{array}{ccc} \text{case1} & \text{case2} & \text{case3} \\ 35\begin{bmatrix} 1 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \\ 0 & 0 & 0 & 1 & \\ & & \vdots & & \end{bmatrix} & 35\begin{bmatrix} 1 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \\ 0 & 0 & 0 & 1 & \\ & & \vdots & & \end{bmatrix} & 35\begin{bmatrix} 1 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \\ 0 & 0 & 0 & 1 & \\ & & \vdots & & \end{bmatrix} \end{array}$$

Figure 5. $\sum_e$ is assumed to be ID and held constant across three cases

Another way of modeling $\sum_e$ is assuming it to have one of the autocorrelated error structures. Assume that the first-order autoregressive structure [AR(1)] is assumed for the covariance structure $\sum_e$. When the covariance structure $\sum_e$ is held constant across cases with the autocorrelation and the variance of level-1 error being estimated as .2 and 30, respectively, these values apply for all three cases. This is illustrated in Figure 6.

$$\begin{array}{ccc} \text{case1} & \text{case2} & \text{case3} \\ 30\begin{bmatrix} 1 & .2 & .2^2 & .2^3 & \\ .2 & 1 & .2 & .2^2 & \cdots \\ .2^2 & .2 & 1 & .2 & \\ .2^3 & .2^2 & .2 & 1 & \\ & & \vdots & & \end{bmatrix} & 30\begin{bmatrix} 1 & .2 & .2^2 & .2^3 & \\ .2 & 1 & .2 & .2^2 & \cdots \\ .2^2 & .2 & 1 & .2 & \\ .2^3 & .2^2 & .2 & 1 & \\ & & \vdots & & \end{bmatrix} & 30\begin{bmatrix} 1 & .2 & .2^2 & .2^3 & \\ .2 & 1 & .2 & .2^2 & \cdots \\ .2^2 & .2 & 1 & .2 & \\ .2^3 & .2^2 & .2 & 1 & \\ & & \vdots & & \end{bmatrix} \end{array}$$

Figure 6. $\sum_e$ is assumed to be AR(1) and held constant across three cases

Those two ways of modeling $\sum_e$ are traditional ways that are often modeled for the single-case analysis. For the proposed approach where the covariance structure $\sum_e$ is allowed to vary across cases, the autocorrelation and the variance of level-1 error will be estimated with as many values as cases. An example of this approach is illustrated in Figure 7. As you see in Figure7, each case has unique autocorrelation and variance when between case variation is modeled for the level-1 error covariance structure.

$$\begin{array}{ccc} \text{case1} & \text{case2} & \text{case3} \\ 30\begin{bmatrix} 1 & .2 & .2^2 & .2^3 \\ .2 & 1 & .2 & .2^2 \cdots \\ .2^2 & .2 & 1 & .2 \\ .2^3 & .2^2 & .2 & 1 \\ & & \vdots & \end{bmatrix} & 55\begin{bmatrix} 1 & .6 & .6^2 & .6^3 \\ .6 & 1 & .6 & .6^2 \cdots \\ .6^2 & .6 & 1 & .6 \\ .6^3 & .6^2 & .6 & 1 \\ & & \vdots & \end{bmatrix} & 12\begin{bmatrix} 1 & .4 & .4^2 & .4^3 \\ .4 & 1 & .4 & .4^2 \cdots \\ .4^2 & .4 & 1 & .4 \\ .4^3 & .4^2 & .4 & 1 \\ & & \vdots & \end{bmatrix} \end{array}$$

Figure 7. $\sum_e$ is assumed to be AR(1) and allowed to vary across three cases

Estimation Methods

Restricted maximum likelihood (REML) estimation. The traditional estimation method to run the three specified models is restricted maximum likelihood (REML) estimation (Patterson & Thompson, 1971; Kenward & Roger, 1997, 2009). This estimation method has been historically and commonly utilized to analyze multilevel models. It has become a standard of variance component estimation in MLM and has provided computation advantages in that it is relatively fast and automated by many software programs (e.g., HLM, MLwiN, SAS, SPSS, R, and Stata). The rationale behind likelihood estimation is that the best way to estimate a parameter is to find the value for which the observed data were most likely to have occurred (Lynch, 2007).

The REML estimation has been commonly used to estimate the traditional models in many single-case applications (e.g., Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens, 2010; Van Noortgate & Onghena, 2003a). Generally, it has been found that the REML method used to estimate multilevel models in single-case data produces correct inferences for fixed effects by adjusting standard errors and degrees of freedom (Kenward & Roger, 1997, 2009), but produces biased variance components. Several methodological research studies of single-case also support these findings (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013a, 2013b; Owens & Ferron, 2012). Specifically, Monte Carlo simulation studies suggest that using REML to estimate a variety of multilevel models and data conditions for single-case data leads to: (1) unbiased fixed effects (i.e., treatment effect) regardless of sample sizes, (2) accurate confidence intervals for the fixed effects (average treatment effect) regardless of sample sizes, as long as Kenward-Roger or Satterthwaite methods are used for the degree of freedom estimates, and (3) biased variance estimates particularly with small sample sizes.

However, for the proposed model where the covariance structure $\sum_e$ is allowed to vary across cases, using the REML estimation has raised a technical issue (Baek, Petit-Bois, & Ferron, 2012). The estimation can be computationally intensive since the level-1 error structure should be estimated for each case. It turns out that as the number of cases increases, the number of parameters increases, and that leads to non-convergence issues. For example, in a recent study of single-case studies (Baek, Petit-Bois, & Ferron, 2012), the multilevel meta-analytic model of single-case data was extended to allow the autocorrelation [AR(1)] and error variance to vary across studies and cases using REML estimation. In this analysis, convergence criteria were not met when the level-1 error structure was allowed to vary across studies or cases. Thus, in order to apply the proposed idea of allowing between variation in level-1 error structure, it is necessary to use an alternative estimation approach that can solve the convergence issue.

Bayesian estimation. Bayesian estimation can be one of the alternative estimation methods to handle the convergence issue. Bayesian estimation method has been considered as a practical method for analyzing data for many areas such as education, social science, psychology, and medical decision making (Lindley & Smith, 1972; Gelman, Carlin, Stern, & Rubin, 2004). Bayesian inference is the process of fitting a probability model, given the observed data, and summarizing uncertainty of parameters by a probability distribution (Gelman, Carlin, Stern, & Rubin, 2004). This method incorporates existing information into the analysis by constructing prior distributions using the existing information (e.g., Howard, Maxwell, & Fleming, 2000). Bayesian estimation can take into account the uncertainty of estimating both fixed effects and variance components by using these constructed prior distributions (Gelman, 2002; Gelman, Carlin, Stern, & Rubin, 2004).

Bayesian estimation methods are well known for their benefits of analyzing social science data (e.g., Gelman & Hill, 2007; Howard, Maxwell, & Fleming, 2000; Kruschke, 2011a, 2011b; Lynch, 2007; Yuan & MacKinnon, 2009). They have great flexibility to construct hypothesis tests and interval estimates, and they also have a benefit to estimate parameters in special cases (e.g., non-normal sampling distributions). Bayesian estimation can also handle inferential and technical challenges of using likelihood estimation in multilevel analysis (Gelman, Carlin, Stern, & Rubin, 2004; Shadish, Rindskopf, & Hedges, 2008; Shadish & Rindskopf, 2007). Studies in multilevel analyses have found that Bayesian methods perform as well or better than likelihood methods regarding bias, efficiency, and coverage (Browne, 2008; Baldwin & Fellingham, 2013). For the multilevel single-case research, the Bayesian approach could provide more accurate results when using small samples or unequal sample sizes per subject (Shadish, Kyse, & Rindskopf, 2013). Convergence issues could also be resolved by using Bayesian estimation methods (Baek, Petit-bois, & Ferron, 2013). Bayesian methods are capable of performing with computationally intensive cases by using Markov Chain Monte Carlo (MCMC) procedures (e.g., Chen & Shao, 1999; Cowles &Carlin, 1996; Gelman, Carlin, Stern, & Rubin, 2004; Gilks, Richardson, & Spiegelhalter, 1996; Tierney, 1994). Baek, Petit-bois, and Ferron (2013) found that more complex multilevel models of single-case data, which failed previously using REML, can reach convergence using the Bayesian estimation method. Bayesian estimation can also be implemented by a variety of software programs, such as MLwinN, R, SAS, and WinBugs.

Bayesian form of the equation for multilevel models. Since Bayesian estimation method is implemented using a probability framework, the multilevel model can also be expressed using probability distributions. Thus the simple traditional two-level single-case model which is represented by equation (4) and (5) can be re-written as seen in the following equation:

$$y_{ij} \sim \text{Normal}(\mu_{ij}, \sigma^2) \qquad (6)$$

$$\mu_{ij} = \alpha_j + \beta_j \mathit{Time}_{ij} + \gamma_j \mathit{Phase}_{ij} + \delta_j \mathit{Time}_{ij} * \mathit{Phase}_{ij}$$

$$\alpha_j \sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha)$$

$$\beta_j \sim \text{Normal}(\mu_\beta, \sigma^2_\beta)$$

$$\gamma_j \sim \text{Normal}(\mu_\gamma, \sigma^2_\gamma)$$

$$\delta_j \sim \text{Normal}(\mu_\delta, \sigma^2_\delta)$$

where, y_{ij} is the observed value (outcome) for the i^{th} observation at the j^{th} case; α_j is the intercept of the baseline for the j^{th} case; β_j is the baseline slope for the j^{th} case; γ_j is the shift in level for the j^{th} case; δ_j is the shift in slope for the j^{th} case. σ^2 is the variance of the within case errors and it is assumed constant across cases in this equation. For the second level equation, μ_α is the average intercept of the baseline; μ_β is the average baseline slope; μ_γ is the average shift in level; μ_δ is the average shift in slope, and σ^2_α, σ^2_β, σ^2_γ, and σ^2_δ are corresponding error variances. These μ_α , μ_β , μ_γ , μ_δ ,σ^2_α , σ^2_β, σ^2_γ, and σ^2_δ are refered to as *hyperparameters* in that they are the upper level of parameters, not the direct parameters (i.e., α_j, β_j, γ_j, δ_j) of the model.

In addition, it is assumed that all regression coefficients, α_j, β_j, γ_j, δ_j, follow a normal distribution. In the Bayesian method, this distribution is called a prior distribution, and all parameters and hyperparameters are required to have a prior distribution.

Prior probability distribution. The prior distribution is a crucial part of Bayesian inference. It represents the plausible distribution for an unknown parameter that is believed prior to observing the specific data (Gelman, 2002; Gelman, Carlin, Stern, & Rubin, 2004). The belief could be obtained from previous research or theoretical rationale. Without using a prior distribution, any Bayesian inference cannot be made.

Reasonable choices of objective prior distributions, *noninformative* prior distributions, will have minor effects on posterior inferences (Berger, 2006; Efron & Morris, 1975; Goldstein, 2006; Gelman, 2002; Jeffreys, 1961; Morris, 1983). The rationale for using noninformative prior distributions is to make the data speak for themselves so that posterior inferences are unaffected by external information out of the current data (Gelman, 2006; Gelman, Carlin, Stern, & Rubin, 2004).

Reasonable noninformative prior distributions have been developed for the parameters of the multilevel models. Typically, enough data is available to estimate fixed effect (i.e, $\mu_{\alpha\,,}\,\mu_\beta$, μ_γ , and μ_δ in Equation 6) and level-1 error variance (σ^2) in multilevel models that one can use any reasonable noninformative prior distribution (Gelman, Carlin, Stern, & Rubin, 2004; Gelman, 2006). A common prior distribution used in applied work for the fixed effects is a noninformative normal distribution, and a noninformative uniform distribution is a commonly used prior distribution for σ.

In general, noninformative normal distributions are constructed with large variance (i.e., 1000^2), so that posterior inferences cannot be influenced by the choice of variance value. Similarly, for the uniform distribution, when the upper limit of σ (standard deviation unit) goes sufficiently large, it yields a proper posterior distribution, and inferences are not sensitive to the choice of the upper limit value. The term *sufficiently large* is subjective in that it will be defined by the scale of the target parameter (i.e., σ). One could obtain a rationale for the proper scale of the target parameter by conducting a marginal analysis (e.g., general regression based analysis). The lower limit of σ is commonly set to be 0 due to the fact that the value of standard deviation could not be negative.

Unlike fixed effects and level-1 error variance, noninformative prior distributions for level-2 variance parameters (i.e., variance of the hyperparameters; σ^2_α, σ^2_β, σ^2_γ, and σ^2_δ in Equation 6) have been more difficult to construct. The choice of noninformative prior distribution for level-2 variance parameters can have a substantially large impact on inferences, especially in the case where the number of *j* (cases; unit of the higher level) is small or the corresponding level-2 variance is close to zero (Gelman, 2002; Gelman, 2006).

Many researchers have suggested various noninformative prior distributions for the hierarchical variance parameters in multilevel models, including uniform, inverse-gamma family, and half-t distributions (Berger & Strawderman, 1996; Daniels & Kass, 1999; Gelman, 2006; Spiegelhalter, Thomas, Best, & Lunn, 2003). For example, Gelman (2006) demonstrated the impact of various proposed noninformative prior distributions for the level-2 variance parameters in multilevel models by using a simple example. He found that the uniform distribution generally works well in that it has little impact on posterior inferences, as long as the number of $j \geq 3$ which is required to ensure a proper posterior density. Thus, he recommended starting with a noninformative uniform prior density for the standard deviation of the level-2 variance.

Convergence criteria. In the Bayesian estimation approach, convergence refers to diagnosing if MCMC techniques reach a proper posterior distribution. MCMC techniques will eventually converge to the posterior distribution, but if iterations have not proceeded long enough, the simulations may not be representative of the population distribution. Therefore, in

Bayesian estimation, one must determine when convergence occurs, and then, how many samples are needed to make accurate posterior inferences after reaching convergence (Gelman, Carlin, Stern, & Rubin, 2004; Cowles &Carlin, 1996; Spiegelhalter, Thomas, Best, & Lunn, 2003).

A number of techniques have been implemented in various software packages to identify these two issues. Various techniques of monitoring convergence are available in WinBUGS software, including trace plots, history plots, Kernel density plots, and Brooks–Gelman–Rubin (BGR) plots. A trace or history plot is one of the intuitive diagnostic criteria which plots the parameter value at time against the iteration number. When more than one chain is assigned simultaneously, the trace and history plots show each chain in a different color. If all the chains overlap one another, we can be confident to say that convergence has been achieved (see Spiegelhalter, Thomas, Best, & Lunn, 2003). A clear sign of non-convergence occurs when we observe some trends in the plots. Kernel density plot shows the final posterior distribution of the estimated parameter. This plot could be another useful diagnostic criterion. When converge occurs, the distribution shows a smooth shape. Generally, as more iterations are performed, the distribution will become smoother. WinBUGS also has the Brooks-Gelman-Rubin (BGR) diagnostic which is computed based on the ratio of between-within chain variances (Brooks & Gelman, 1997; Brooks & Roberts, 1998; Cowles & Carlin, 1996; Gelman & Rubin, 1992). The intuition is that the variance within the chains should be the same as the variance across the chains. BGR plots have three lines: green lines represent the normalized width of the central 80% interval of the pooled, blue lines represent the normalized average width of the 80% intervals within the individual, and red lines represent the BGR statistic, R. When R converges to 1, and both the pooled and within interval widths converge with stability, we consider convergence has occurred. Convergence for analyses of this study will be visually inspected by these different diagnostic criteria.

Even if the simulations have reached convergence, the early iterations could still be influenced by the starting point rather than the population distribution. To eliminate the effect of the starting point on posterior distribution, it is generally recommended to discard the first half of each chain and focus on the second half as a conservative choice (Gelman, Carlin, Stern, & Rubin, 2004; Spiegelhalter, Thomas, Best, & Lunn, 2003). The practice of discarding early iterations in MCMC is referred to as *burn-in*. The final inferences, after discarding early iterations, will be made based on the assumption that the distributions of the simulated values are close to the population distribution.

Summary

Single-case studies are essential to intensively study the effect of a treatment on a single case over time. Single-case designs have growing interest over many disciplines including education, psychology, and social science due to several advantages that single-case designs have. They provide information about individual effects as well as group effects (Barlow, Nock, & Hersen, 2009). They also allow the study of special population groups that particularly have a low population prevalence rate (Van den Noortgate & Onghena, 2003a). In addition, the characteristics of these designs allow a reduction in the gap between research and practice, and provide a mechanism for cases to serve as their own control (Morgan & Morgan, 2001).

Several non-parametric and parametric methods have been proposed to analyze single-case data including visual analysis, nonoverlap statistics, randomization tests, and regression based methods. In single-case data, it is often considered that the errors are autocorrelated as opposed to be independent, and the possible dependency of the errors should be taken into account in the model. There are several methods available to take autocorrelation into consideration. Particularly, regression based methods can take autocorrelation into consideration, using the GLS regression method for one case or multilevel modeling for multiple cases. By using a multilevel framework, researchers are able to model various dependent error structures, and complex models (e.g., heterogeneity of variance, and the nesting of cases within studies).

Although the multilevel model has the flexibility to handle dependency of the errors, it should be noted that a critical assumption has typically been made in the multilevel approach. Past applications of multilevel modeling to single-case data (e.g., Van Noortgate & Onghena, 2003a, 2003b) as well as methodological studies of multilevel models with single-case data (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens, 2010) have assumed the level-1 error structure is the same for all cases.

It is plausible that the level-1 error structure may not be same across cases (Baek & Ferron, 2011; Baek & Ferron, 2013). Failure to account for variation that exists in a level-1 error structure can impact the inferences of a study. Thus, it is important to examine the consequences of both not modeling between case variation and modeling between case variation in the level-1 error structure. Despite the importance of this issue, neither the effects of non-modeled between case variation effects nor the performance of modeled between case variation effects in the level-1 error structure have been systematically examined.

There are several estimation methods available to make it feasible to allow the level-1 error structure to vary across cases including restricted maximum likelihood (REML) and Bayesian methods. The REML method is the most commonly used method to analyze multilevel models, and has been implemented by several software procedures that allow easy access.

However, the REML has inferential and technical issues associated with analyzing complex multilevel models of single-case data such as non-convergence with more complex models. The Bayesian method has the potential to resolve this issue found with REML. It was found that a complex multilevel model that fails to converge using REML can be run by using the Bayesian approach (Baek, Petit-Bois, & Ferron, 2012).

Therefore, this study will examine the consequences of modeling and not modeling between case variation in level-1 error structure on parameter estimations and inferences for single-case data using Bayesian estimation. Specifically, two level multilevel models where the level-1 error structures are modeled in different ways (i.e., ID, AR(1) constant across cases, and AR(1) varies across cases) will be compared in terms of the quality of the fixed effects (i.e., the overall average baseline intercept, the overall baseline slope, and the overall average treatment effects (shift in level and shift in slope)) and the variance components (i.e., the between case variance in the average baseline intercept, the between case variance in the average baseline slope, the between case variance in the average treatment effect, and the level-1 error variance, and the autocorrelation). This will be achieved by investigating credible interval coverage rates, credible interval widths, RMSE, and bias of the point estimates as a function of specific design and data factors.

CHAPTER THREE: METHODS

This chapter outlines the methods for this study including the purpose, research questions, design, sample and analysis conditions, data generation, and outcome measures.

Purpose

The purpose of this simulation study was to extend the MLM modeling in single-case design to allow between case variation in the level-1 error structure which allows the level-1 error and autocorrelation to vary across cases. This study identified the consequences of not modeling and modeling between case variation in the level-1 error structure for single-case studies using the Bayesian estimation approach. Specifically, two level multilevel models where the level-1 error structures were modeled in different ways (i.e., not modeling between case variation vs. modeling between case variation) were examined in terms of the accuracy of the estimates of the parameters. More specifically, this study investigated credible interval coverage rates, credible interval widths, bias of the point estimates, and root mean squared error (RMSE) as a function of specific design, data, and analysis factors such as number of cases, series length per case, true level-1 error structure, variation in the level-2 errors, and methods to modeling level-1 error structure.

Research Questions

1. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **fixed treatment effect** in single-case design?

    1) to what extent are the ***bias and RMSE for the fixed treatment effects*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation in the level-2 errors)?

    2) to what extent are the ***credible interval coverage and width for the fixed treatment effects*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation in the level-2 errors)?

2. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **variance components** in single-case design?

    1) to what extent are the ***bias and RMSE for the variance components*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation in the level-2 errors)?

    2) to what extent are the ***credible interval coverage and width for the variance components*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation in the level-2 errors)?

Design

This study was conducted with a 2x2x3x2x2 factorial design. These factors included (1) number of cases (4 and 8); (2) series length per case (10 and 20); (3) true level-1 error structure (level-1 error structure as constant across cases (homogeneous), level-1 error structure as varying across cases(moderately heterogeneous and severely heterogeneous); (4) variation in the level-2 errors (most of the variance at level-1 and most of the variance at level-2); (5) analysis methods to modeling level-1 error structure (not modeling between case variation (Model 1), and modeling between case variation(Model 2)). For each of the 48 conditions, 1,000 data sets were generated using SAS IML (SAS Institute Inc., 2008) and analyzed using WinBUGS software. The dependent variables were bias (the average difference between the known parameter value and the parameter estimate for both the fixed effects and the variance components), credible interval coverage (the proportion of 95% credible intervals (equal tailed credible interval) containing both the fixed effects estimates and the variance components), credible interval width (the average difference between the upper and lower limits of the 95% credible intervals (equal tailed credible interval) for both the fixed effects and the variance components), and RMSE (the square root of the average squares of the errors).

Limiting the number of conditions to 48 was partially based on the result of a preliminary pilot test that was conducted prior to the study. The pilot test was conducted to verify the accuracy of the simulation program, and to estimate the approximate amount of time required to run the simulation. For checking the accuracy of the program, a small number of the replications was run for some of the conditions. Datasets and outputs from the analyses were examined to ensure the correct dataset and models were being created and analyzed. For estimating the approximate amount of time to run the simulation, several conditions were run with 1000 replications. The result of the pilot test shows that the amount of time required for each condition varied from the least amount of time being 4 days to the longest amount of time being over two weeks for a condition. The series length per case and the number of cases are two main factors that most affect the amount of time required. As the series length per case and the number of cases increase, the amount of time required to run a simulation increases substantially. Based on this finding, only a limited number of conditions were selected to meet reasonable time period to finish this study.

Sample

The sample for this study was generated through Monte Carlo simulation methods. Three factors were manipulated in this study: (1) data factors, (2) design factors, and (3) analysis factor. The data factors addressed two conditions: true level-1 error structure (how to generate the level-1 error structures) and variation in the level-2 errors. For the true level-1 error structure, three different types of data sets were generated, homogeneous, moderately heterogeneous, and severely heterogeneous error structures. For the homogeneous error structure, the level-1 error structure was generated as constant across cases. For the moderately and severely heterogeneous error structures, the level-1 error structure was generated as varying across cases. Design factors addressed specific values of the following two conditions: number of cases and series length per case. The analysis factor addressed how to model the level-1 error structure, Model 1 and Model 2. For Model 1, the level-1 error structure was assumed and analyzed as constant across cases. For Model 2, the level-1 error structure was assumed and analyzed as varying across cases. The data, design, and analysis factors which were used to define the simulated data are further defined below.

Data factors

True level-1 error structure. Two different types of data sets were generated depending on how the level-1 error structure was modeled, homogeneous error structure and heterogeneous error structures. The general equations used to generate data are presented in equations (7) and (8).

Level-1 equation:

$$y_{ij} = \beta_{0j} + \beta_{1j}\ Phase_{ij} + \beta_{2j}\ Time_{ij} + \beta_{3j}\ Time_{ij} *Phase_{ij} + e_{ij} \qquad (7)$$

Level-2 equation:

$$\beta_{0j} = \theta_{00} + u_{0j} \qquad (8)$$

$$\beta_{1j} = \theta_{10} + u_{1j}$$

$$\beta_{2j} = \theta_{20} + u_{2j}$$

$$\beta_{3j} = \theta_{30} + u_{3j}$$

where y_{ij} was the observed value (outcome) at the ith observation for the jth case. β_{0j} was the baseline intercept for the jth case and β_{1j} was the difference between the baseline level and the treatment level (shift in level) for the jth case when $Time_{ij}$ was equal to 0. β_{2j} was the baseline slope for the jth case, and β_{3j} was the change in slopes between the baseline phase and the treatment phase (shift in slope). For the interaction term ($Time_{ij}$ $*Phase_{ij}$), $Time_{ij}$ was centered so that 0 corresponds to the first observation of the treatment phase. e_{ij} was the residual that indicates within case variation (level-1 errors) and was assumed to be multivariate normally distributed N(0,Σ_e). In this study, Σ_e was assumed to follow first-order autoregressive error structure, AR(1). For the level-2 equation, θ_{00} was the average baseline intercept and θ_{10} was the average shift in level at $Time_{ij}$ which was equal to 0, θ_{20} was the average baseline slope, and θ_{30} was the average shift in slope. u_{0j} , u_{1j} , u_{2j} and u_{3j} were level-2 errors and were assumed to be multivariate normally distributed $N(0,\Sigma_u)$. In this study, the fixed effect value was fixed for both data sets so that the average baseline intercept (θ_{00}) and the average baseline slope (θ_{20}) were 1, and the shift in level (θ_{10}) was 2 and the shift in slope (θ_{30}) was .2.

Although both homogeneous error structure and heterogeneous error structure data sets were generated using this same general equation, they were distinguished by how the level-1 error structure was generated. For the homogeneous error structure, the level-1 error structure was generated using the ARMASIM function in SAS version 9.3 (SAS Institute, 2008) with a level-1 error variance of 1.0 and autocorrelation values of .2. This led to all cases included in the study having the same value of level-1 error variance and autocorrelation for each condition. For the moderately heterogeneous error structure, the level-1 error structure was also be generated using the ARMASIM function, but values of autocorrelation and level-1 error variance were generated from a normal distribution using the RANNOR random number generator, and from a uniform distribution using the RANUNI random number generator in SAS version 9.3 (SAS Institute, 2008), respectively. For the autocorrelation, the normal distribution followed a mean of .2 and a standard deviation of .1 for the moderately heterogeneous, and the normal distribution followed a mean of .2 and a standard deviation of .2 for the severely heterogeneous error structure. The mean value of the autocorrelation .2 had been selected based on the literature review of single-case designs. According to the survey conducted by Shadish and Sullivan (2011), the average autocorrelation value of the studies reviewed was .2, after correcting for sampling errors. The values of the standard deviation of .1 and .2 were selected based on a consideration of possible range of autocorrelation distribution. The mean of .2 with standard deviation of .1 creates a distribution that 99% of the autocorrelation values fall between -.1 and .5. The mean of .2 with standard deviation of .2 creates a distribution that 99% of the autocorrelation values fall between - .4 and .8. The range of these values is covered the possible autocorrelation values typically found in behavior research (Huitema, 1985; Matyas & Greenwood, 1996; Shadish & Sullivan, 2011). For the level-1 error variance, standard deviation unit was used. The uniform distribution of the level-1 error standard deviation with a lower bound of .7 and an upper bound of 1.3 led the uniform distribution to follow a mean of 1 and a standard deviation of .17 for the moderately heterogeneous, and the uniform distribution of the level-1 error standard deviation with a lower bound of .4 and an upper bound of 1.6 led the uniform distribution to follow a mean of 1 and a standard deviation of .35 for the severely heterogeneous. This process was led to every case included in the study to have their unique value of level-1 error standard deviation and autocorrelation within a specified range. The level-1 error standard deviation were generated in the way the largest level-1 error variance $((1.3)^2)$ value can be either as much as 3.5 times of the smallest level-1 error variance value $((.7)^2)$ or as much as 16 times $((1.6)^2)$ of the smallest level-1 error variance value $((.4)^2)$. The motivation for this rationale was based on the analyses of real datasets. Baek and Ferron (2013) found that when they allowed the level-1 error variance to vary across cases in real datasets, the largest level-1 error variance tended to be about average four times the smallest, and ranged up to 16 times the smallest.

For all data sets, the level-2 errors were generated from a normal distribution using the RANNOR random number generator in SAS version 9.3 (SAS Institute, 2008). For each of the 24 conditions (not included the analysis methods design), 1,000 data sets of homogeneous, moderately heterogeneous, and severely heterogeneous data sets were generated which led to a total of 72,000 datasets being generated.

Variation in the level-2 errors. The variation in the level-2 errors had two levels (most variance at the level-1 and most variance at the level-2). The previous simulation studies either had the most variance at level-1(Ferron et al., 2009; Van den Noortgate, 2008), or had most of the variance at the higher levels (level-2 or level-3) (Van den Noortgate, 2008). Their simulation studies were motivated by analyses of real datasets where it was found that in some studies the largest variance component was at level-1 whereas in other studies the largest variance components were at level 2. Based on these finding, both cases were incorporated into this study. The average value of level-1 error variance was fixed to 1.0. The first category will model the data having most of the variance at the level-1, so that the level-2 error variances in intercept, phase, time, and interaction had the values of .5, .5, .05, and .05, respectively. It was assumed that there was no covariance among level-2 errors. The second category modeled the data having most of the variance at the level-2, so that the level-2 error variance in intercept, phase, time, and interaction had the values of 2, 2, .2, and .2, respectively.

Design factors

Number of cases. The number of participants had two levels (small and large). The small category included 4 participants, and the large category included 8 participants.

These numbers had been selected based on previous findings of single-case studies. Farmer, Owens, Ferron, and Allsopp (2010) found that the average number of participants per single-case study are less than or equal to 7. Another study that reviewed published single-case studies found that the number of participants or sample size per single-case study falls between 1 and 13, with an average of 3.64 (Shadish & Sullivan, 2011). Some applied studies that synthesized published single-case studies also found that the average number of participants per study was 3.25 (Petit-Bois, 2012) and 4.60 (Baek, Petit-Bois, & Ferron, 2012). In addition, Kazdin (2011) suggests that a minimum of three or more baselines are recommended to see a treatment effect. He states that 8 or 9 baselines (participants, settings, and behaviors) are needed in order to see clear treatment effects.

Previous Monte Carlo simulation studies have been conducted for single-case studies using 4 or 8 participants (Owens & Ferron, 2011; Petit-Bois, in press), and 4 or 7 participants (Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012).

Series length per case. The series length per case had two levels (small and moderate). The small category included series lengths of 10, and the moderate category included series lengths of 20. Previous studies were used to determine the series lengths for this study. Shadish and Sullivan (2011) found that 90% of the studies reviewed had 49 or fewer observations. In addition, previous simulation studies in this area used series lengths of 10, 20 and 30 (Ferron et al., 2009; Ferron, Farmer, & Owens, 2010; Owens & Ferron, 2011), or 10 and 30 (Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012). Only two levels of the series length per case were chosen due to the great impact on the amount of time to run the simulation. These selected values cover small to moderate series lengths found in the previous studies.

Analysis factor

Two different methods of modeling level-1 error structure were applied to the generated data (both homogeneous and heterogeneous error structures). The first method was modeling the level-1 error structure to be constant across cases (Model 1). The second method was modeling the level-1 error structure to vary across cases (Model 2). This cross effect provides in-depth information about the performance of the proposed idea. More detailed information about Model 1 and Model 2 is in the following section.

Analysis of Each Simulated Data Set

Equations for the specified models (Model 1 and Model 2)

Each data set was analyzed using the two different models. The two level models were estimated using the Bayesian estimation method via WinBUGS software version 1.4.3 which uses a Gibbs sampler. The equations of two-level single-case design (equations (7) and (8)) used for this study can also be expressed using Bayesian forms (probability distributions) as shown in below. Equation (9) was for Model 1 that assumed the first-order autoregressive structure for the level-1 error structure where the autocorrelation and the within error variance were assumed constant across cases. This equation is an extension from equation (6) in that the equation includes the autocorrelation parameter (ρ).

$$y_{\mathrm{ij}} \sim \mathrm{Normal}(\theta_{\mathrm{ij},}\, \sigma^2) \qquad (9)$$

$$\mu_{\mathrm{ij}} = \alpha_{\mathrm{j}} + \beta_{\mathrm{j}} \mathit{Time}_{ij} + \gamma_{\mathrm{j}} \mathit{Phase}_{ij} + \delta_{\mathrm{j}} \mathit{Time}_{ij} {*} \mathit{Phase}_{ij}$$

$$\theta_{0\mathrm{j}} = \mu_{0\mathrm{j}}$$

$$\theta_{\mathrm{ij}} = \mu_{\mathrm{ij}} + \rho\, (y_{(\mathrm{i}-1)\mathrm{j}} - \mu_{(\mathrm{i}-1)\mathrm{j}}) \quad (i \geq 1)$$

$$\alpha_{\mathrm{j}} \sim \mathrm{Normal}(\mu_{\alpha}, \sigma^2{}_{\alpha})$$

$$\beta_{\mathrm{j}} \sim \mathrm{Normal}(\mu_{\beta}, \sigma^2{}_{\beta})$$

$$\gamma_{\mathrm{j}} \sim \mathrm{Normal}(\mu_{\gamma}, \sigma^2{}_{\gamma})$$

$$\delta_{\mathrm{j}} \sim \mathrm{Normal}(\mu_{\delta}, \sigma^2{}_{\delta})$$

where y_{ij} was the observed value (outcome) for the i^{th} observation at the j^{th} case, and follows normal distribution as a prior distribution with the mean of θ_{ij} instead of μ_{ij}, and variance of σ^2 ; θ_{ij} was defined by adding the correlated error term between the adjacent two time points to the $\mu_{ij,}$ where ρ represented the autocorrelation, and $(y_{(i-1)j} - \mu_{(i-1)j})$ represented the error term in the *i-1* time point. When i=0, θ_{0j} was same as μ_{0j} ; α_j was the intercept of the baseline for the j^{th} case; β_j was the baseline slope for the j^{th} case; γ_j was the shift in level for the j^{th} case; δ_j was the shift in slope for the j^{th} case. σ^2 was the error variance that leads to within-case variation. It was assumed that all regression coefficeints, $\alpha_{j,}$ $\beta_{j,}$ $\gamma_{j,}$ δ_j, follow a common prior distribution (Gelman, Carlin, Stern, & Rubin, 2004; Gelman, 2006). In this study, normal distributions were assigned as prior distributions for all parameters. More detailed information about how to model the prior distributions is in the following section.

For the second level equation, μ_α was the average intercept of the baseline; μ_β was the average baseline slope; μ_γ was the average shift in level; μ_δ was the average shift in slope, and σ^2_α, σ^2_β, σ^2_γ, and σ^2_δ are corresponding error variances.

Model 2 could be further developed from Model 1 with modeling between case variation in the level-1 error structure which can be accomplished by changing σ^2 to σ^2_j and ρ to ρ_j which indicated the values were specified to the j^{th} case. Model 2 was defined in the same way that Model 1 was defined where intercept, baseline slope, shift in level, and shift in slope were included and they were all allowed to vary across cases. Model 1 and Model 2 were distinguished only in the way to model the level-1 error structure. In Model 2, the level-1 error variance and autocorrelation were allowed to vary across cases (j) as follows:

$$y_{ij} \sim \text{Normal}(\theta_{ij,}\ \sigma_{\mathbf{j}}^2) \qquad (9)$$

$$\mu_{ij} = \alpha_j + \beta_j Time_{ij} + \gamma_j Phase_{ij} + \delta_j Time_{ij} * Phase_{ij}$$

$$\theta_{0j} = \mu_{0j}$$

$$\theta_{ij} = \mu_{ij} + \rho_{\boldsymbol{j}} \, (y_{(i-1)j} - \mu_{(i-1)j}) \quad (i \geq 1)$$

$$\alpha_j \sim \mathrm{Normal}(\mu_\alpha, \sigma^2_\alpha)$$

$$\beta_j \sim \mathrm{Normal}(\mu_\beta, \sigma^2_\beta)$$

$$\gamma_j \sim \mathrm{Normal}(\mu_\gamma, \sigma^2_\gamma)$$

$$\delta_j \sim \mathrm{Normal}(\mu_\delta, \sigma^2_\delta)$$

$$\sigma_j \sim \mathrm{Uniform}(L_\sigma, U_\sigma)$$

$$\rho_j \sim \mathrm{Normal}\ (\mu_\rho, \sigma^2_\rho)\ I\ (-1 < \rho_j < 1)$$

Prior distributions for the parameters

A common prior distribution used in applied work for μα , μβ , μγ , and μδ is a noninformative normal distribution with a mean of 0 and a variance of 10002, and σ is a uniform distribution with the lower limit of 0 and the upper limit of 100. Thus, these prior distributions were constructed for the fixed effect (i.e., μα , μβ , μγ , and μδ) and level-1 error standard deviation (σ) in this study.

$$\mu_{\alpha\,,}\ \mu_\beta\ ,\ \mu_\gamma\ ,\ \mu_\delta \sim \mathrm{Normal}(0,\ 1000^2)$$

$$\sigma \sim \mathrm{Uniform}(0,\ 100)$$

For the fixed effect, noninformative normal distributions were constructed with large variance (i.e., 1000^2), so that posterior inferences could not be influenced by the choice of variance value. Similarly, for the level-1 error variance, the uniform distribution was constructed with the large upper limit of σ (standard deviation unit of the level-1 error variance). The value of 100 was considered as sufficiently large because the true value of σ was set as 1 in this study. The lower limit of σ was set to 0 due to the fact that the value of the standard deviation cannot be negative.

In addition, uniform distributions were assigned to be the priors for the level-2 error variance parameters (i.e., σ^2_α , σ^2_β, σ^2_γ, and σ^2_δ) by Gelman (2006)'s recommendation. Specifically, the noninformative prior distributions for the standard deviation unit of the level-2 error variance (σ_α , σ_β, σ_γ, and σ_δ) were assigned to be the uniform distribution with the lower limit of 0 and the upper limit of 100.

$$\sigma_\alpha , \sigma_\beta, \sigma_\gamma, \sigma_\delta \sim \text{Uniform}(0, 100)$$

For autocorrelation, ρ, a reasonable noninformative prior distribution can be a normal distribution. Shadish and Sullivan (2011) summarize the characteristics of single-case designs using 809 published studies. The characteristics include types of designs, outcome variables, cases per study, series length per case, number of phases, and autocorrelations. In their report, the histogram of the autocorrelation among the published studies seems to follow a normal distribution ranging from -.931 to .786. Thus, the noninformative prior for ρ that follows a normal distribution with a mean of 0 and a standard deviation (σ) of 1000 was assigned. However, since ρ is a correlation parameter, the scale of this parameter should be the same as a correlation scale, from -1 to 1. Therefore, the scale of the prior distribution for ρ was stationary restricted so that its range falls between -1 and 1 (Gamerman & Lopes, 2006).

$$\rho \sim \text{Normal }(0, 1000^2) \text{ I } (-1 < \rho < 1)$$

Since no one has worked through the proposed idea that the level-1 error structure could vary across cases, no literature was found to define priors for σ_j and ρ_j. This study had suggested one possible theoretical way to construct the priors for σ_j and ρ_j as follows:

$$\begin{aligned} \sigma_j \sim \text{Uniform}(L_\sigma, U_\sigma) \text{ with } L_\sigma &\sim \text{Uniform}(0, 100) \\ U_\sigma &\sim \text{Uniform}(L_\sigma, 100) \end{aligned}$$

$$\begin{aligned} \rho_j \sim \text{Normal }(\mu_\rho, \sigma^2_\rho) \text{ I } (-1 < \rho_j < 1) \text{ with } \mu_\rho &\sim \text{Normal }(0, 1000^2) \\ \sigma_\rho &\sim \text{Uniform}(0, 100) \end{aligned}$$

The prior for σ_j could simply be assumed to follow the same prior that σ follows, which is the uniform distribution with the lower limit of L_σ and the upper limit of U_σ. The lower limit of L_σ can be assumed to follow a uniform distribution with the lower limit of 0 and the upper limit of 100. The upper limit of U_σ can be also assumed to follow a uniform distribution but with the lower limit of 0, and the upper limit of L_σ since the U_σ value should be bigger than the L_σ value. The mean and the standard deviation of the uniform distribution for σ_j will be computed using the following formula: $\frac{L\sigma+U\sigma}{2}$ and $\sqrt{\frac{|U\sigma-L\sigma|^2}{12}}$, respectively.

A reasonable way to construct the prior for ρ_j is to assume the same prior used to construct ρ. One can assume that ρ_j follows the same prior that ρ follows, which is the normal distribution with a mean of μ_ρ and a variance of σ^2_ρ but with the restricted range between -1 and 1. The μ_ρ and σ_ρ could be further defined as a normal distribution with a mean of 0 and a variance of 1000^2 for $\mu_{\rho\,,}$ and a uniform distribution with the lower limit of 0 and the upper limit of 100 for σ_ρ.

Convergence criteria for the analysis

Pilot simulation data were generated to test convergence and to make decisions about the number of iterations, and the burn-in period. A data set per each condition of the design factors (24 data) was created and run with two models (Model 1 and Model 2). This ended up testing all 48 conditions. The various diagnostic criteria were used in monitoring convergence, including trace plots, history plots, Kernel density plots, and Brooks–Gelman–Rubin (BGR) plots for the created data set using two different MCMC chains. The specific information about each criterion is illustrated in Figures 8 through 10.

Trace or history plots. One of the intuitive diagnostic criteria is a trace plot or history plot which plots the parameter value at time against the iteration number. Trace plot is dynamic, being redrawn each time the screen is redrawn, and history plot is showing a complete trace for the targeted variables. When more than one chain is assigned simultaneously, the trace and history plots show each chain in a different color. If all the chains overlap one another, we can be confident to say that convergence has been achieved (see Spiegelhalter, Thomas, Best, & Lunn, 2003). A clear sign of non-convergence occurs when we observe some trends in the plots. An example of trace and history plots is illustrated in Figure 8. In the figure, two chains are assigned simultaneously, and overall the convergence looks reasonable since both chains appear to be overlapping each other.

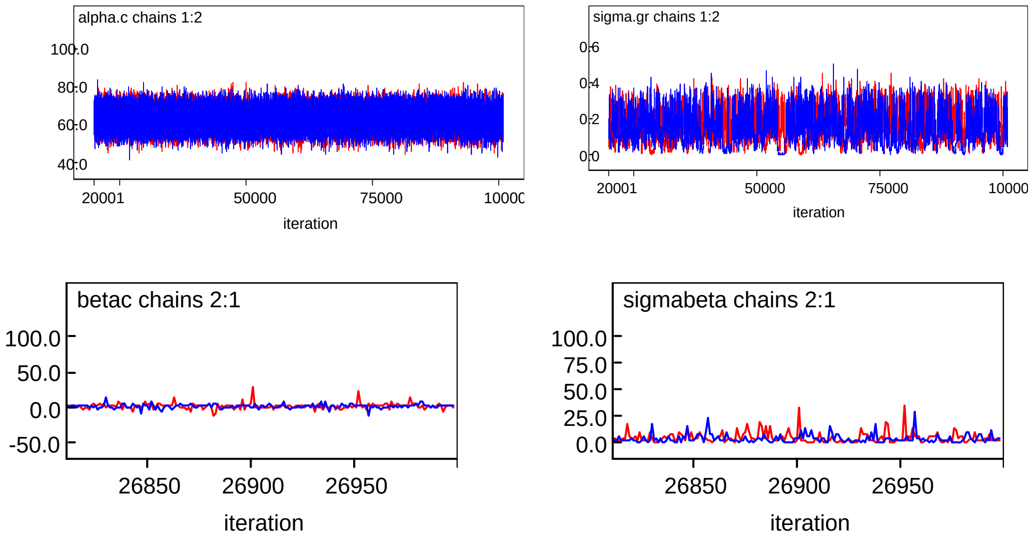


Figure 8 An Example of Trace and History plots (first raw: history plots, second raw: trace plots)

Kernel density plots (Posterior distributions of each parameter). Kernel density plot shows the final posterior distribution of the estimated parameter. This plot could be another useful diagnostic criterion. When converge occurs, the distribution shows a smooth shape.

Generally, as more iterations are performed, the distribution would become smoother. Figure 9 shows an example of the Kernel density plots. The convergence looks reasonable in that the distributions show a smooth shape. From the plot, we can also see the range of possible values for each parameter and which values are more likely than others.

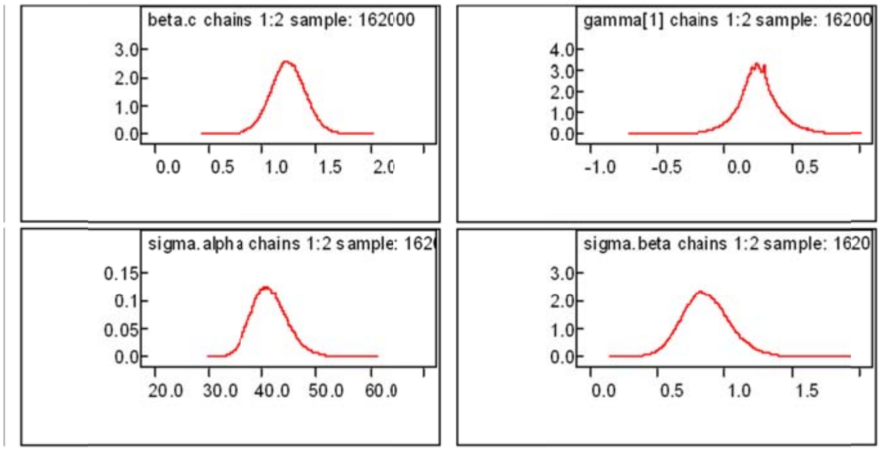


Figure 9 An example of Kernel density plots

Brooks–Gelman–Rubin (BGR) plots. Brooks-Gelman-Rubin (BGR) diagnostic is computed based on the ratio of between-within chain variances (Brooks & Gelman, 1997; Brooks & Roberts, 1998; Cowles & Carlin, 1996; Gelman & Rubin, 1992). The intuition is that the variance within the chains should be the same as the variance across the chains. BGR plots have three lines; Green lines represent the normalized width of the central 80% interval of the pooled, blue lines represent the normalized average width of the 80% intervals within the individual, and red lines represent the BGR statistic, R. When R converges to 1, and both the pooled and within interval widths converge with stability, we consider convergence occurred. Figure 10 shows an example of BGR plots. In the figure, the convergence looks reasonable since three lines converge to one with stability.

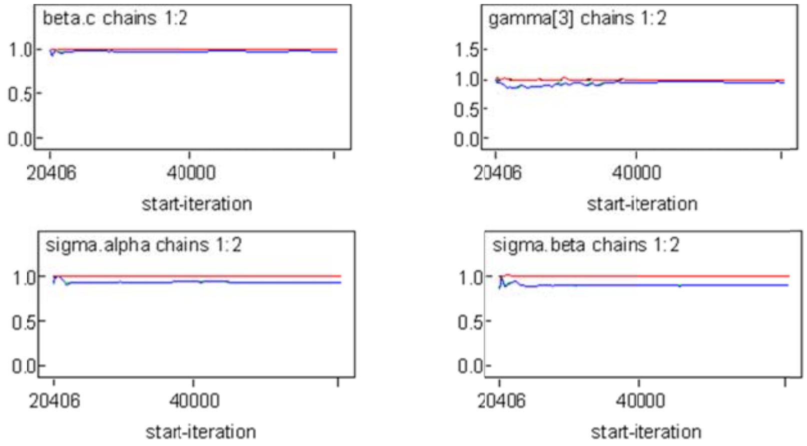


Figure 10 An example of BGR plots

MC errors. Monte Carlo error (*MC error*) will also be tracked to check the computational accuracy of the posterior estimates. This indicates a difference between the mean of the sampled values (the estimated posterior mean for each parameter) and the true posterior mean. Typically, the simulation should be run until the *MC error* for each parameter is less than 5% of the sample standard deviation (*sd*) to obtain a reliable estimate of the parameter.

Analysis to Estimate Bias of the Point Estimates, Credible Interval Coverage, Credible Interval Width, and Root Mean Squared Error

Bias, credible interval coverage, credible interval width, and RMSE were the dependent variables for the six independent variables (Data, design, and analysis factors). Bias for the average treatment effects and average variances of treatment effects parameters (shift in level, shift in slope, level-1 error variances) was computed as the average difference between the known value of parameters and the estimated posterior mean value of the parameters. The equation of the bias is shown below:

$$bias = \frac{\sum_{n=1}^{1000}(\hat{\gamma} - \gamma)}{1000}$$

The deviation between the known value of parameters and the estimated value of the parameters $(\hat{\gamma} - \gamma\ )$ was first aggregated across 1000 replications within each condition [$\sum_{n=1}^{1000}(\hat{\gamma} - \gamma)$] and then was divided by 1000 to obtain an average bias value. Bias for the level-1 error variance and autocorrelation parameters were also computed as the average difference between the known value of parameters and the estimated posterior mean value of the parameters. However, since the level-1 error variance and the autocorrelation parameters were generated to vary across cases for heterogeneous error structure data sets, and estimated to vary across cases for Model 2, bias for the level-1 error variance and autocorrelation parameters were computed as the average difference between the known value of parameters for each case and the estimated posterior mean value of the parameters for each case. The equation of the bias is shown below:

$$bias = \frac{\sum_{n=1}^{1000}\left(\frac{\sum_{i=1}^{m}(\widehat{\gamma_i} - \gamma_i)}{m}\right)}{1000}$$

The deviation between the known value of parameters for each case and the estimated value of the parameters $(\widehat{\gamma_i} - \gamma_i)$ was first aggregated across the number of cases per each replication $[(\sum_{i=1}^{m}(\widehat{\gamma_i} - \gamma_i)]$ and then divided by the number of cases *m* to obtain an average bias value per each replication $[(\frac{\sum_{i=1}^{m}(\widehat{\gamma_i}-\gamma_i)}{m})]$. This average bias value per each replication was then aggregated across 1000 replications within each condition $[\sum_{n=1}^{1000}(\frac{\sum_{i=1}^{m}(\widehat{\gamma_i}-\gamma_i)}{m})]$ and then divided by 1000 to obtain an average bias value.

Relative bias for parameters whose known value is anything other than 1.0 or 0 was also computed which can be represented as a percentage of the known parameter value. Since relative bias is represented by percentages rather than a value, this statistic allows comparisons of bias among parameters that have different scales of the value. Relative bias for the average treatment effects and average variances of treatment effects parameters (shift in level, shift in slope, level-1 error variances) was computed as the average difference between the known value of parameters and the estimated value of the parameters divided by the known parameter values. The equation of the relative bias is shown below:

$$relative\ bias = \frac{\sum_{n=1}^{1000}(\frac{\hat{\gamma} - \gamma}{\gamma})}{1000}$$

The deviation between the known value of a parameter and the estimated value of the parameter divided by the known value of the parameter $(\frac{\hat{\gamma}-\gamma}{\gamma})$ was first aggregated across 1000 replications within each condition [$\sum_{n=1}^{1000}(\frac{\hat{\gamma}-\gamma}{\gamma})$] and then was divided by 1000 to obtain an average relative bias value. Relative bias for the level-1 error variance and autocorrelation parameters was computed as the average difference between the known value of parameters for each case and the estimated value of the parameters for each case divided by the known parameter values for each case. The equation of the relative bias is shown below:

$$relative\ bias = \frac{\sum_{n=1}^{1000}(\frac{\sum_{i=1}^{m}\frac{(\hat{\gamma}_i - \gamma_i)}{\gamma_i}}{m})}{1000}$$

The deviation between the known value of a parameter for each case and the estimated value of the parameter for each case divided by the known value of the parameter for each case $(\frac{(\hat{\gamma}_i-\gamma_i)}{\gamma_i})$ was first aggregated across the number of cases per each replication [($\sum_{i=1}^{m}\frac{(\hat{\gamma}_i-\gamma_i)}{\gamma_i}$)] and then divided by the number of cases *m* to obtain an average relative bias value per each condition [$\frac{\sum_{i=1}^{m}\frac{(\hat{\gamma}_i-\gamma_i)}{\gamma_i}}{m}$]. This average relative bias value per each condition was then aggregated across 1000 replications within each condition [$\sum_{n=1}^{1000}(\frac{\sum_{i=1}^{m}\frac{(\widehat{\gamma_\iota}-\gamma_i)}{\gamma_i}}{m})$] and then divided by 1000 to obtain an average relative bias value.

The root mean squared error for the average treatment effects and average variances of treatment effects parameters (shift in level, shift in slope, level-1 error variances) was computed as the square root of the average sums of the squares of the errors. The equation of the RMSE is shown below:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{1000}(\hat{\gamma}-\gamma)^2}{1000}}$$

The squared deviation between the known value of a parameter and the estimated value of the parameter [$(\hat{\gamma}-\gamma)^2$] was first aggregated across 1000 replications within each condition [$\sum_{n=1}^{1000}(\hat{\gamma}-\gamma)^2$] and then was divided by 1000, and the average RMSE value was obtained through the square root of the entire equation. The root mean squared error for the level-1 error variance and the autocorrelation was computed as the square root of the average sums of the squares of the errors for each case. The equation of the RMSE is shown below:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{1000}(\frac{\sum_{i=1}^{m}(\widehat{\gamma_\iota}-\gamma_i)^2}{m})}{1000}}$$

The squared deviation between the known value of a parameter for each case and the estimated value of the parameter for each case [$(\widehat{\gamma_\iota}-\gamma_i)^2$] was first aggregated across the number of cases per each replication [$\sum_{i=1}^{m}(\widehat{\gamma_\iota}-\gamma_i)^2$] and then divided by the number of cases m to obtain an average squared deviation value per each replication. This average squared deviation per each replication was then aggregated across 1000 replications within each condition

[$\sum_{n=1}^{1000}(\frac{\sum_{i=1}^{m}(\widehat{\gamma_\imath}-\gamma_i)^2}{m})$] and then divided by 1000, and the average RMSE value was obtained through the square root of the entire equation.

Credible interval coverage was computed as proportion of the 95% credible interval (equal tailed credible interval) that contains the known parameter value. The credible interval width was computed as the average difference between the upper and lower limits of the 95% credible intervals (equal tailed credible interval). These statistics were aggregated across 1000 replications within each condition to represent the average values of the statistics.

Analyses to Examine Relationships between Data, Design, and Analysis Factors and Bias of the Point Estimates, Credible Interval Coverage, and Credible Interval Width, and Root Mean Squared Error

Box and whisker plots along with general linear modeling (GLM) were examined to evaluate the bias estimate and RMSE of each parameter. Box and whisker plots illustrated the distribution of the bias and the RMSE estimate of the each parameter across the simulation conditions. GLM illustrated the explained variability of the bias and the RMSE estimates associated with each parameter as a function of the main effects of and interaction effects between the design, data, and analysis factors to inform the source of bias and error. A main effect only model was built first, then two-way or three-way interactions were added in the model. If a main effect only model explained a significant proportion of the variability (at least 94% of the total variability), then no further models were investigated. However, if the model failed to explain the minimum variability, then interactions were included in the model. The effects size, eta-squared (η^2), was also calculated to determine the proportion of variability associated with each effect. The eta-squared value of the each effect was compared to Cohen's

(1988) criteria to determine the size of each effect. According to the criteria, a small effect size is .01, a medium effect size is .06, and a large effect size is .14 or greater. Finally the line graphs were created for a factor that has a medium or larger effect ($\eta^2 \geq .06$) to illustrate the relationship between the different level of the factor and the dependent variables.

CHAPTER FOUR: RESULTS

This chapter provides the results of the research questions. The chapter starts with describing how the results were obtained, and then displays convergence information (trace plots, history plots, Kernel density plots, and Brooks–Gelman–Rubin (BGR) plots along with MC error). Then the outcome measures (bias, RMSE, credible interval coverage and width) of the fixed treatment effects and the variance components are provided in sequential order. The following research questions were addressed:

3. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **fixed treatment effects** in single-case design?
    1) to what extent are the ***bias and RMSE for the fixed treatment effects*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variance of level-2 errors)?
    2) to what extent are the ***credible interval coverage and width for the fixed treatment effects*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variance of level-2 errors)?
4. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **variance components** in single-case design?

1) to what extent are the ***bias and RMSE for the variance components*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variance of level-2 errors)?

2) to what extent are the ***credible interval coverage and width for the variance components*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variance of level-2 errors)?

There were 48 conditions simulated using the five factors in this Monte Carlo study. These factors were (1) number of cases (4 and 8); (2) series length per case (10 and 20); (3) true level-1 error structure (homogeneous , moderately heterogeneous, and severely heterogeneous; (4) variation in the level-2 errors (most of the variance at level-1 and most of the variance at level-2); and (5) analysis methods to modeling level-1 error structure (not modeling between case variation (Model 1), and modeling between case variation(Model 2). This yielded a 2x2x3x2x2 factorial design.

A small set of data sets were first generated to test convergence and to make decisions about the number of iterations, and the burn-in period. A data set per each condition of the design factors (24 conditions) was created and run with two models (Model 1 and Model 2). The various diagnostic criteria were used in monitoring convergence, including trace plots, history plots, Kernel density plots, and Brooks–Gelman–Rubin (BGR) plots for the created data sets using two different MCMC chains. The initial values of the first chain were randomly given for all parameters and the initial values of the second chain were generated for all parameters using a *gen inits* option in WinBUGS software. In WinBUGS, the initial values are generated by sampling either from the prior or from an approximation to the prior given in the model. The MC error was also tracked for all parameters to check the computational accuracy of the posterior estimates. Specifically, the MC error of each parameter was examined if it was less than 5%.

Next, the outcome measures (bias, RMSE, credible interval coverage and width) were evaluated for the fixed treatment effects and the variance components. In addition, relative bias was calculated for the parameter where its value was not equal to 1. The relationship between five factors (number of cases, series length per case, true level-1 error structure, variation in the level-2 errors, and analysis methods to modeling level-1 error structure) and outcome measures (bias, RMSE, confidence interval coverage and width) were then evaluated using PROC GLM in SAS. Models were built to find medium effects or larger (eta-squared values (η^2) were equal to or greater than .06). The η^2 value is measuring the degree of association between the outcome measures and the main and interaction effects of the independent variables (five factors). The $\eta^2$is the proportion of variability of each outcome measure that is associated with each of the effects in the simulation study. It is computed as the ratio of the effect variance (SS_{effect}) to the total variance (SS_{total}).

$$\eta^2 = SS_{effect} / SS_{total}$$

The computed η^2 values were interpreted using Cohen's (1988) standards with a small effect size as .01; a medium effect as .06; and a large effect as .14 or greater. Each model was first built as a main effects only model, and if this model explained at least 94% of the total variability then no interaction effects were included. However, if the model explained less than 94% of the total variability, then interactions (two or higher order interactions, sequentially) were added until the model explained at least 94% of the total variability. For the independent variables (both main and interaction effects) that showed η^2values of .06 and larger, box plots and line graphs were created to further examine the association with outcomes of interest.

The results of the fixed treatment effects and variance components were also looked by three different types of specifications in the level-1 error structure: under-specified (i.e., Model 1 when the data were generated to be heterogeneous), correctly-specified (i.e., Model 1 when the data were generated to be homogeneous, or Model 2 when the data were generated to be heterogeneous), and over-specified (i.e., Model 2 when the data were generated to be homogeneous).

Convergence

In order to meet convergence criteria, a very long run of iterations was required because of the complex models used in this study. As the complexity of the model to be estimated increased (i.e., more parameters to estimate), longer iteration time was required. Therefore, when the data were analyzed with Model 2, it required more iterations than when the data were analyzed by Model 1. In addition, the parameters that had the most difficulty meeting the convergence criteria were the level-2 error standard deviation parameters, especially the level-2 error standard deviation of phase parameter. It was more difficult to meet the convergence criteria when the number of cases was small (4), than large (8). One possible reason that the level-2 error standard deviation parameters presented more difficulty in meeting the convergence criteria is because number of units at level-2 (case) is relatively small compared to the number of units at level-1(series length).

After checking all simulated data sets for convergence analyses (24 data sets), it was decided to use a burn-in of 2,000 iterations and to run an additional 500,000 iterations, but to use only 50,000 samples of the 500,000 iterations to form the posterior distribution for the main analyses. *Thinning* is a technique that can help reduce storage requirements when very long iteration chains need to run. The samples from every kth iteration are stored by using the value of *thin* k. In this study, 50,000 samples were used to form the posterior distribution and *thin* was set to be 10, so a total of 500,000 (10*50,000) iterations were actually run, of which 50,000 samples (every 10th) were stored.

The 50,000 samples were twice the required sample to form the posterior distribution. The required sample was 25,000 samples (after thinning to select 1 in every 10 iterations) based on estimates of the parameters and the models that required the longest iteration. They were the level-2 error standard deviations parameters estimated by Model 2 with the number of cases 4. Once the required sample size, 25,000 was selected based on the various convergence criteria and MC error statistics, the final sample size 50,000 was selected as double of the required sample size to be make sure that all simulated samples would reach the convergence criteria.

More detailed information about each convergence criteria follows. In the generated data sets for the convergence test (24 data sets), more than 10 parameters for Model 1 and more than 18 parameters for Model 2 were estimated that yield a total of over 336 parameters to be estimated. Therefore, only convergence results of some of the parameters were provided in detail. Since the level-2 error standard deviations were the most difficult parameters to reach convergence criteria, the results of the convergence criteria were provided for those parameters along with some of the fixed treatment effect parameters.

Trace and History Plots

The trace and history plots of the level-2 error standard deviation for phase and the interaction, and the average treatment effect for phase parameters were illustrated in Figure 11.

In this analysis, two chains were assigned simultaneously, and overall the convergence looks reasonable since both chains appear to be overlapping each other.

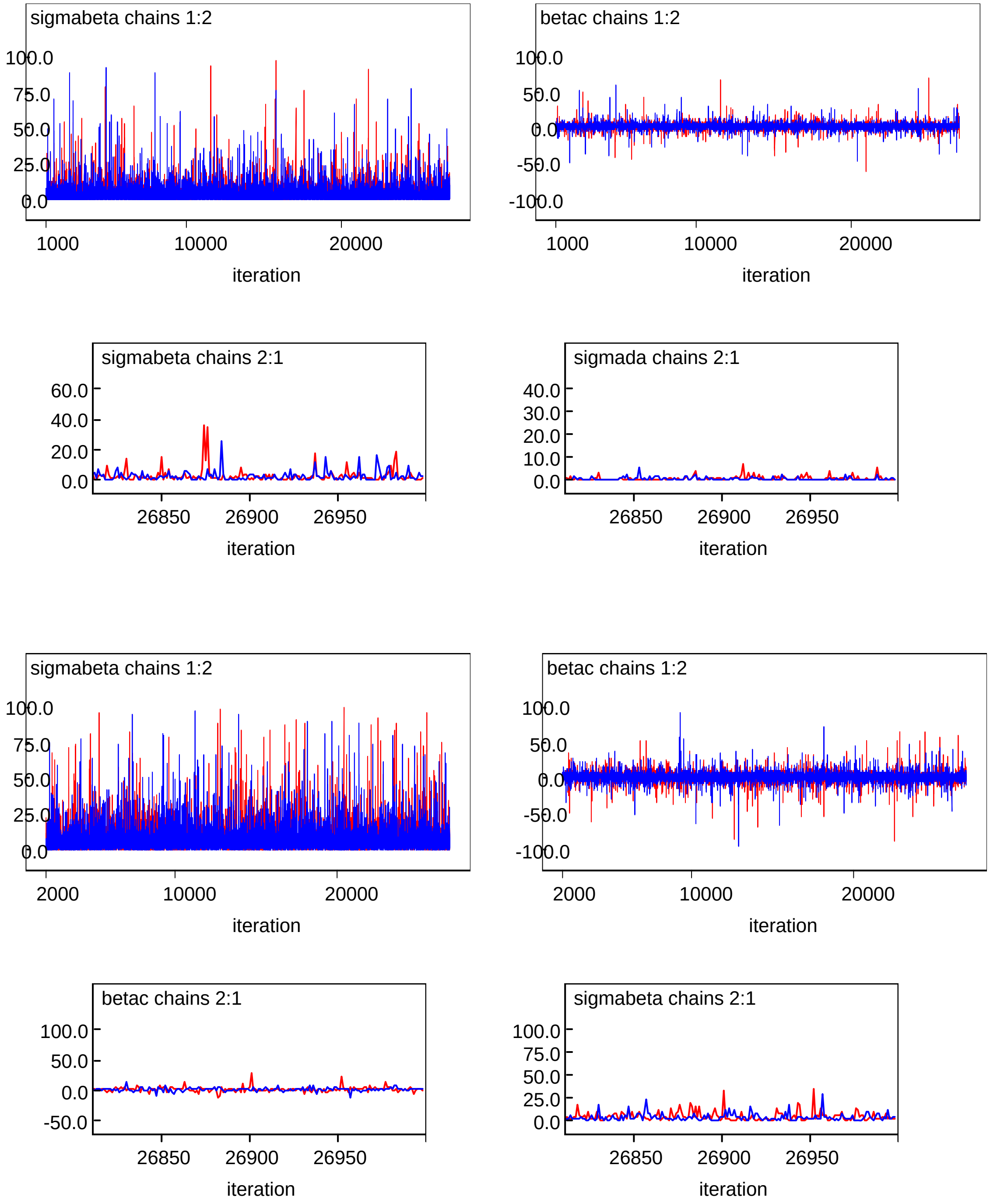


Figure 11. Trace and history plots of estimated parameters (sigmabeta: Level-2 error standard deviation for phase; betac: Average treatment effect for phase; sigmada: Level-2 error standard deviation for interaction)

The trace and history plots of the rest of the parameters look similar to Figure 11. The plots from the first two rows were obtained when Model 1 was used to estimate the parameters for the condition where the number of cases equaled 4 and the series length per case was 10 (First row: History plots; second row: Trace plots). The rest of the plots were from when Model 2 was used to estimate the parameters for the same condition (Third row: History plots; Last row: Trace plots).

Kernel Density Plots (Posterior distributions of each parameter)

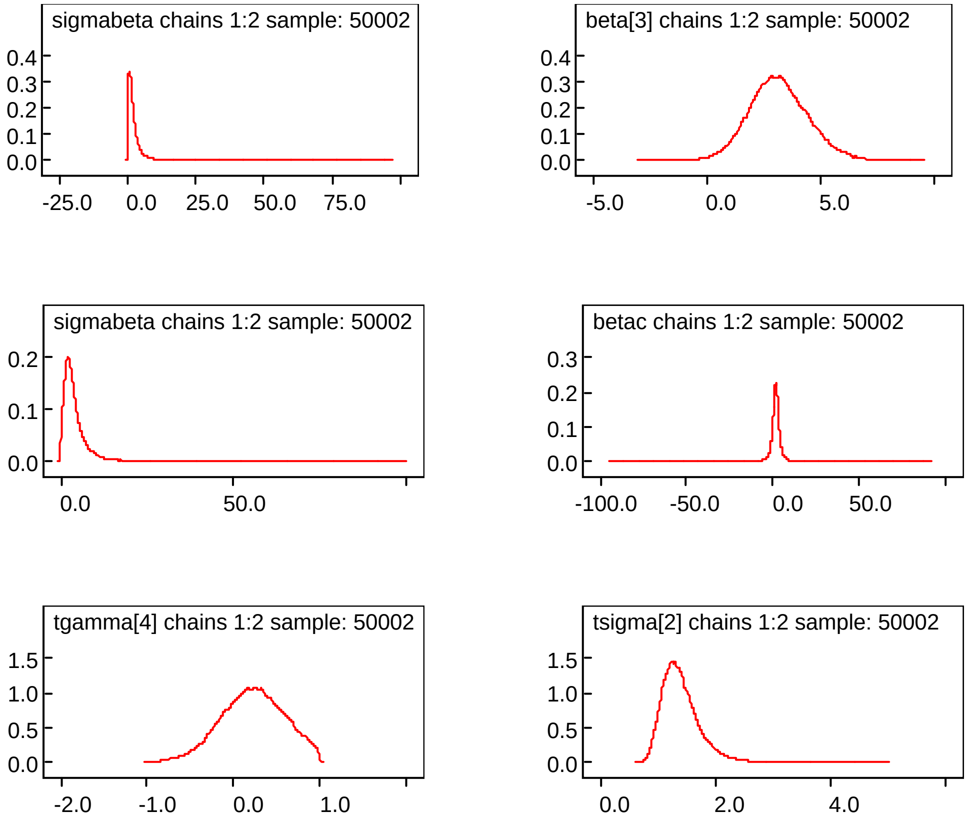


Figure 12. Kernel density plots of estimated parameters (sigmabeta: Level-2 error standard deviation for phase; beta[3] :Individual treatment effect for phase for the case who had id number 3; betac: Average treatment effect for phase; tgamma[4]: autocorrelation for the case who had id number 4; tsigma[2]; level-1 error standard deviation for the case who had id number 2)

Figure 12 shows the Kernel density plots of the level-2 error standard deviation for phase, the individual treatment effect of phase for the case who had the id number 3, the average treatment effect for phase, the autocorrelation for the case who had the id number 4, and the level-1 error standard deviation for the case who had the id number 2. The plots were created from 50,000 samples.

Overall, the convergence looks reasonable in that the distributions are smooth. The density plots of the rest of the parameters all show a smooth shape. The two plots of the first row were from the analysis of Model 1 for the condition where the number of cases was 4 and the series length per case was 10. The rest of the plots were from the analysis of Model 2 for the same condition.

Brooks–Gelman–Rubin (BGR) Plots

Figure 13 shows the BGR plots of the level-2 error standard deviation for phase and the interaction, the average treatment effect for phase, and the autocorrelation for the case who had the id number 3.

Overall, the convergence looks reasonable for most of the parameters since three lines converge to one with stability. The two plots of the first row were from Model 1 and the condition when the number of cases was 4 and the series length per case was 10, whereas the rest of the plots were from Model 2 for the same condition.

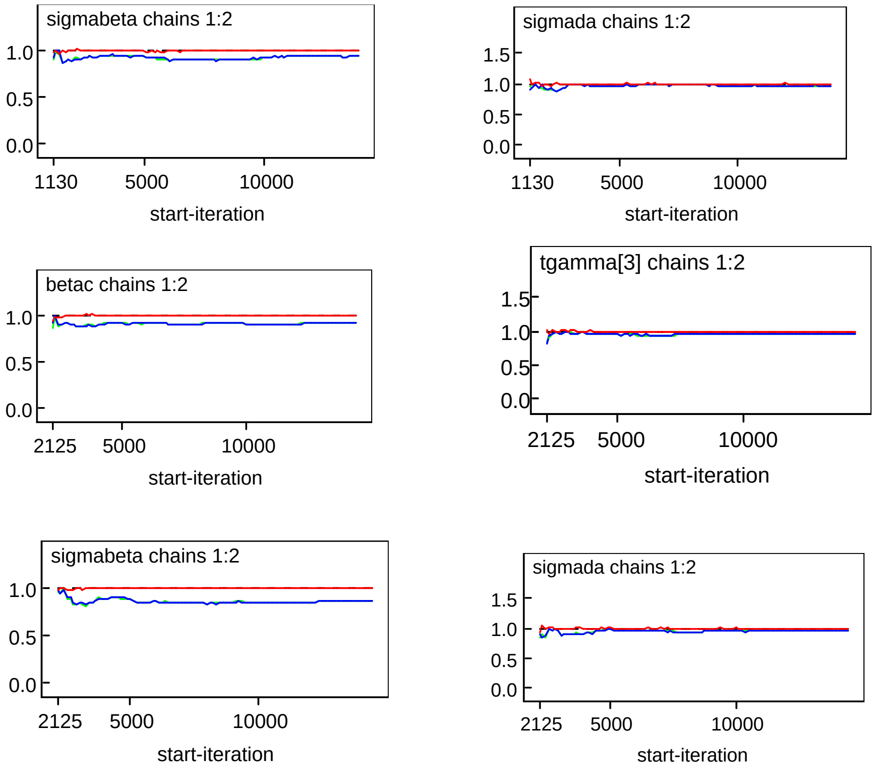


Figure 13. BGR plots of some parameters (sigmabeta: Level-2 error standard deviation for phase; sigmada: Level-2 error standard deviation for interaction; betac: Average treatment effect for phase; tgamma[3]: autocorrelation for the case who had id number 3)

MC Error

The MC error was also tracked for all parameters to check the computational accuracy of the posterior estimates. For example, in the condition of the number of cases equal 4 and the series length per case equal 10 estimated by Model 1, the MC error of the all parameters ranged from .001 (level-1 error standard deviation) to .02 (level-2 error standard deviation for phase). For the same condition estimated by Model 2, the MC error of the all parameters ranged from .002 (level-1 error standard deviation for the case who had the id number 2) to .03 (level-2 error standard deviation for phase).

As each of the 48 conditions were run with the selected number of burn-in and iterations (2,000 burn-in and 500,000 more iterations), the convergence rate that indicated a complete analysis of each condition (1000 samples per each condition) was also tracked for each condition. In the WinBUGS software, several types of trap messages can be popped up during a running analysis which indicates an error that cannot be solved by WinBUGS, as a result, the running analysis cannot be completed. In the analyses of the current study, the 'undefined real result' trap message was obtained occasionally throughout the analysis of each condition. This message indicates numerical overflow which can be caused by several reasons. One possible reason is that the initial values generated may be numerically too extreme, especially when 'noninformative (vague)' priors are used. Another possible reason is that the analysis faces on the numerical difficulties in sampling. For more information about trap messages, please refer to WinBUGS user manual version 1.4 (Spiegelhalter, Thomas, Best, & Lunn, 2003).

The trap message does not explicitly provide a reason that the 'undefined real result' error occurred, therefore, it was assumed the combinations of these possible reasons caused this error in the current study. The error had occurred in every condition. Since the analysis of the targeted sample could not be completed when this error occurred, not all of the 1000 samples were analyzed. Therefore, the total number of samples that were analyzed were tracked per each condition, which was indicated by the convergence rate. The Convergence rate was over 97% for all 48 conditions.

Fixed Treatment Effects

The first research question involves the estimates of the fixed treatment effects and the consequences of modeling and not modeling between case variation in the level-1 error structure.

More specifically, (1) the bias and the RMSE for the fixed treatment effects as function of the design and data factors, and (2) the credible interval coverage and width for the fixed treatment effects as function of the design and data factors.

Bias

The distribution of bias values of the fixed effect for treatment effects (shift in level and shift in slope) are illustrated in Figures 14 through 21. Relative bias values for the treatment effect are provided in Appendix A. The full information about the η^2values for the GLM models is also provided in Appendix B.

Average treatment effect for phase (shift in level). The average bias values of the treatment effect for phase were close to 0 across the two models (Model 1 and Model 2) with little variation (Figure 14). The type of model explained little of the variability ($\eta^2 = .00078$), but the average bias value for Model 2 ($M = 0.0003$, $SD = 0.024$) where between case variation was modeled in the level-1 error structure was slightly smaller than the average bias value for Model 1($M = 0.0016$, $SD = 0.024$) where between case variation was not modeled in the level-1 error structure.

The average bias values of the treatment effect for phase across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 15). The average bias values were all close to 0 across the two models within the three true level-1 error structures with little variability. The different types of the true level-1 error structures explained little of the variability ($\eta^2 = .01196$) which indicates similarity of the average bias values across the three true level-1error structures. Specifically, the smallest average bias value was found when the true level-1 error structure was moderately heterogeneous and estimated by Model 2 (*M*= .0003, *SD* = .028), and the largest average bias value was found when the true level-1 error structure was homogeneous and estimated by Model 1 (*M*= .0050, *SD* = .014).

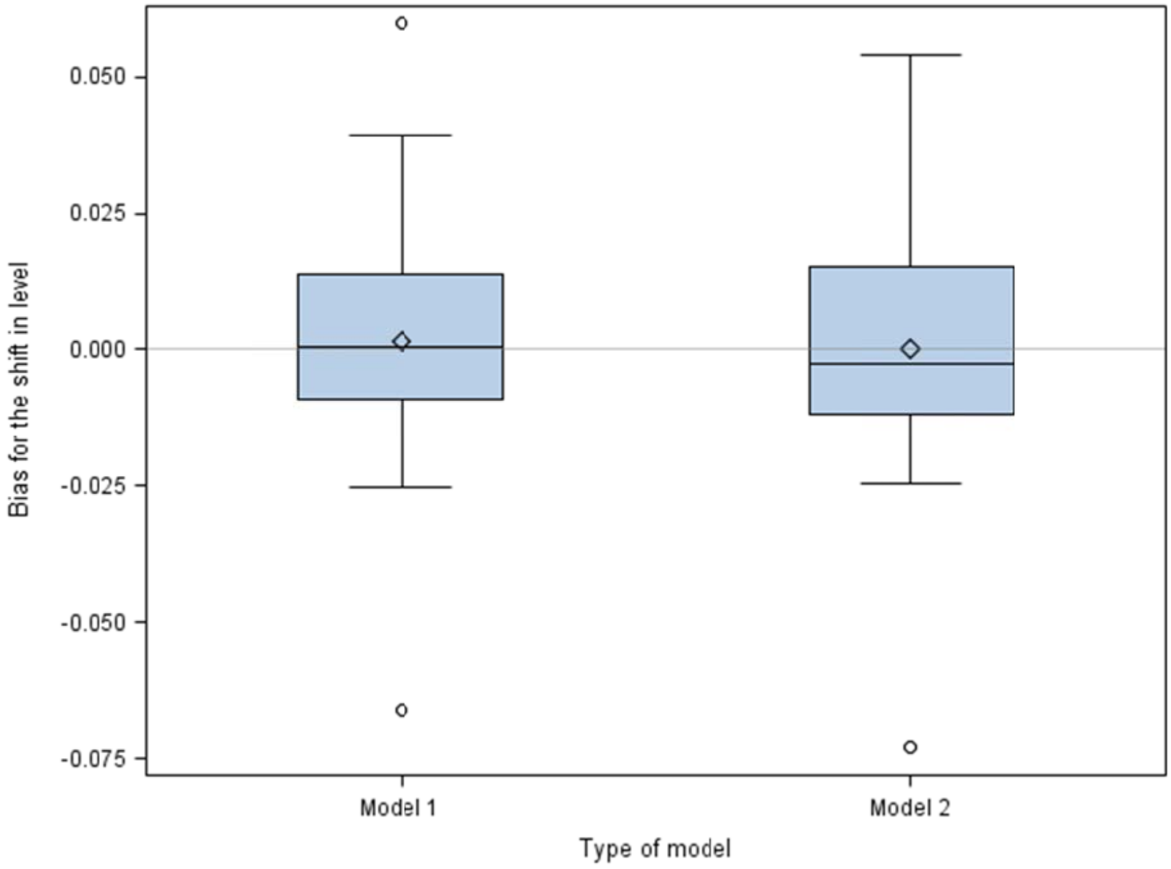


Figure 14. Box plots illustrating the distribution for the average bias values for the shift in level across Model 1 which did not model between case variation, and Model 2 which models between case variation.

In addition, there was very little difference across the two models within the three different types of the true level-1 error structures. The smallest average bias difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2-M_1|$ = 0.0007), and the biggest average bias difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2-M_1|$ = 0.0016).

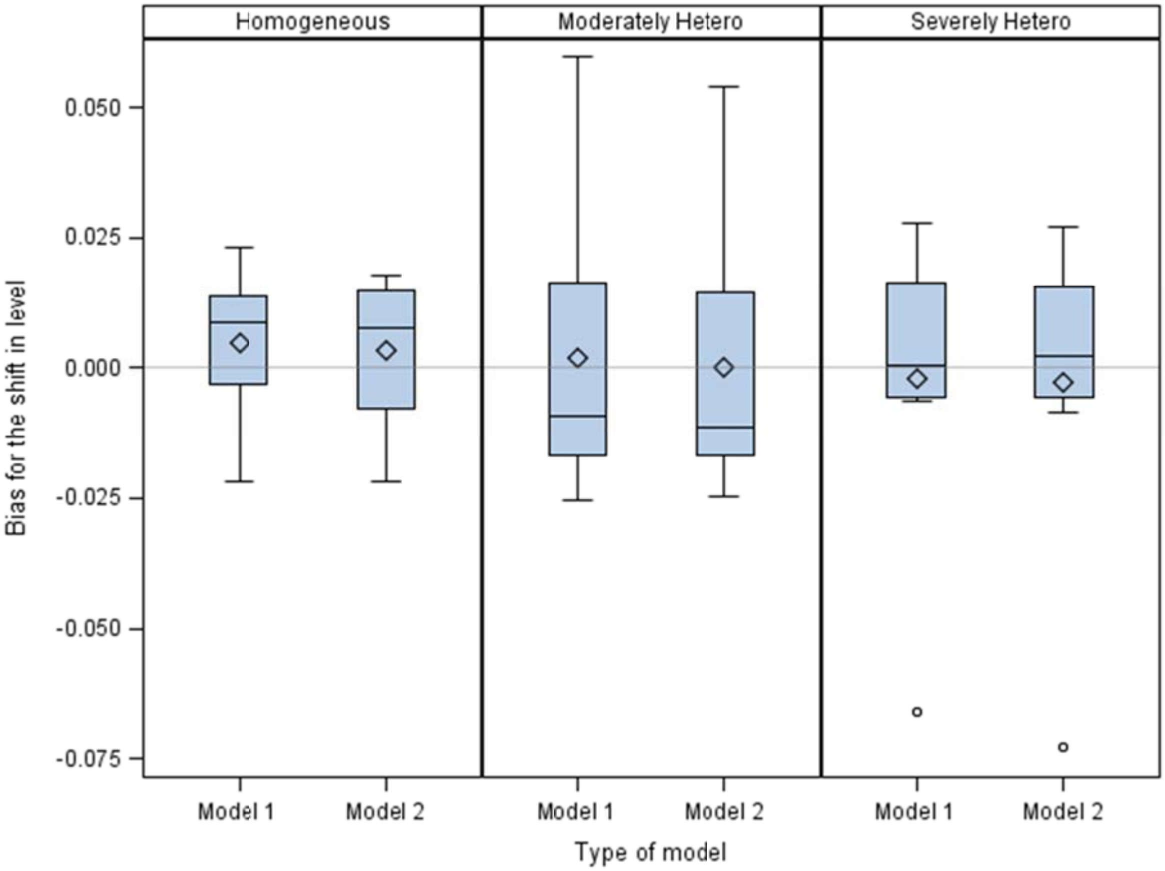


Figure 15. Box plots illustrating the distribution for the bias values for the shift in level across the two models within the three true level-1 error structures.

In order to further explore if any design factor had a significant effect on the bias for the shift in level, GLM models were run. The model explained 99% of variability after including 4-way interactions, and indicated the 4-way interaction among the number of cases, the series length per case, the true level-1 error structure, and the variation in the level-2 errors had a medium effect ($\eta^2 = .10$). The relationship for the average bias for the shift in level as a function of the number of cases, the series length per case, the true level-1 error structure, and variation in the level-2 errors is illustrated with a line graph in Figure 16.

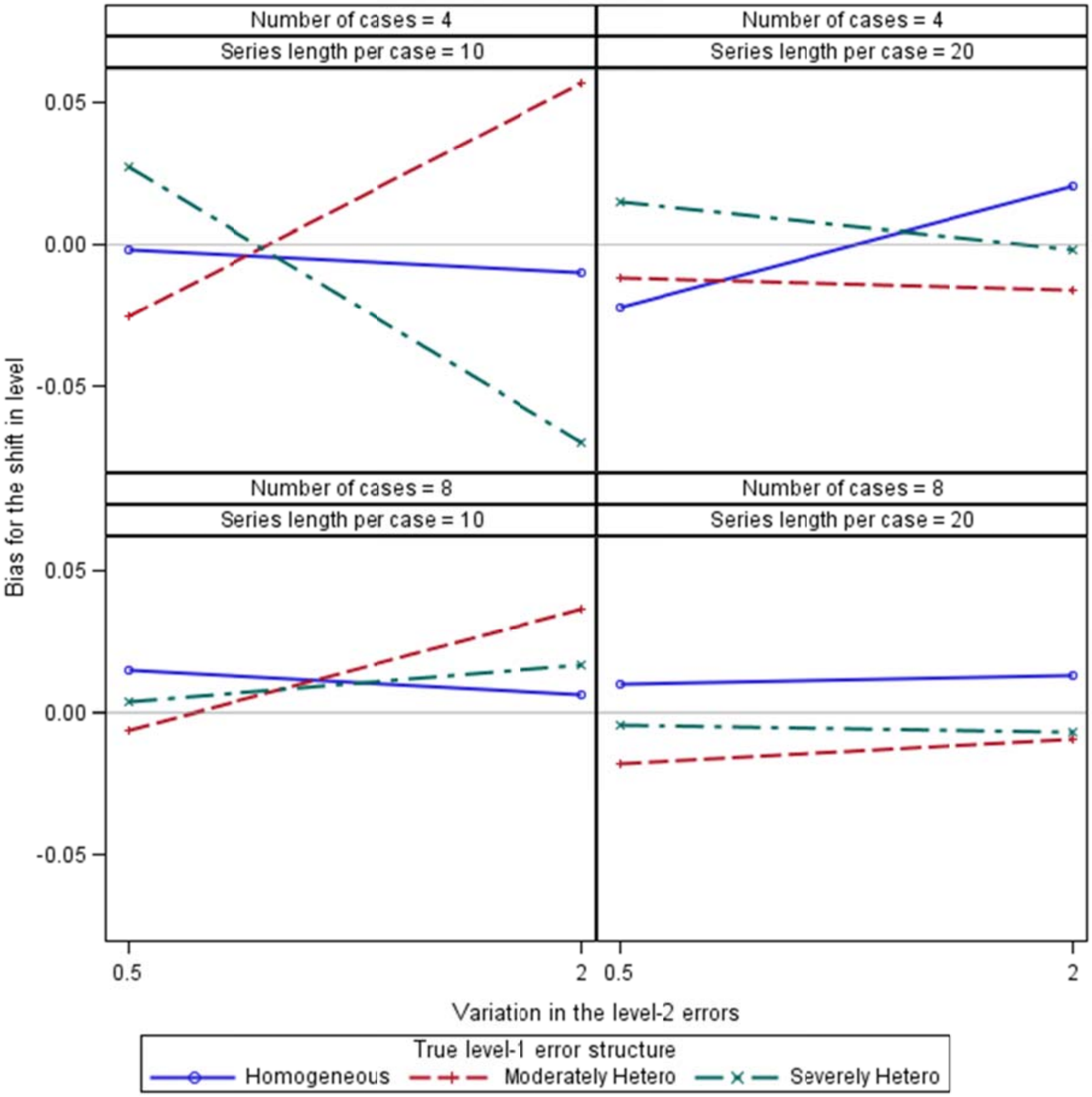


Figure 16. Line graphs illustrating the relationship of the bias in the shift in level and the four-way interaction among the number of cases, the series length per case, the variation in the level-2 errors, and the true level-1 error structure.

The line graph shows that there was some variability of the average bias values across the true level-1 error structures. When the true level-1 error structure was homogeneous, the average bias values were relatively similar across the number of cases, the series length per case, and the variation in the level-2 errors. The smallest bias value was found when the number of cases was

4, the series length per case was 10, and the variation in the level-2 errors was such that most variance was at level-1 which was a bias of .5 (*M*= -0.001, *SD* = 0.004). The largest bias value was found when the number of case was 4, the series length per case was 20, and the variation in the level-2 errors was such that most variance was at level-1 which was a bias of .5 (*M*= -0.022, *SD* < 0.001). However, when the true level-1 error structure was the moderately or the severely heterogeneous, the average bias values were impacted by the level of factors. Specifically, the average bias values were varied the most across the variation in the level-2 errors when the number of cases and the series length per case were small which was 4 and 10 respectively. When the variation in the level-2 errors shifted from most variance at level-1 (0.5) to most variance at level-2 (2), the average bias values increased for both the moderately or the severely heterogeneous error structure (from *M*= -0.025, *SD* = 0.001; *M*= 0.028, *SD* = 0.001, respectively to *M*= 0.057, *SD* = 0.004; *M*= -0.069, *SD* = 0.005, respectively).

Average treatment effect for interaction (shift in slope). The average bias values for the treatment effect for interaction were very similar and close to 0 across the two models (Model 1 and Model 2) with little variation (η^2 = .00059) (Figure 17). The average bias value for Model 1 was *M* = 0.0035, *SD* = 0.008, and the average bias value for Model 2 was *M* = 0.0031, *SD* = 0.008.

The average bias values for the treatment effect for interaction across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 18). The average bias values were very similar and close to 0 across the two models within the three true level-1 error structures.

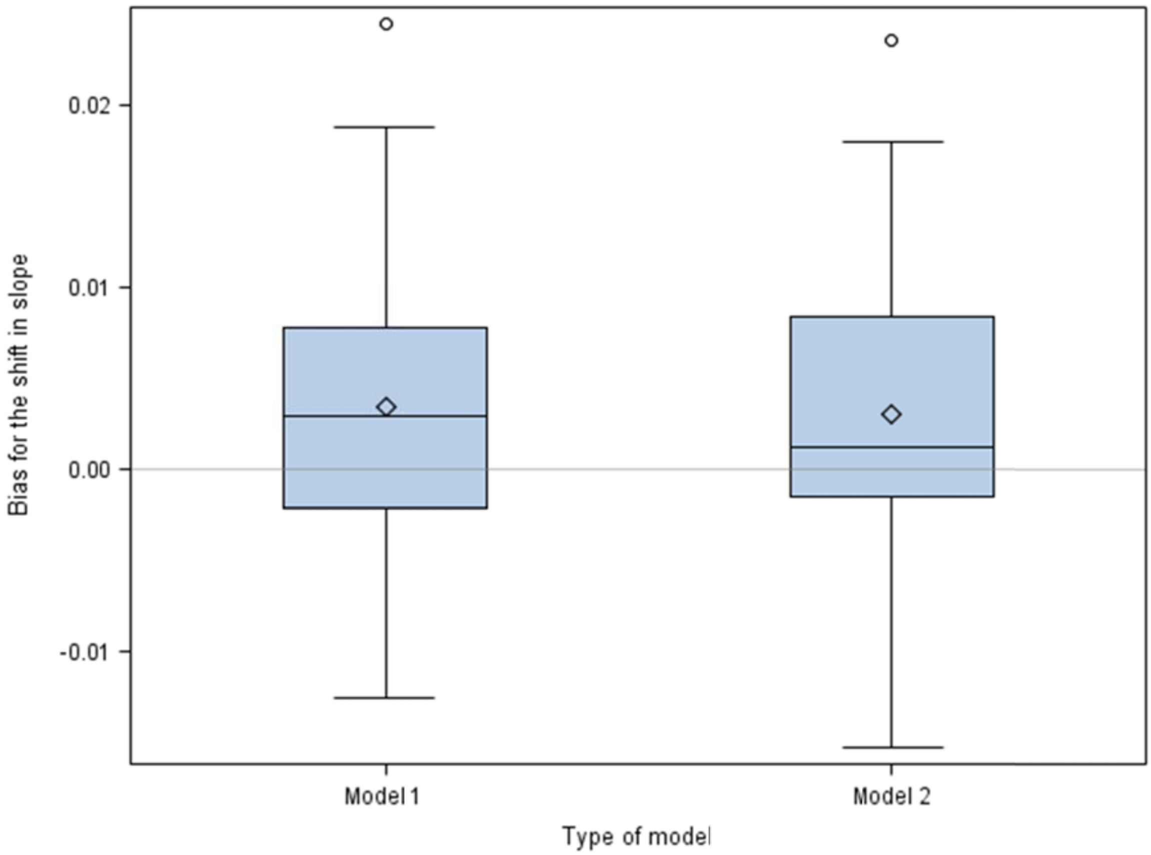


Figure 17. Box plots illustrating the distribution for the average bias values for the shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation .

The smallest average bias difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1|$ = 0.0001), and the biggest average bias difference between the two models was found when the true level-**1** error structure was severely heterogeneous ($|M_2\text{-}M_1|$ = 0.001).

However, some variability of the average bias values was found across the true level-1 error structures. The different types of the true level-1 error structures explained a large amount of the variability (η^2 = .25) which indicates substantial differences of the average bias across the true level-1error structures. Specifically, the smallest average bias value was found when the true level-1 error structure was severely heterogeneous (Model 1: *M*= 0.0002, *SD* = 0.007; Model 2: *M*= -0.0007, *SD* = 0.008), and the largest average bias was found when the true level-1 error structure was homogeneous (Model 1: *M*= 0.0088, *SD* = 0.009; Model 2: *M*= 0.0089, *SD* = 0.009).

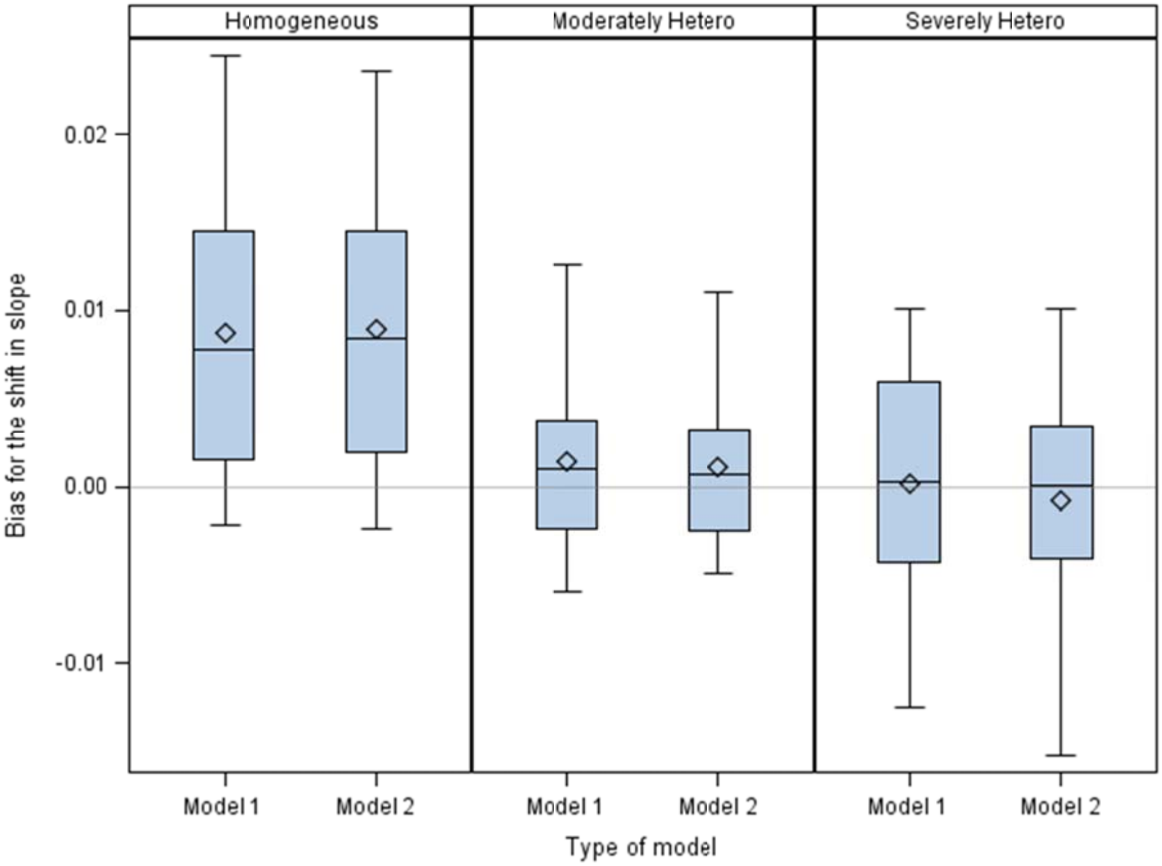


Figure 18. Box plots illustrating the distribution for the bias values for the shift in slope across the two models within the three true level-1 error structures.

In order to further explore if any design factor had a significant effect on the bias for the shift in slope, GLM models were run. The model explained 95% of the variability after including 3-way interactions. The GLM model found three interaction effects that had a medium effect, including the 3-way interaction among the number of cases, the series length per case, and the true level-1 error structure (η^2 = .10), the 3-way interaction among the number of cases, the series length per case, and the variation in the level-2 errors ($\eta^2 = .09$), and the 3-way interaction among the series length per case, the true level-1 error structure, and the variation in the level-2 errors ($\eta^2 = .08$). These three interaction effects were illustrated using a line graphs in Figure 19, 20, and 21.

The relationship for the average bias for the shift in slope as function of the number of cases, the series length per case, and the true level-1 error structure is illustrated with line graph in Figure 19. The graph shows that there was some variability of the average bias values across the true level-1 error structures. When the true level-1 error structure was moderately or severely heterogeneous error structure, the average bias value was decreased (close to 0) as the number of cases increased from 4 to 8, regardless of the series length per case. Specifically, when the number of cases increased from 4 to 8, the average bias value in the moderately heterogeneous error structure decreased from *M*= 0.0074, *SD* = 0.005 to *M*= -0.0022, *SD* = 0.004 for the series length per case of 10, and from *M*= 0.0001, *SD* = 0.005 to *M*= -0.0001, *SD* = 0.001 for the series length per case of 20. The average bias value in the severely heterogeneous error structure decreased from *M*= -0.0052, *SD* = 0.010 to *M*= 0.0005, *SD* = 0.002 for the series length per case of 10, and from *M*= 0.0022, *SD* = 0.009 to *M*= 0.0013, *SD* = 0.005 for the series length per case of 20. However, when the true level-1 error structure was homogeneous, the average bias values were positively biased, and relatively higher than when the true level-1 error structure was moderately or severely heterogeneous error structure. In addition, the difference of the average bias across the number of cases was changed depending on the series length per case. Specifically, when the series length per case was 10, the average bias value was increased from *M*= 0.0091, *SD* = 0.011 to *M*= 0.0162, *SD* = 0.009 as the number of cases was increased from 4 to 8. When the series length per case was 20, the average bias value was decreased from *M*= 0.0093, *SD* = 0.002 to *M*= 0.0008, *SD* = 0.004 as the number of cases was increased from 4 to 8.

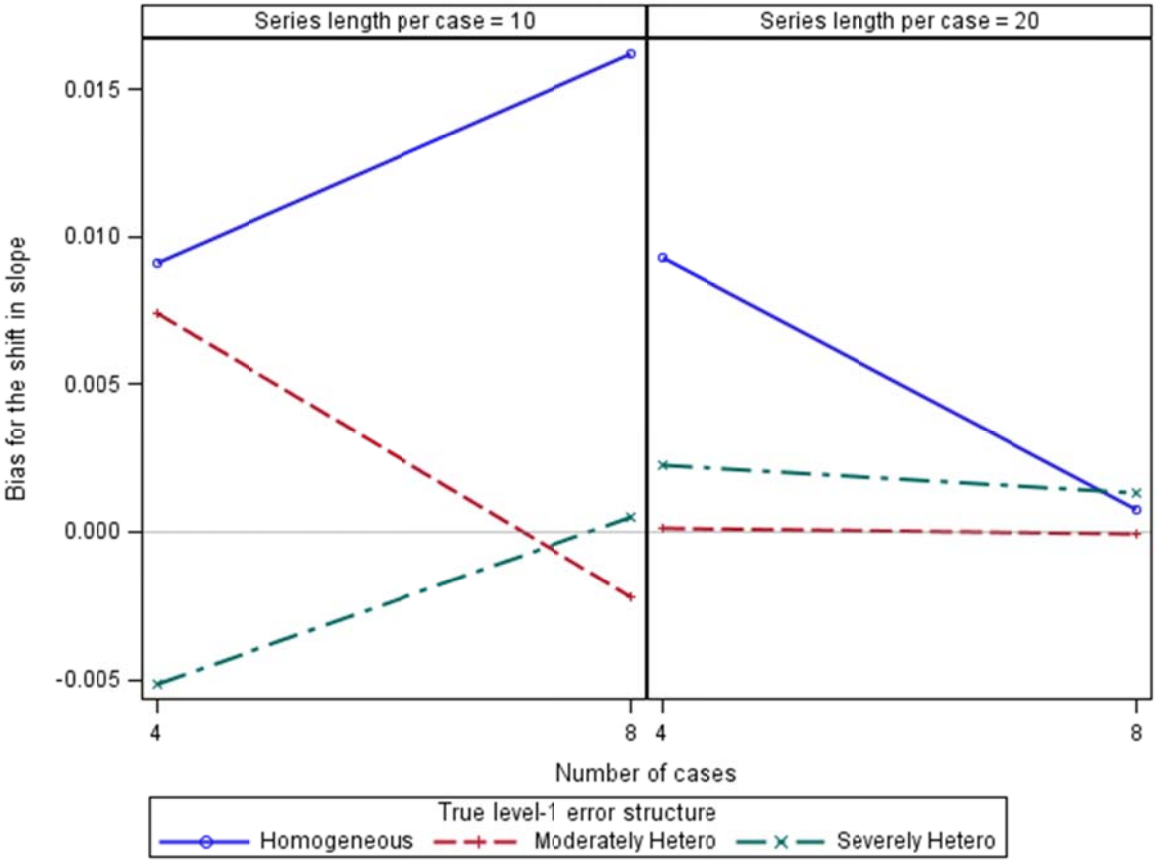


Figure 19. Line graph depicting average bias for the shift in slope as a function of the three-way interaction effect between the number of cases, the series length per case, and the true level-1 error structure.

The relationship for the average bias for the shift in slope as function of the number of cases, the series length per case, and the variation in the level-2 errors is illustrated with a line graph in Figure 20. The graph shows that when the number of cases was 8, the average bias values were increased for both series length per case 10 and 20, regardless of the variation in the level-2 errors. Specifically, the average bias value was increased from *M*= 0.0027, *SD* = 0.004 to

M= 0.0069, *SD* = 0.014 for the series length per case of 10, and from *M*= -0.0020, *SD* = 0.001 to *M*= 0.0033, *SD* = 0.002 for the series length per case of 20.

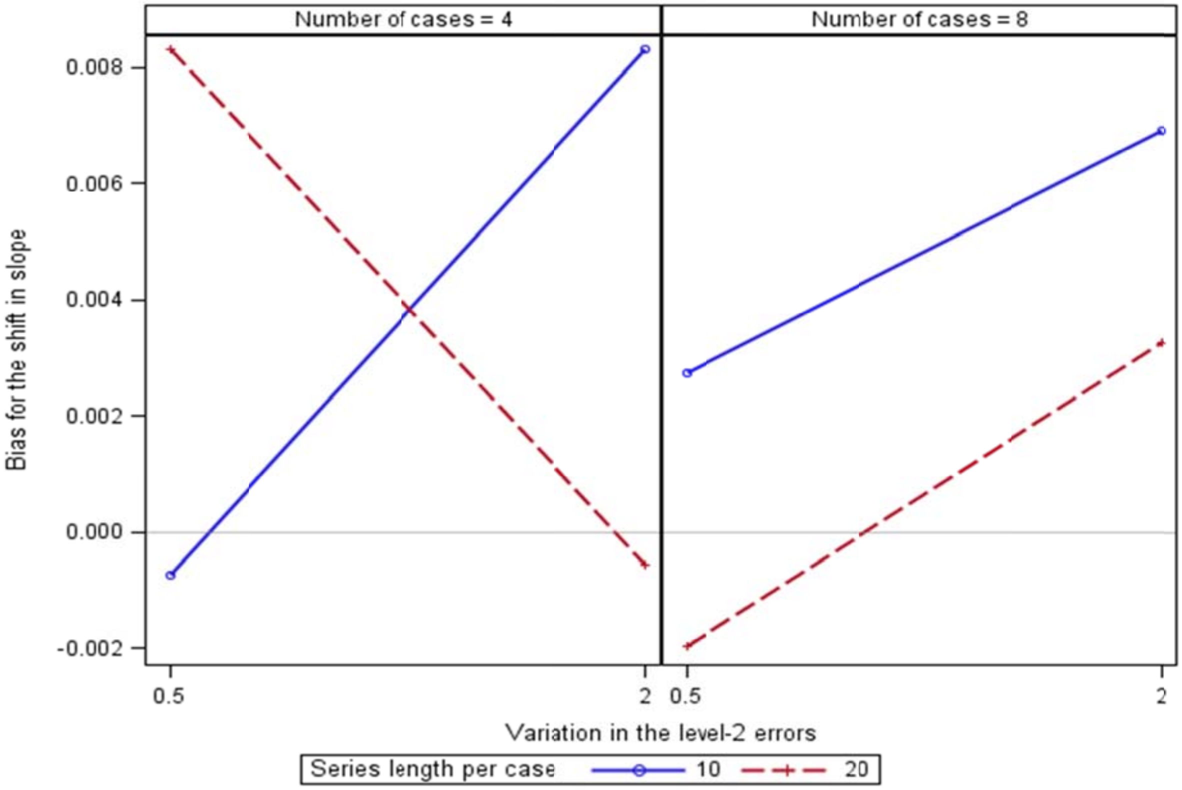


Figure 20. Line graph depicting average bias for the shift in slope as a function of the three-way interaction effect among the number of cases, the series length per case, and the variation in the level-2 errors.

However, when the number of cases was 4, the average bias values with the series length per case were dependent on the variation in the level-2 errors. Specifically, when the variation in the level-2 errors was such that most variance was at level-1 (0.5), the average bias value with the series length per case of 10 was smaller and negatively biased than the average bias value with the series length per case of 20 which was relatively high and positively biased, *M*= -0.0007, *SD* = 0.012 and *M*= 0.0083, *SD* = 0.003 for the series length per case of 10 and 20, respectively. When the variation in the level-2 errors was such that most variance was at level-2

(2), the average bias value with the series length per case of 10 was relatively larger and positively biased (*M*= 0.0083, *SD* = 0.008) than the average bias value with the series length per case of 20 which was relatively small and negatively biased (*M*= -0.0006, *SD* = 0.007).

The relationship for the average bias for the shift in slope as a function of the variation in the level-2 errors, the series length per case, and the true level-1 error structure is illustrated with a line graph in Figure 21. The graph shows that there was some variability of the average bias values across the true level-1 error structures. However, the pattern of the variability across the true level-1 error structures was changed depending on the series length per case, and the variation in the level-2 errors. When the series length per case was 20, the average bias values were changed relatively little across the variation in the level-2 errors for all three true level-1 error structures. The average bias value was changed from *M*= 0.0042, *SD* = 0.008 to *M*= 0.0058, *SD* = 0.002 for the homogeneous error structure, from *M*= 0.0016, *SD* = 0.003 to *M*= -0.0016, *SD* = 0.003 for the moderately heterogeneous error structure, and from *M*= 0.0037, *SD* = 0.007 to *M*= -0.0002, *SD* = 0.006 for the severely heterogeneous error structure. However, when the series length per case was 10, the average bias values changed more, either decreased or increased, across the variation in the level-2 errors for all three true level-1 error structures. The average bias value was decreased from *M*= 0.006, *SD* = 0.006 to *M*= -0.001, *SD* = 0.005, and from *M*= -0.007, *SD* = 0.007 to *M*= 0.003, *SD* = 0.003 for the moderately and the severely heterogeneous error structure, respectively. The average bias value was increased from *M*= 0.004, *SD* = 0.005 to *M*= 0.021, *SD* = 0.003 for the homogeneous error structure.

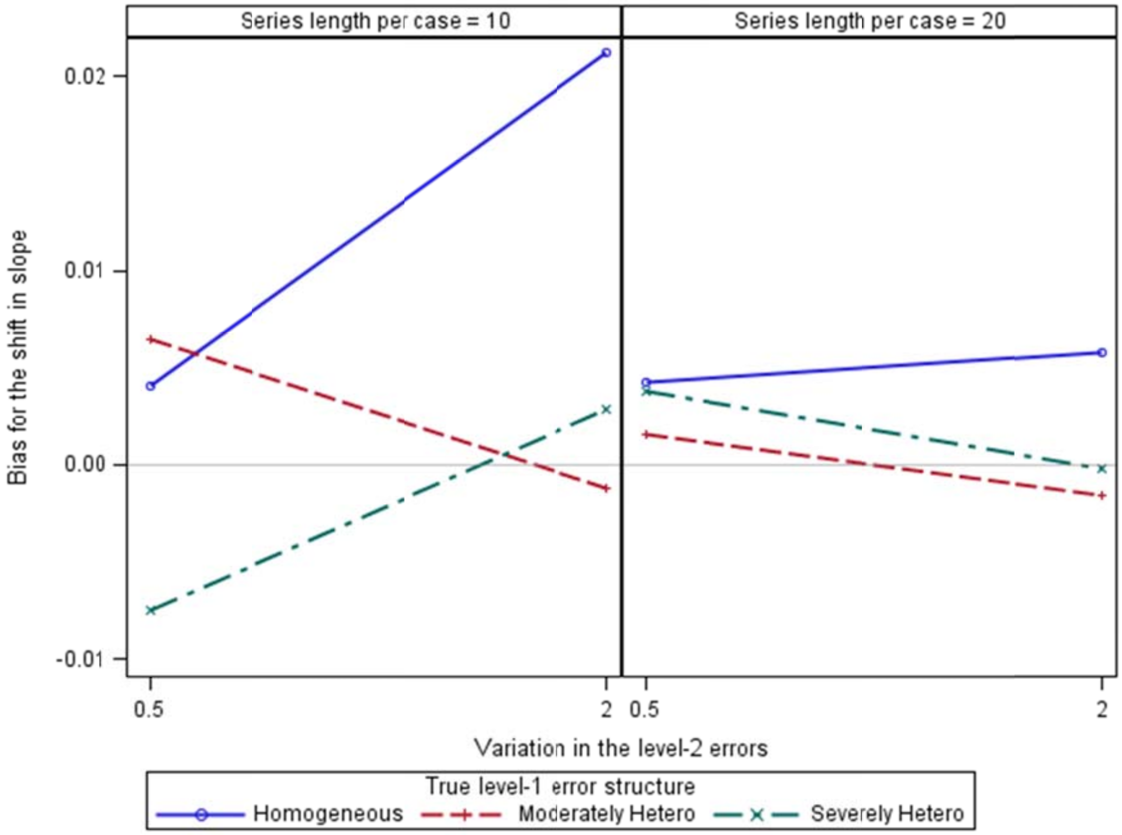


Figure 21. Line graph depicting average bias for the shift in slope as a function of the three-way interaction effect among the variation in the level-2 errors, the series length per case, and the true level-1 error structure.

Root Mean Squared Error (RMSE)

The distribution of RMSE values of the fixed effect for treatment effects (shift in level and shift in slope) are illustrated in Figures 22 through 31. The full information about the η^2 values for the GLM models is also provided in Appendix B.

Average treatment effect for phase (shift in level). The average RMSE values of the treatment effect for phase were very similar across the two models (Model 1 and Model 2) with little variation (Figure 22). The average RMSE value for Model 1 was *M* = 0.68, *SD* = 0.19, and the average RMSE value for Model 2 was *M* = 0.68, *SD* = 0.28. The type of model explained little of the variability (η^2 = .00025).

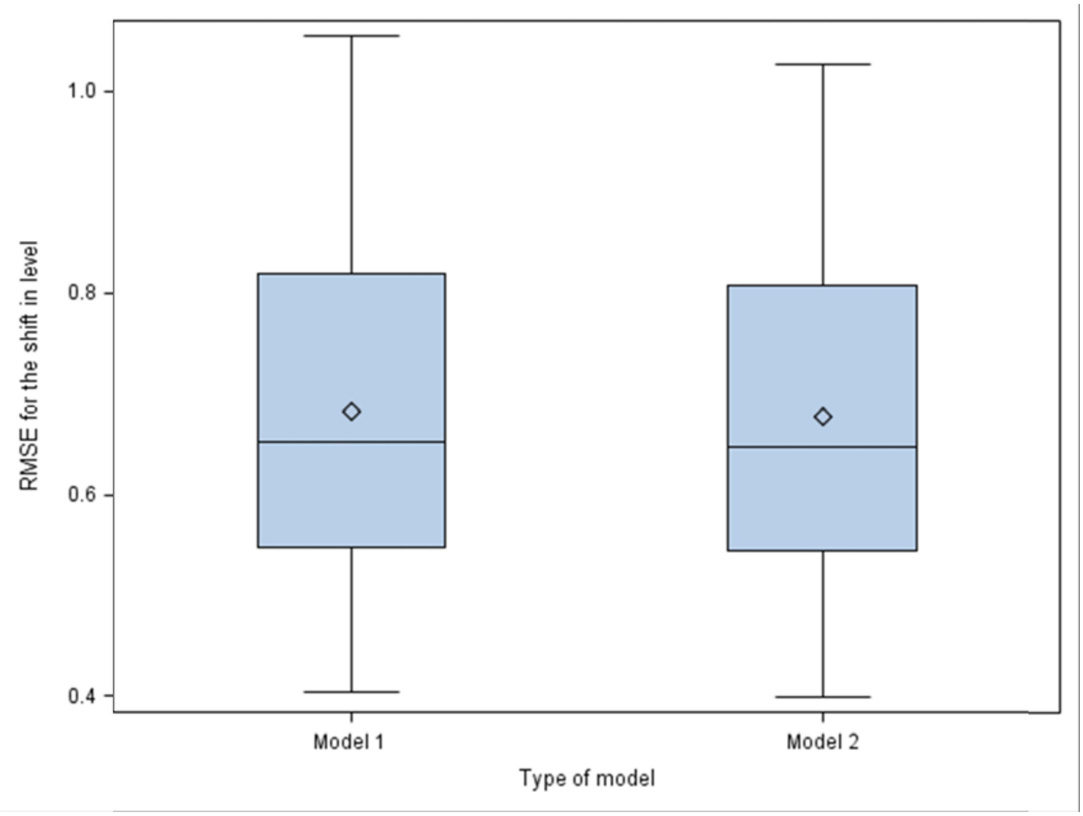


Figure 22. Box plots illustrating the distribution for the RMSE values for the shift in level across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average RMSE values of the treatment effect for phase across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 23). The average RMSE values were very similar across the two models within the three true level-1 error structures, and there were no or little differences across the three true level-1 error structures (η^2 = .001). The smallest average RMSE difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2$-$M_1|$ = 0.001), and the biggest average RMSE difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2$-$M_1|$ = 0.018).

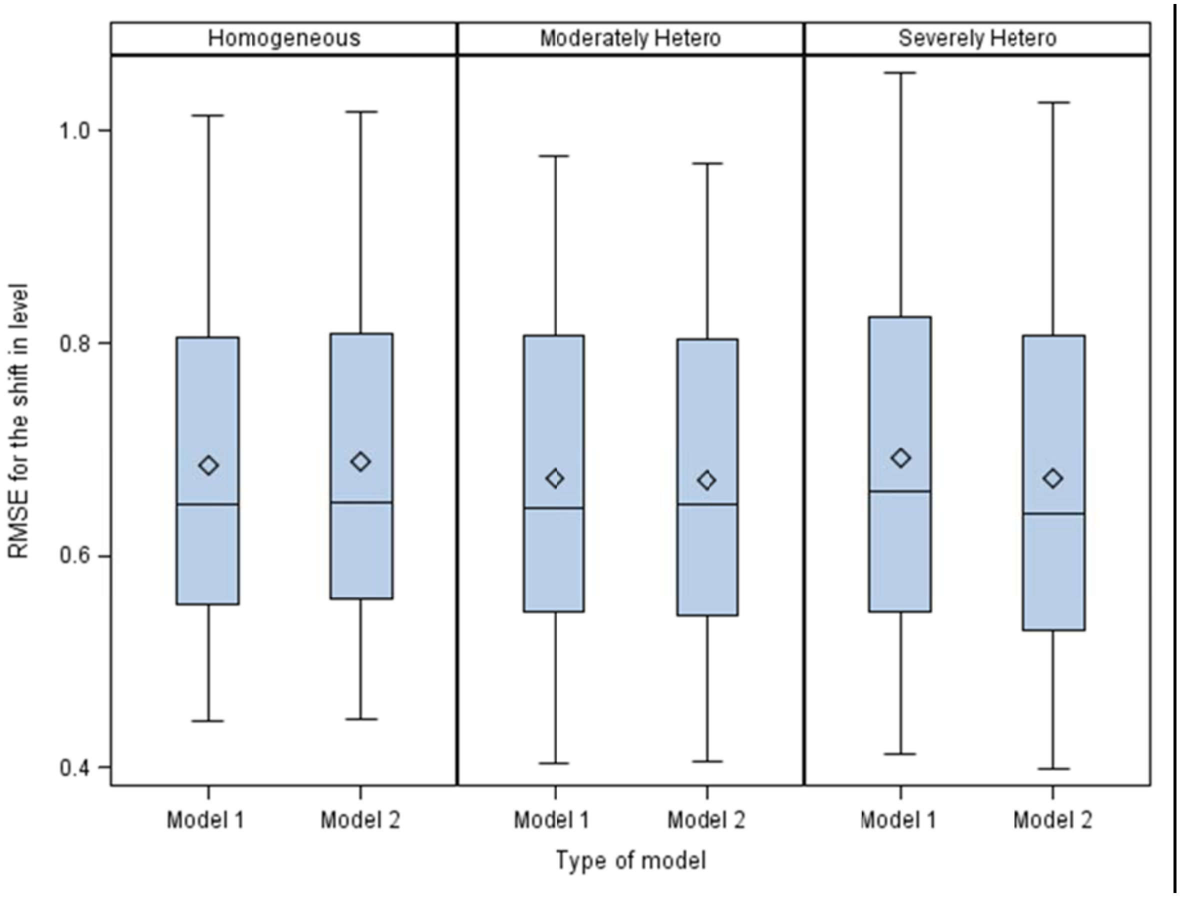


Figure 23. Box plots illustrating the distribution for the RMSE values for the shift in level across the two models within the three true level-1 error structures.

In order to further explore if any design factor had a significant effect on the RMSE for the shift in level, a GLM model was run. The main effects only model explained 97% of the variability and found that three of the design factors had a medium or large effect, including the number of cases (η^2 = .48), variation in the level-2 errors (η^2 = .38), and the series length per case (η^2 = .11). These three main effects are illustrated using a box plots in Figure 24, 25, and 26.

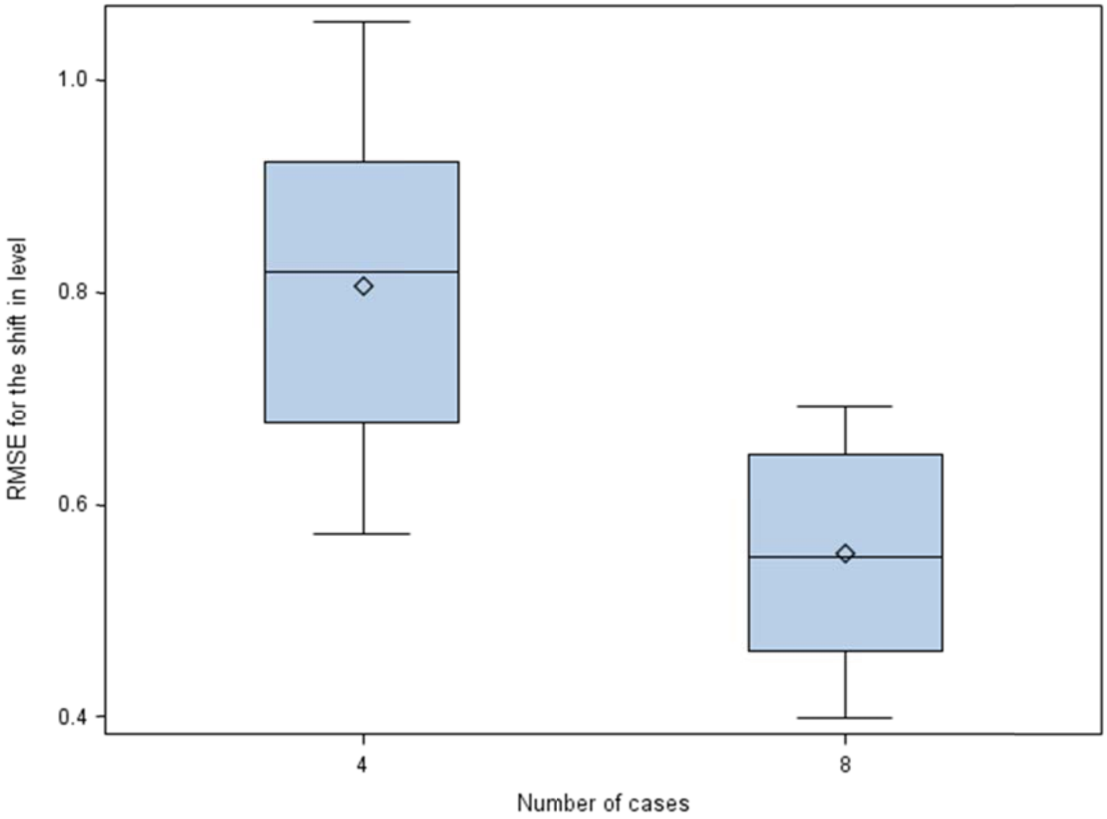


Figure 24. Box plots depicting the estimated RMSE values for the shift in level as a function of the number of cases.

As illustrated in Figure 24, as the number of cases increased from 4 to 8, the average RMSE value decreased from *M* = 0.81, *SD* = .16 to *M* = 0.55, *SD* = .10. Similarly, Figure 25 shows that the average RMSE value was smaller when the variation in the level-2 errors was such that most variance was at level-1 (0.5) (*M* = 0.57, *SD* = .13) than when the variation in the level-2 errors was such that most variance was at level-2 (2) (*M* = 0.79, *SD* = .16). In addition, Figure 26 portrays that as the series length per case increased from 10 to 20, the average RMSE value decreased from *M* = 0.74, *SD* = .19 to *M* = 0.62, *SD* = .16.

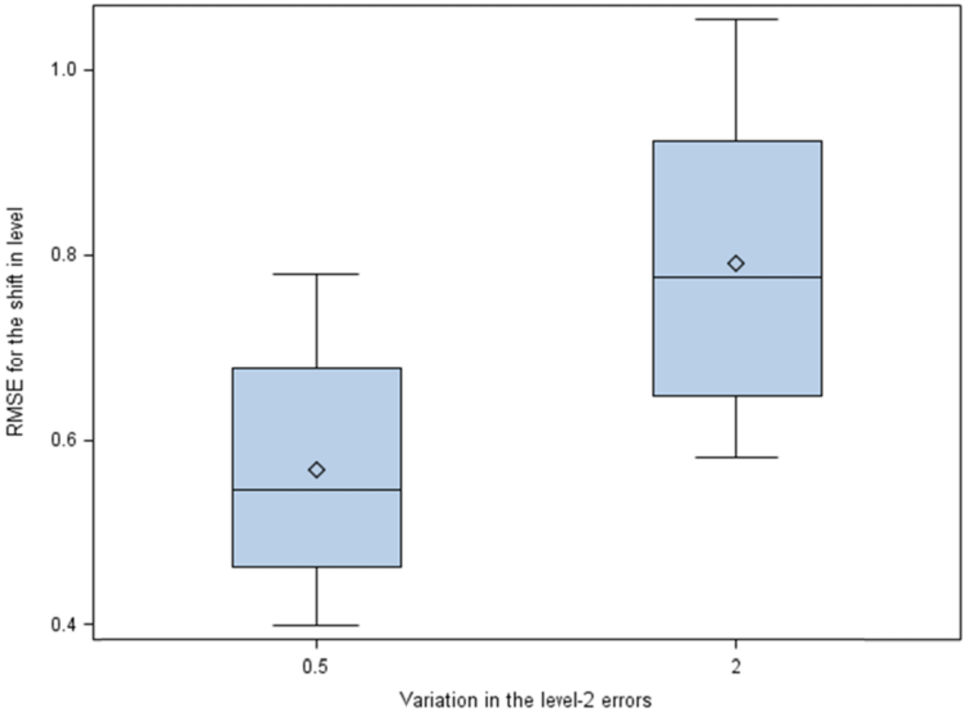


Figure 25. Box plots depicting the estimated RMSE values for the shift in level as a function of the variation in the level-2 errors.

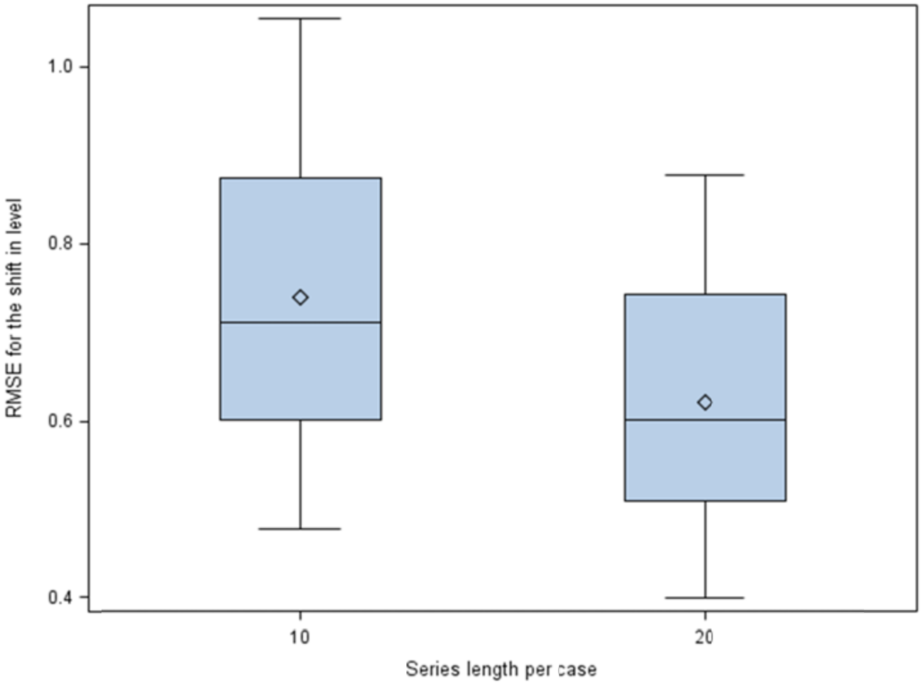


Figure 26. Box plots depicting the estimated RMSE values for the shift in level as a function of the series length per case.

Average treatment effect for interaction (shift in slope). The average RMSE values for the treatment effect for interaction were very similar across the two models (Model 1 and Model 2) with little variation (Figure 27). The average RMSE value for Model 1 was *M* = 0.23, *SD* = 0.08, and the average RMSE value for Model 2 was *M* = 0.22, *SD* = 0.08. The type of model explained little of the variability (η^2 = .0001).

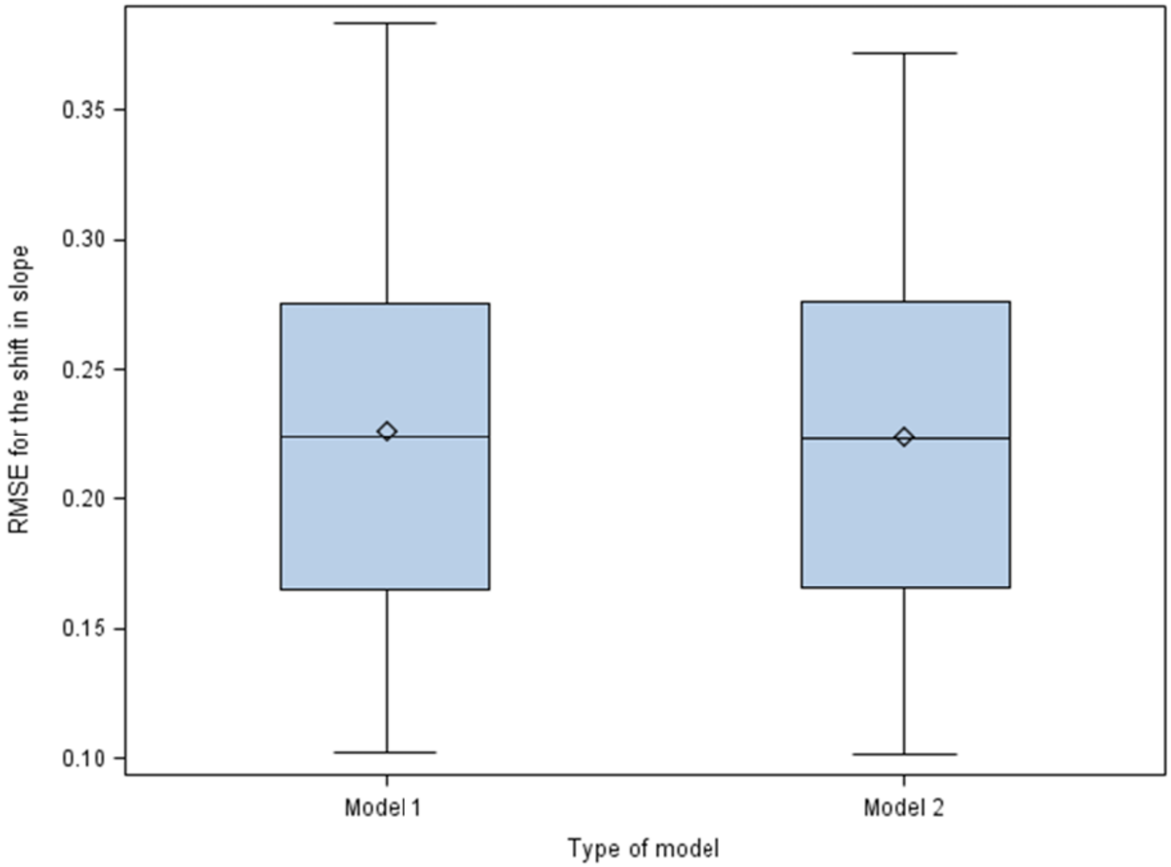


Figure 27. Box plots illustrating the distribution of the RMSE values for the shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average RMSE values of the treatment effect for interaction across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 28). The average

RMSE values were very similar across the two models within the three true level-1 error structures, and there were no or little differences across the three true level-1 error structures ($\eta^2 = .001$). The smallest average RMSE difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2\text{-}M_1| = 0.0004$), and the biggest average RMSE difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1| = 0.0039$).

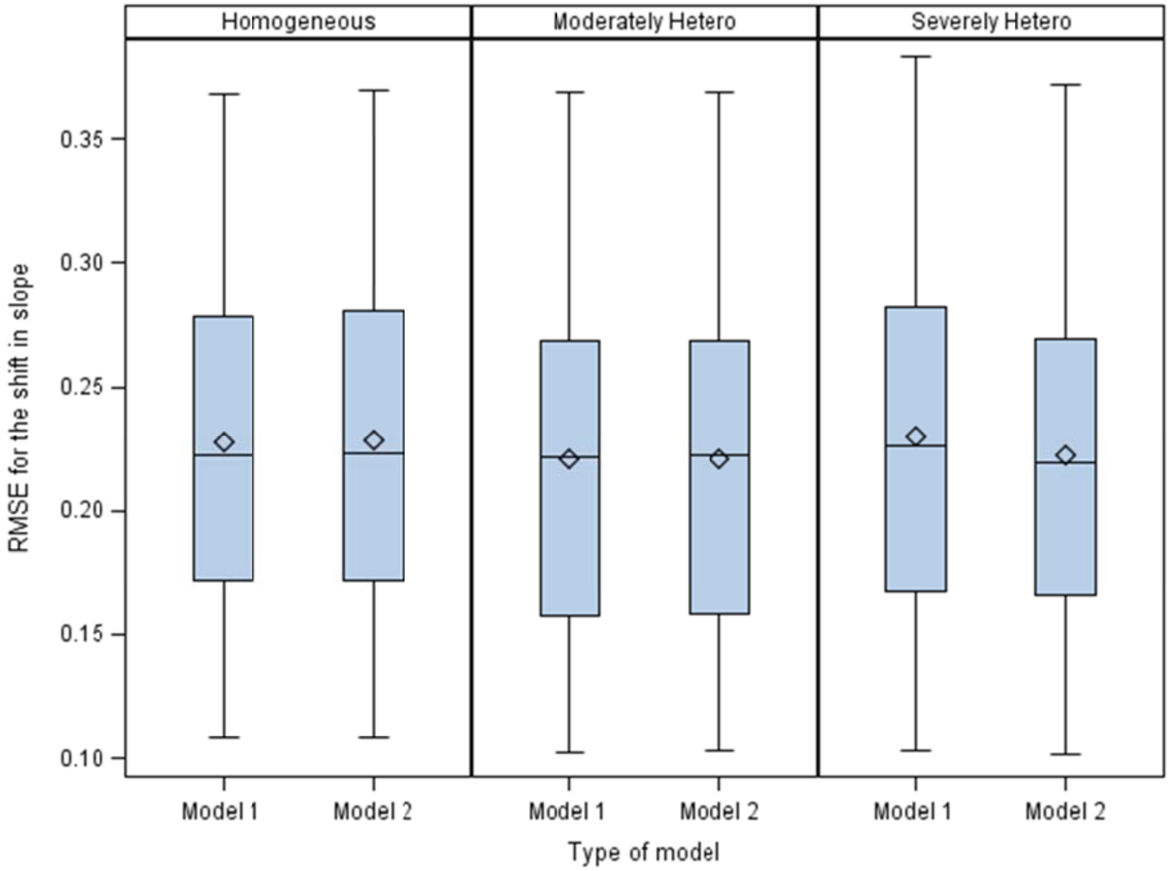


Figure 28. Box plots illustrating the distribution of the RMSE values for the shift in slope across the two models within the three true level-1 error structures.

In order to further explore if any design factor had a significant effect on the RMSE values for shift in slope, a GLM model was run. The main effects only model explained 97% of the variability. Similar to the GLM result of the phase effect, it was found that three of the design factors had a large effect, including the series length per case (η^2 = .47), the number of cases (η^2 = .28), and the variation in the level-2 errors (η^2 = .22). These three main effects are illustrated using a box plots in Figure 29, 30, and 31.

As illustrated in Figure 29, as the series length per case increased from 10 to 20, the average RMSE value decreased from *M* = 0.28, *SD* = .06 to *M* = 0.17, *SD* = .05. Similarly, Figure 30 shows that as the number of cases increased from 4 to 8, the average RMSE value decreased from *M* = 0.27, *SD* = .08 to *M* = 0.18, *SD* = .06.

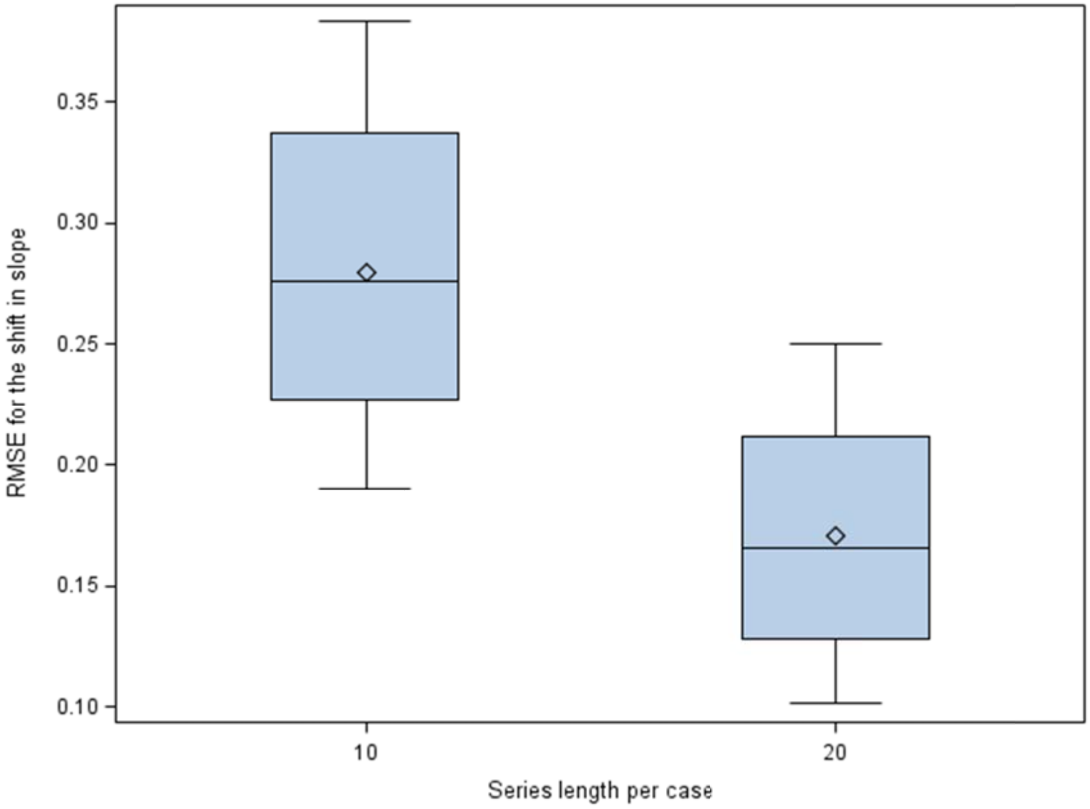


Figure 29. Box plots depicting the estimated RMSE values for the shift in slope as a function of the series length per case.

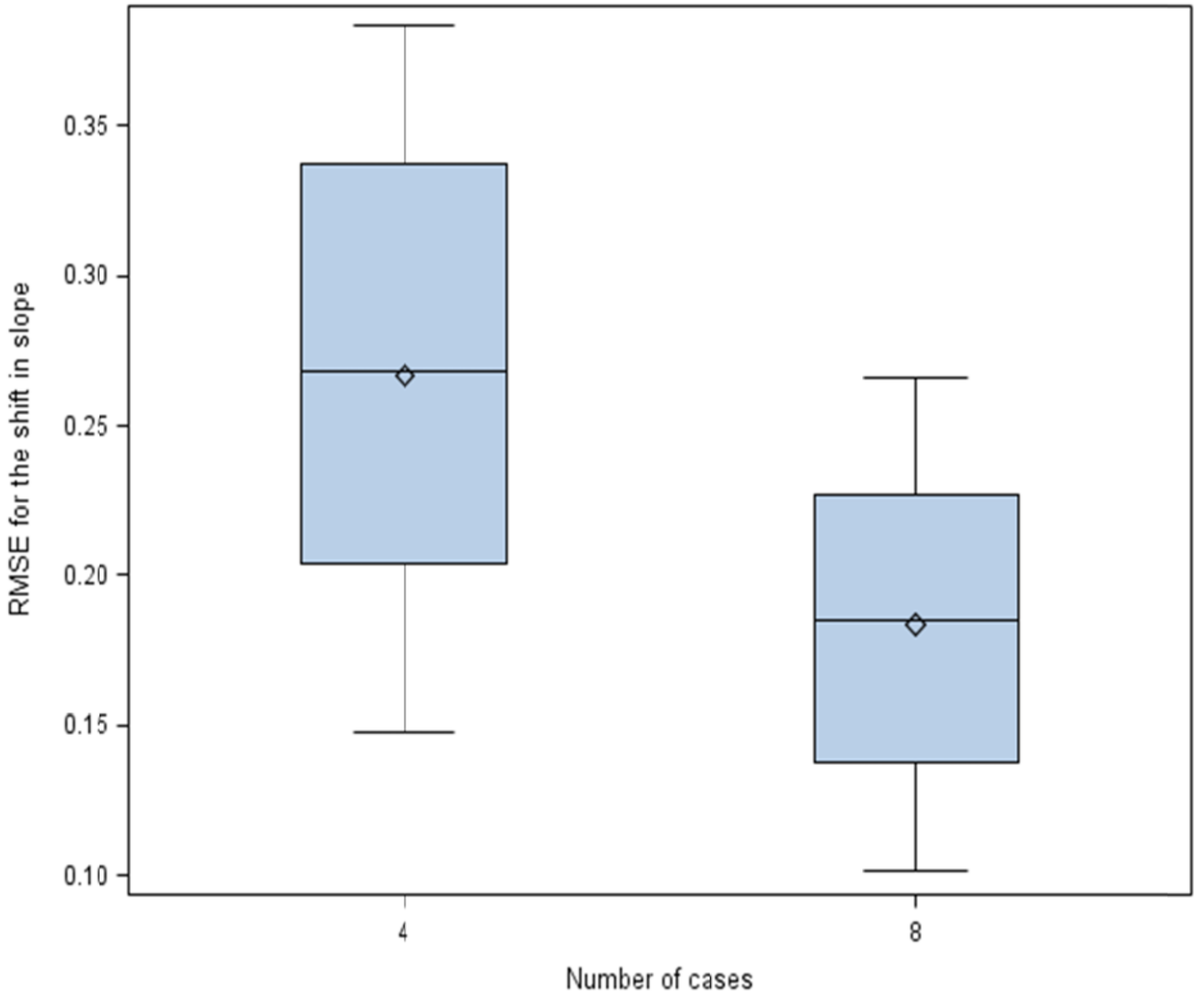


Figure 30. Box plots depicting the estimated RMSE values for the shift in slope as a function of the number of cases.

In addition, Figure 31 portrays that as the variation in the level-2 errors shifted from most of the variance at the level-1 error (.5) to most of the variance at the level-2 error (2), the average RMSE value increased from *M* = 0.19, *SD* = .07 to *M* = 0.26, *SD* = .07.

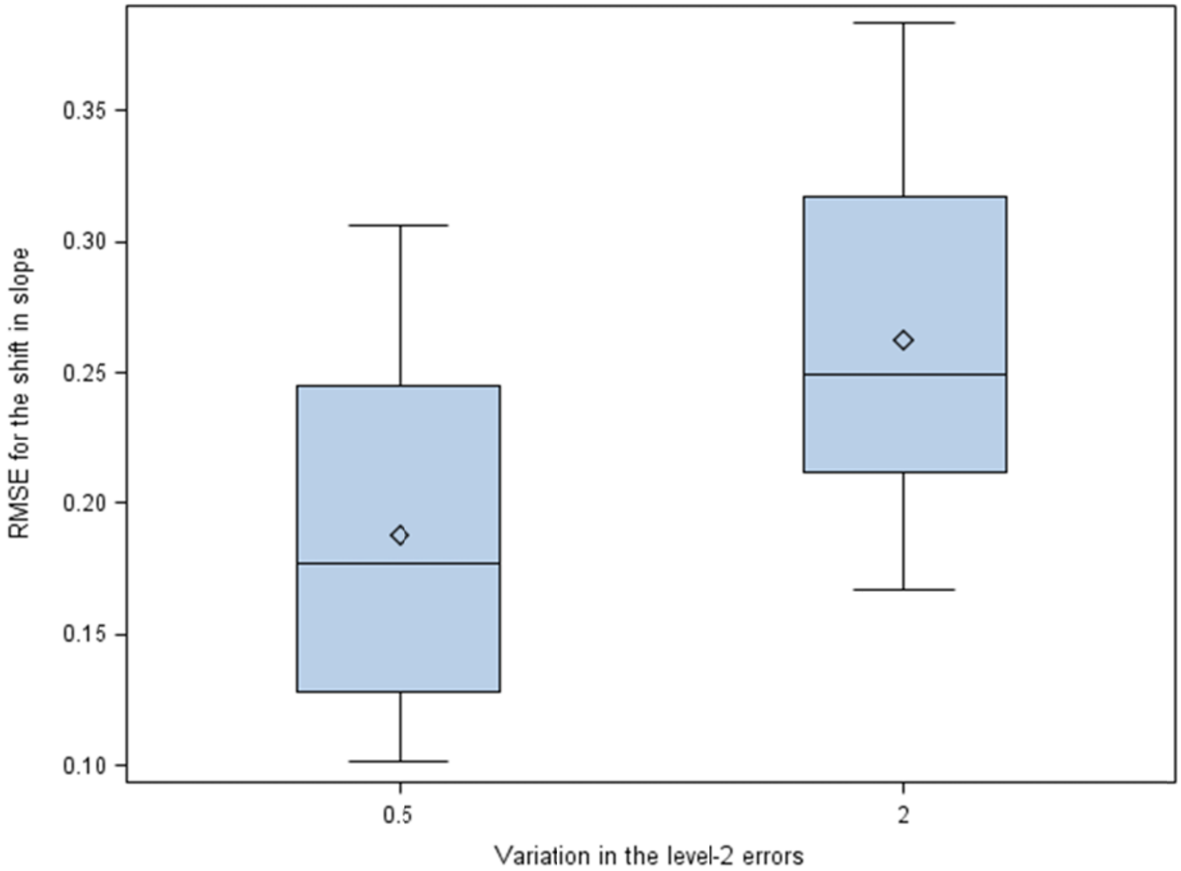


Figure 31. Box plots depicting the estimated RMSE values for the shift in slope as a function of the variation in the level-2 errors.

Credible Interval Coverage

The distribution of CI coverage of the fixed effect for treatment effects (shift in level and shift in slope) are illustrated in Figures 32 through 38. The full information about the η^2 values for the GLM models is also provided in Appendix B.

Average treatment effect for phase (shift in level). The average 95% credible interval (CI) coverage values of the treatment effect for phase exceeded 95% for both models (Model 1 and Model 2) (Figure 32). The average CI coverage for Model 1 was *M* = 0.98, *SD* = 0.01, and the average CI coverage for Model 2 was *M* = 0.98, *SD* = 0.01. The type of model explained little of the variability (η^2 = .001).

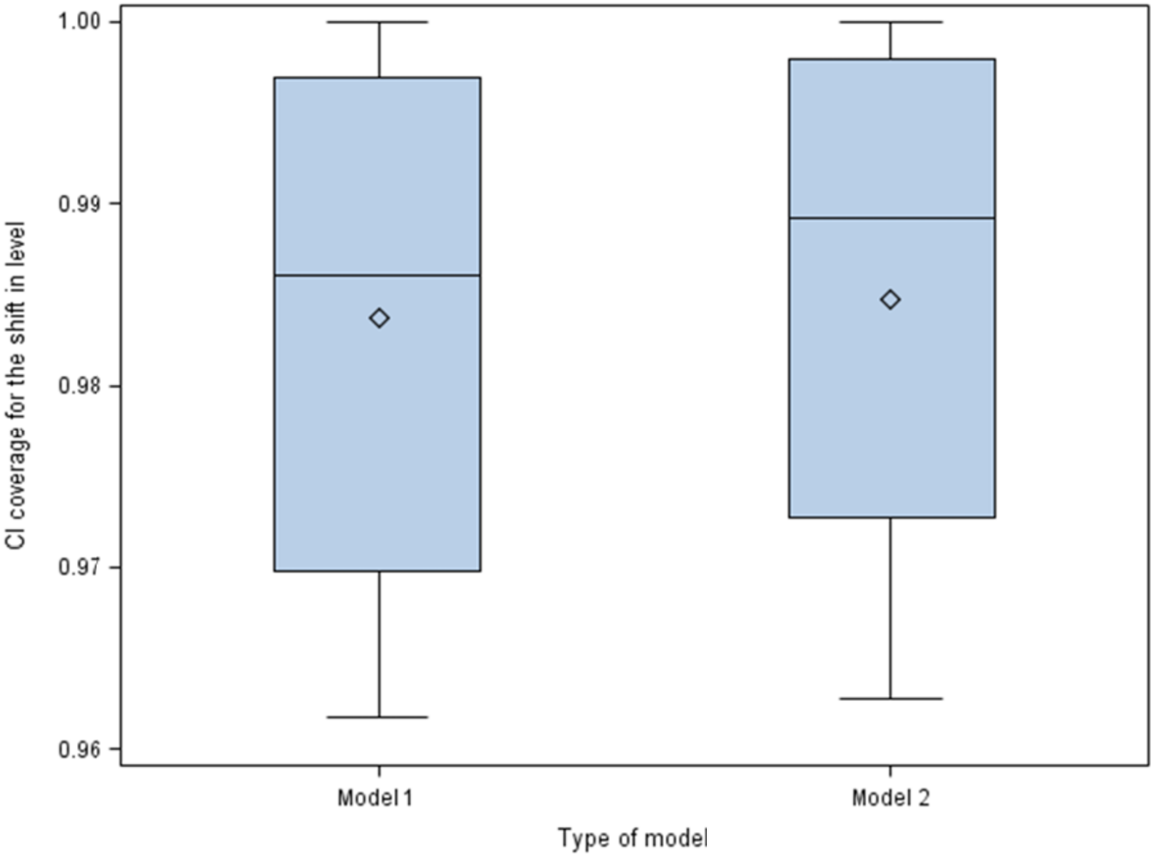


Figure 32. Box plots illustrating the distribution for the CI coverage for the shift in level across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI coverage values of the treatment effect for phase across the two models were also examined within the three different types of the true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 33). The average CI coverage values were very similar across the two models within the three true level-1 error structures, and there were no or little differences across the true level-1 error structures, with little of the variability explained by the different types of the true level-1 error structures (η^2 = .008). The smallest average CI coverage difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1|$ = 0), and the biggest average CI coverage difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2\text{-}M_1|$ = 0.002).

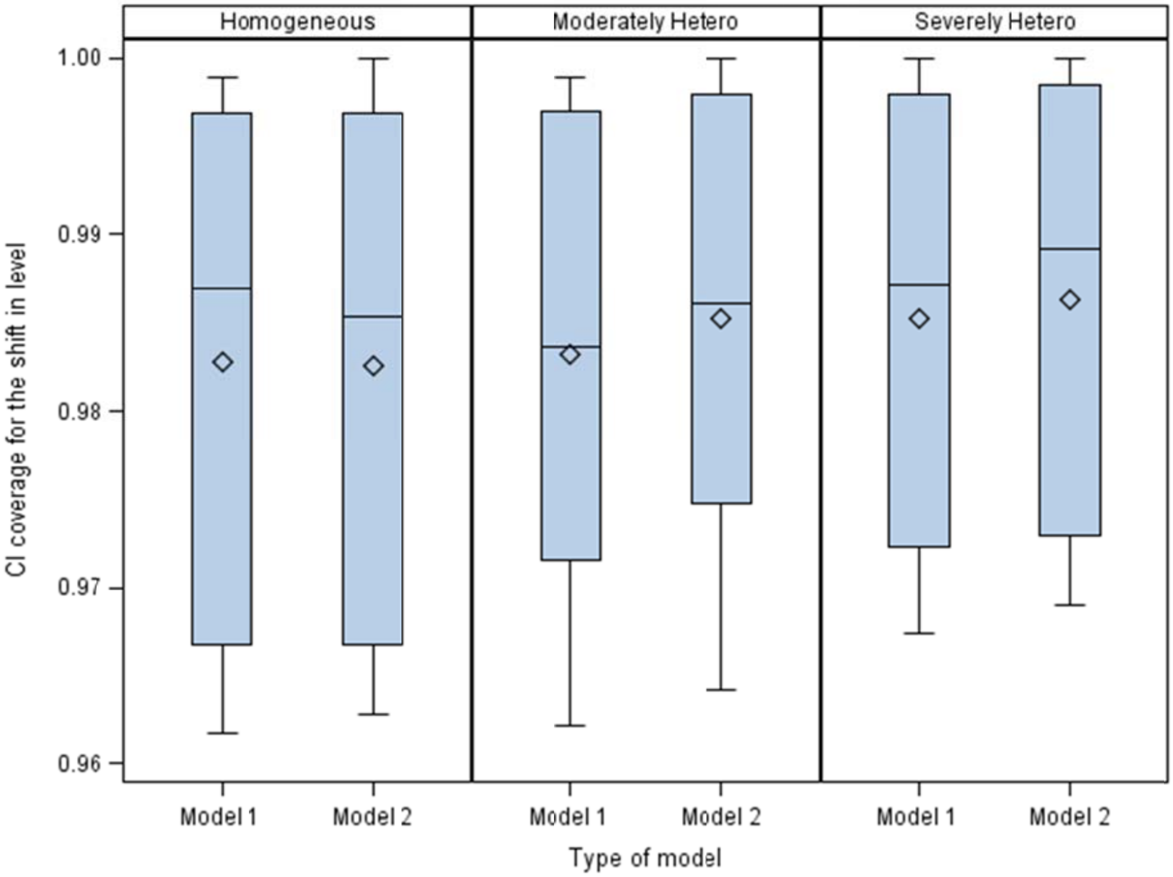


Figure 33. Box plots illustrating the distribution for the CI coverage for the shift in level across the two models within the three true level-1 error structures.

In order to further explore the variability in the CI coverage for phase effect, a GLM model was run. The main effects only model explained 94% of the variability. It was found that one of the design factors, the number of cases, had a large effect (η^2 = .88). This main effect is illustrated using a box plots in Figure 34.

As illustrated in Figure 34, as the number of cases increased from 4 to 8, the average CI coverage approached the nominal level, .95 (from M = 0.997, SD = .002 to M = 0.971, SD = .006).

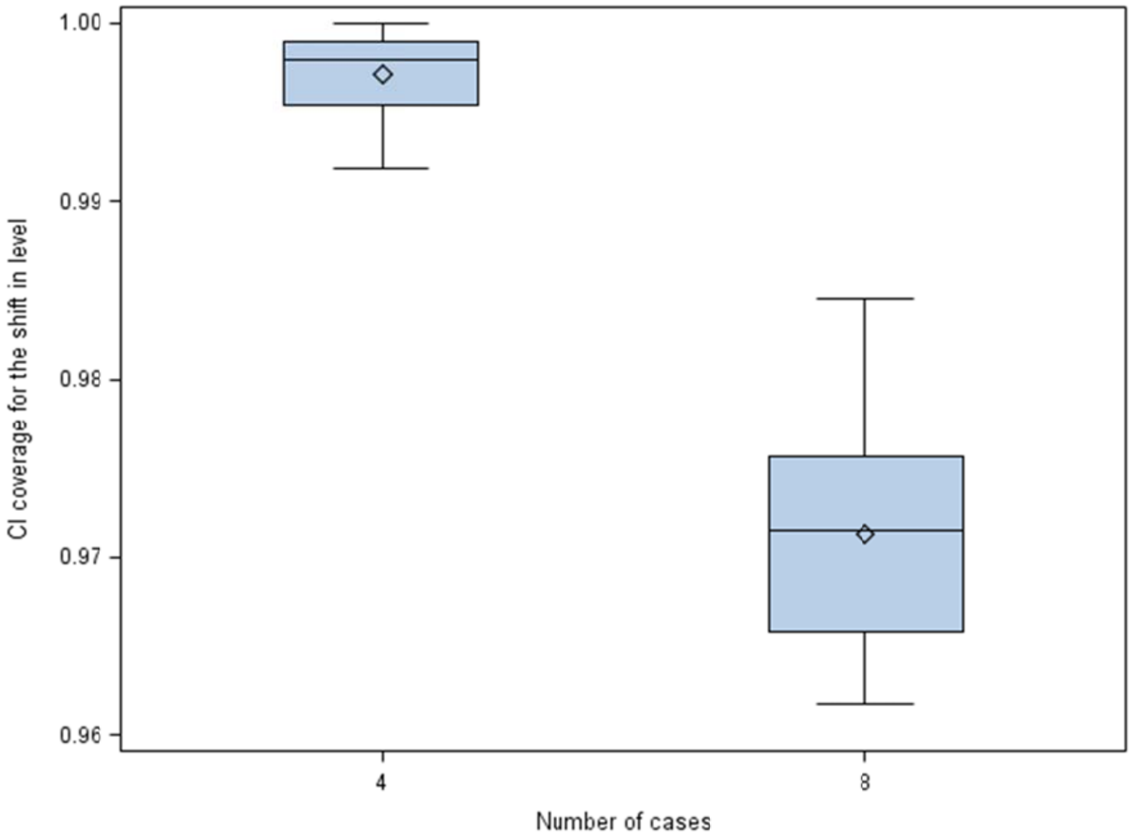


Figure 34. Box plots depicting the estimated CI coverage for the shift in level as a function of the number of cases.

Average treatment effect for interaction (shift in slope). The average credible interval (CI) coverage values for the treatment effect for interaction exceeded .95 for the two models (Model 1 and Model 2) (Figure 35). The average CI coverage value for Model 1 was *M* = 0.985, *SD* = 0.01, and the average CI coverage value for Model 2 was *M* = 0.986, *SD* = 0.01. The type of model explained little of the variability (η^2 = .0003).

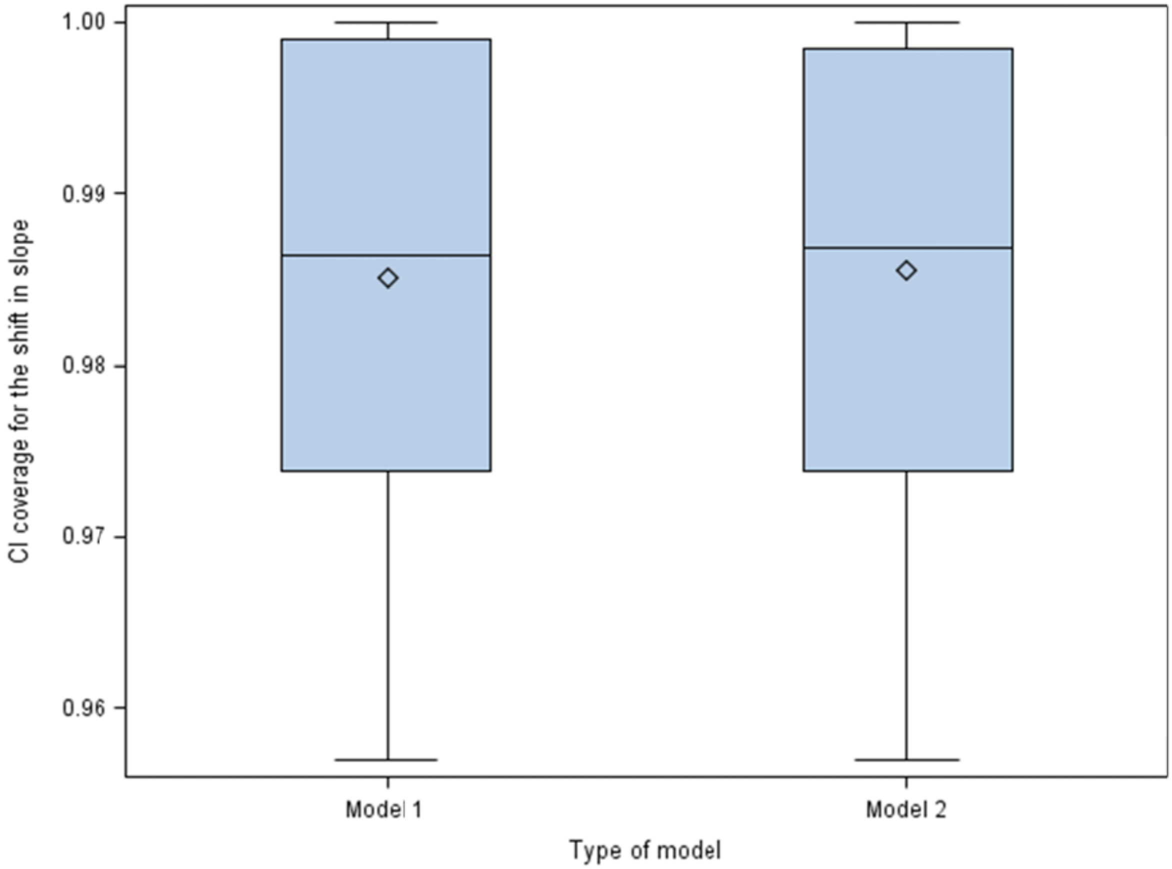


Figure 35. Box plots illustrating the distribution for the CI coverage for the shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI coverage values of the treatment effect for interaction across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 36). The average CI coverage values were very similar across the two models within the three true level-1 error structures, and there were no or little differences across the true level-1 error structures, with little of the variability explained by the different types of the true level-1 error structures (η^2 = .0005). The smallest average CI coverage difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2$-$M_1|$ = 0.0000), and the biggest average CI coverage difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1| = 0.0009$).

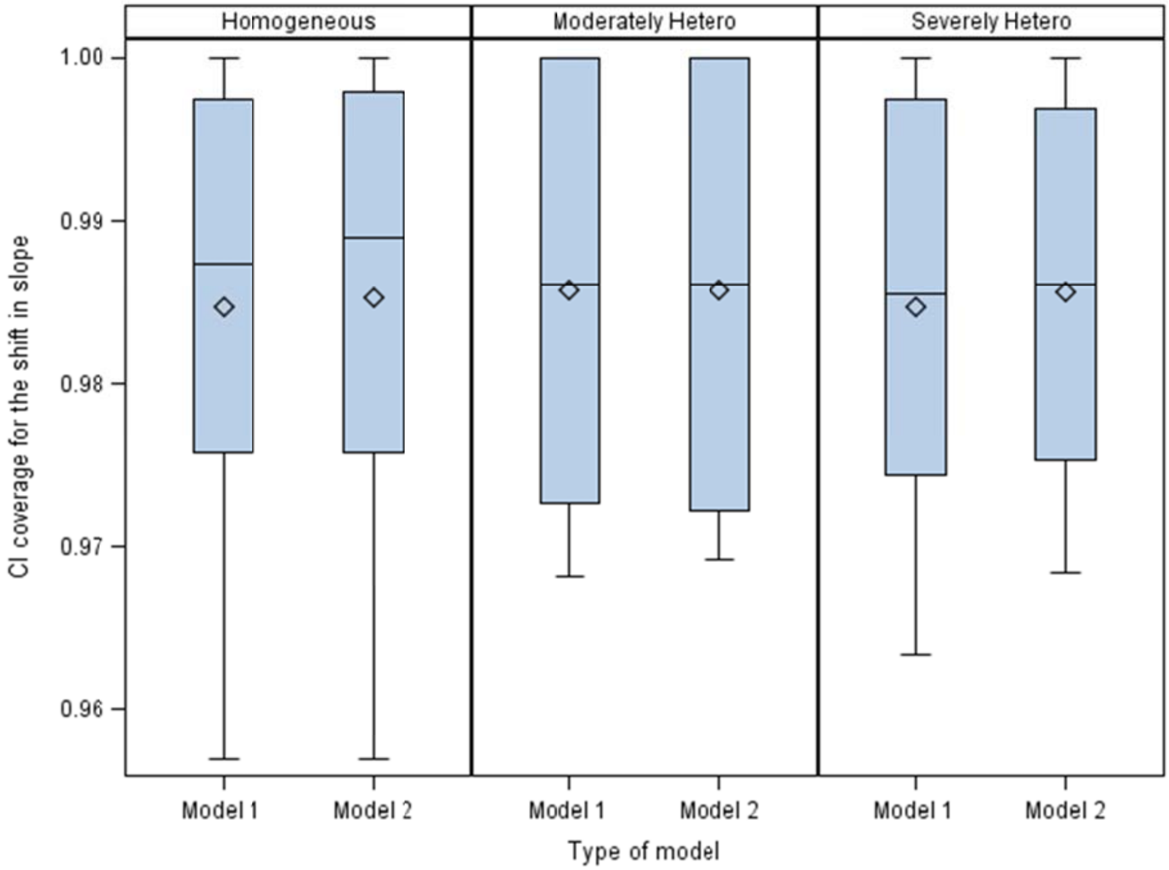


Figure 36. Box plots illustrating the distribution of the CI coverage for the shift in slope across the two models within the three true level-1 error structures.

In order to further explore the variability in the CI coverage for the treatment effect for interaction, GLM models were run. The model explained 96% of variability after including 2-way interactions. The GLM model found two of the design factors that had a medium or large effect, including the number of cases ($\eta^2 = .83$) and the series length per case ($\eta^2 = .08$). These two main effects are illustrated in Figures 37 and 38.

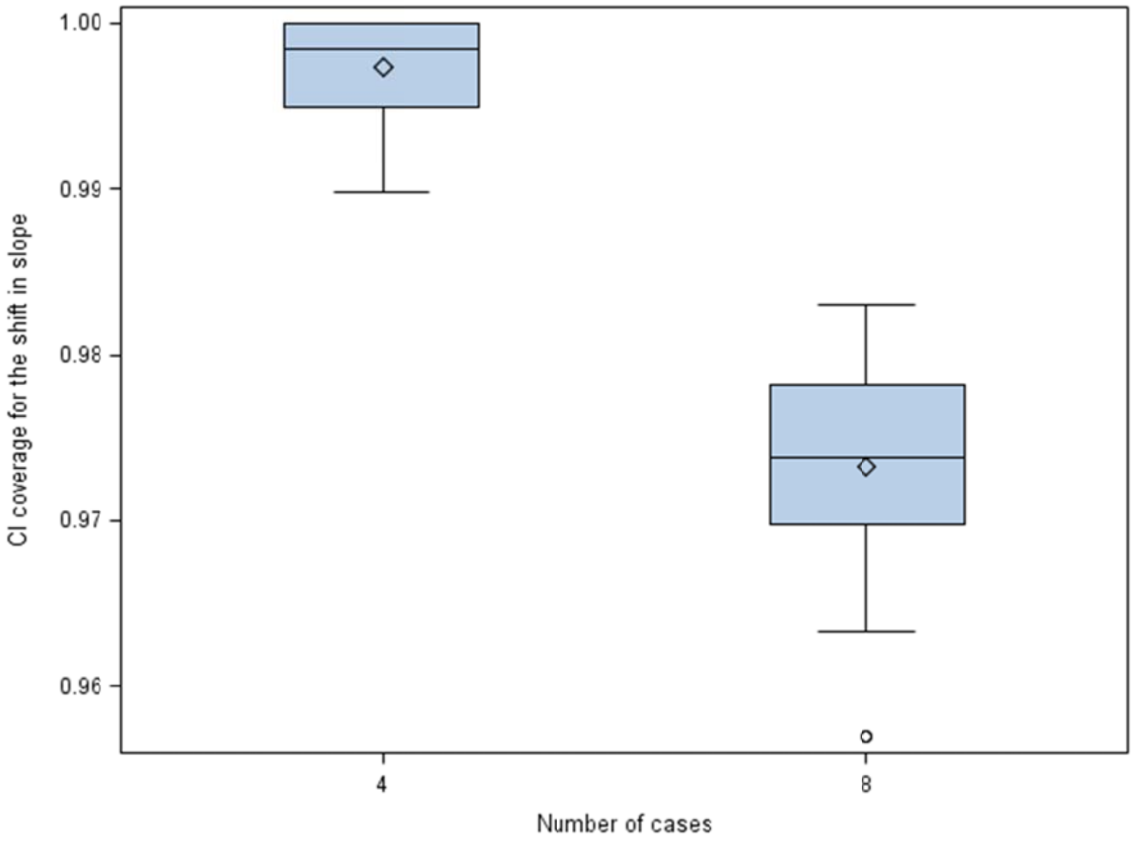


Figure 37. Box plots depicting the estimated CI coverage for the shift in slope as a function of the number of cases.

Similar to the GLM result of the phase effect, Figure 37 portrays that as the number of cases increased from 4 to 8, the average CI coverage approached the nominal level, .95 (from *M* = 0.997, *SD* = .003 to *M* = 0.973, *SD* = .007).

Similarly, Figure 38 depicts that as the series length per case increased from 10 to 20, the average CI coverage approached the nominal level, .95 (from M = 0.989, SD = .01 to M = 0.982, SD = .01).

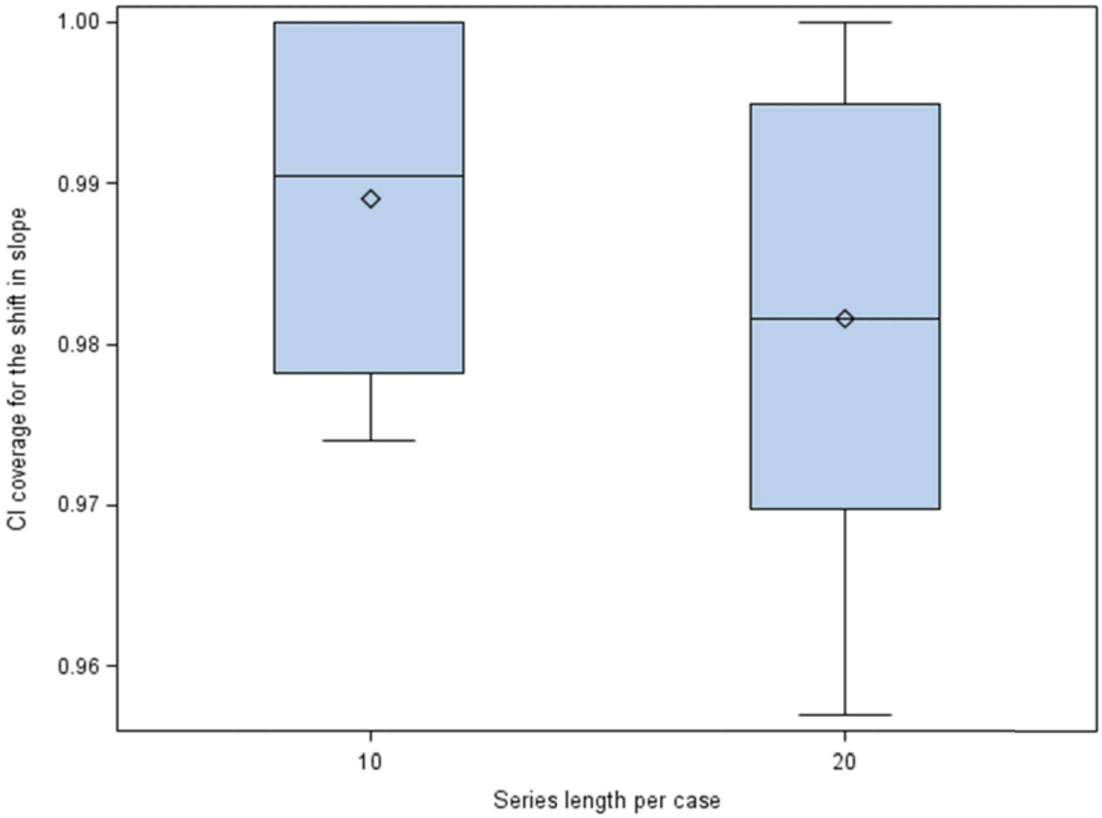


Figure 38. Box plots depicting the estimated CI coverage for the shift in slope as a function of the series length per case.

Credible Interval Width

The distribution of CI width of the fixed effect for treatment effects (shift in level and shift in slope) are illustrated in Figures 39 through 47. The full information about the η^2 values for the GLM models is also provided in Appendix B.

Average treatment effect for phase (shift in level). The average credible interval (CI) width values of the treatment effect for phase were very similar across the two models (Model 1 and Model 2) (Figure 39). The average CI width value for Model 1 was $M = 4.96$, $SD = 2.47$, and the average CI width value for Model 2 was $M = 4.95$, $SD = 2.47$. The type of model explained little of the variability ($\eta^2 < .00001$).

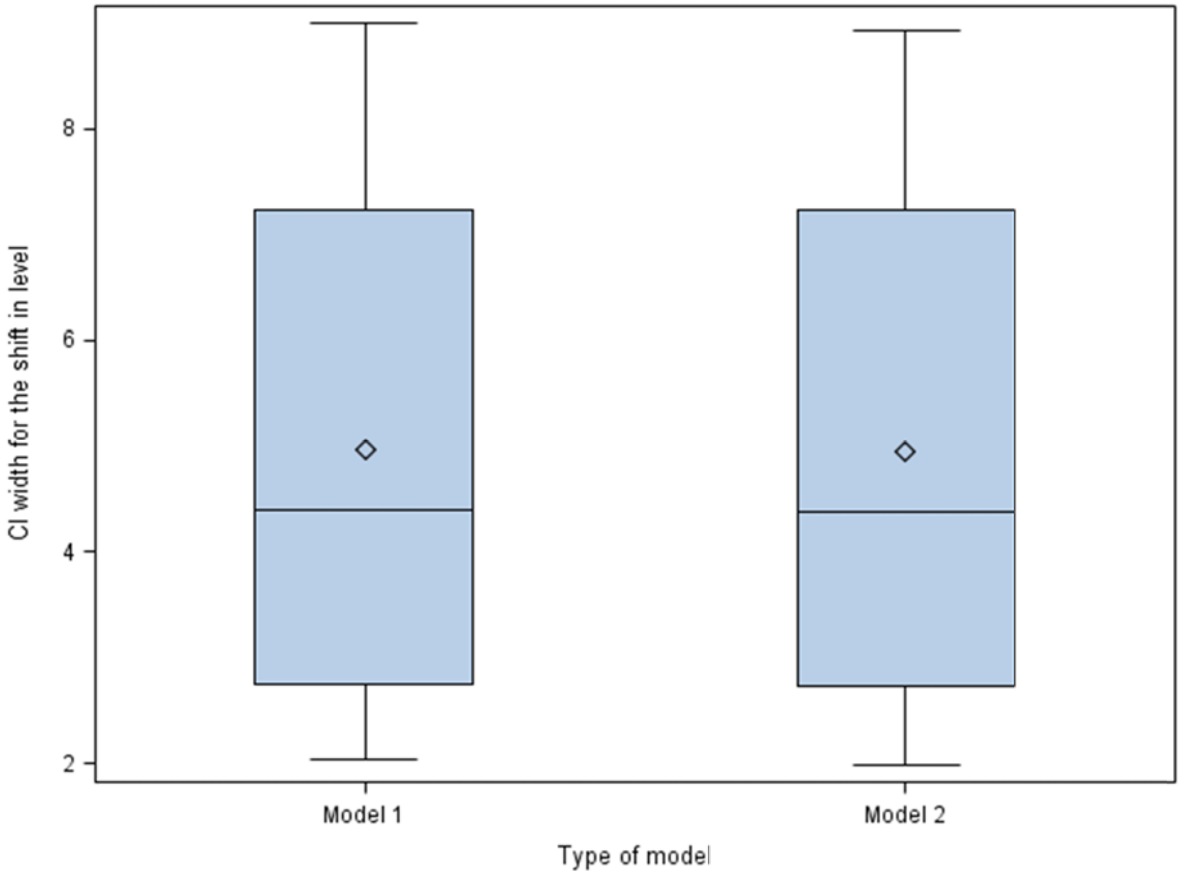


Figure 39. Box plots illustrating the distribution for the CI width for the shift in level across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI width values of the treatment effect for phase across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 40). The average CI width values were very similar across the two models within the three true level-1 error structures, and there were no or little differences across the three true level-1 error structures, with little of the variability explained by the different types of the true level-1 error structures (η^2 = .0002). The smallest average CI width difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2\text{-}M_1|$ = .01), and the biggest average CI width difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2$-$M_1|$ = .06).

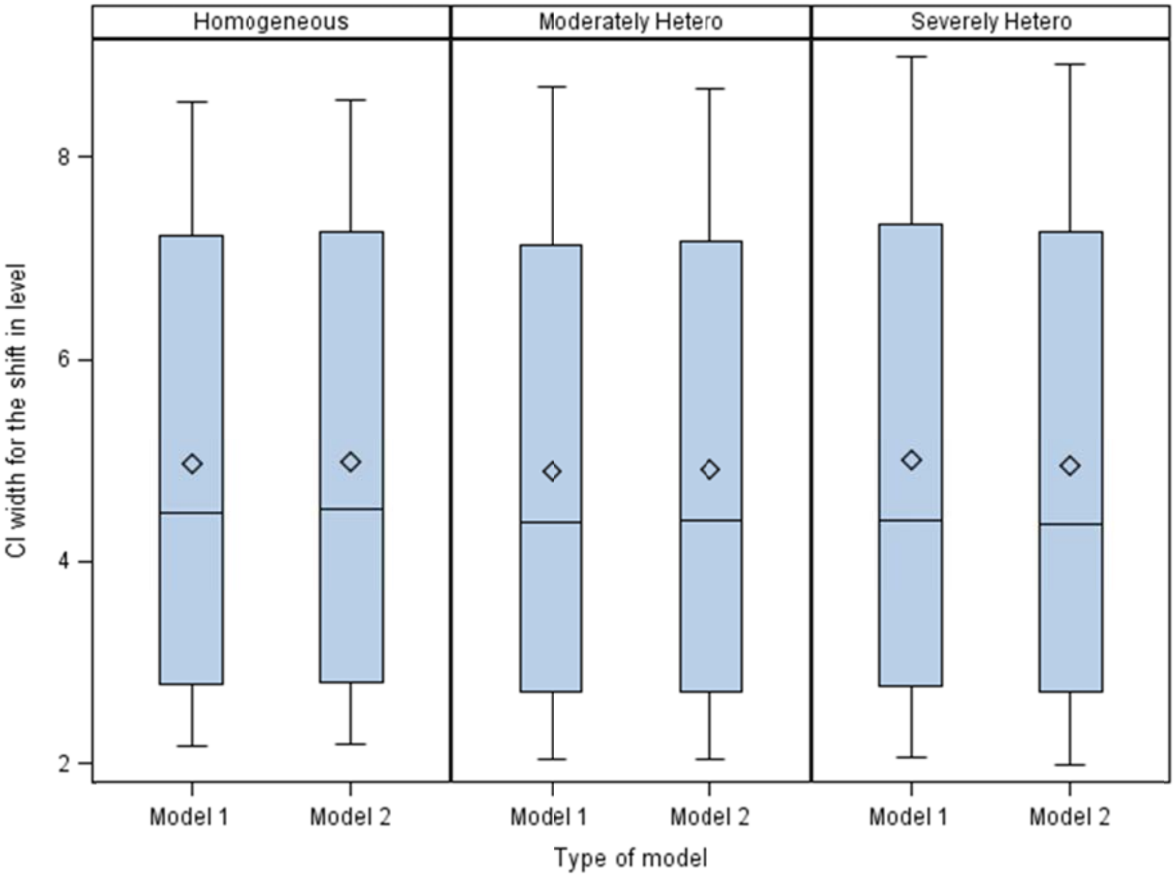


Figure 40. Box plots illustrating the distribution of the CI width for the shift in level across the two models within the three true level-1 error structures.

In order to further explore the variability in the CI width of the treatment effect for phase, a GLM model was run. The main effects only model explained 97% of the variability. It was found that two of the design factors had a medium or large effect, including the number of cases (η^2 = .84) and the variation in the level-2 errors (η^2 = .12). These main effects are illustrated in Figure 41 and 42.

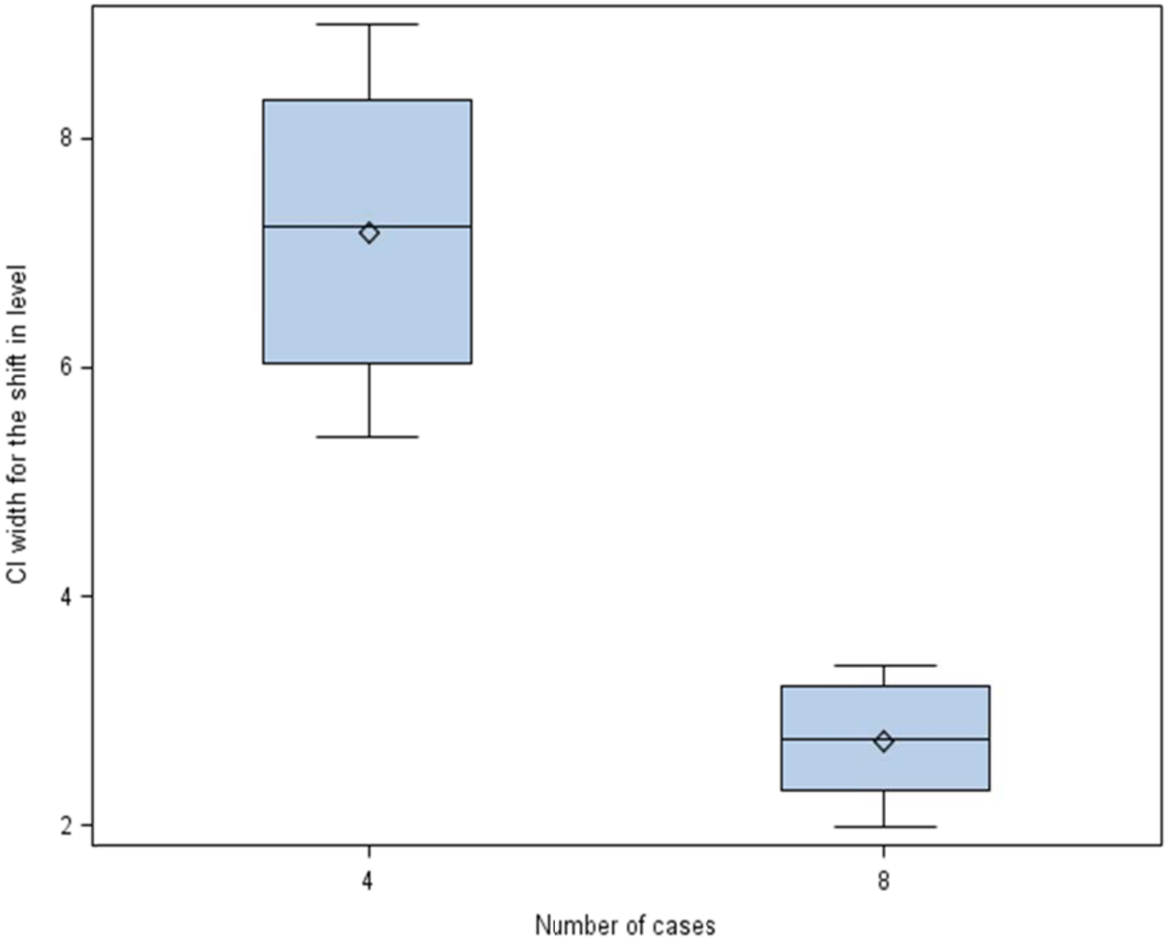


Figure 41. Box plots depicting the estimated CI width for the shift in level as a function of the number of cases.

As illustrated in Figure 41, as the number of cases increased from 4 to 8, the average CI width decreased from *M* = 7.17, *SD* = 1.30 to *M* = 2.74, *SD* = 0.50. Similarly, as the variation in the level-2 errors shifted from most of the variance at the level-1 error (.5) to most of the variance at the level-2 error (2), the average CI width increased from *M* = 4.13, *SD* = 1.92 to *M* = 5.78, *SD* = 2.67.

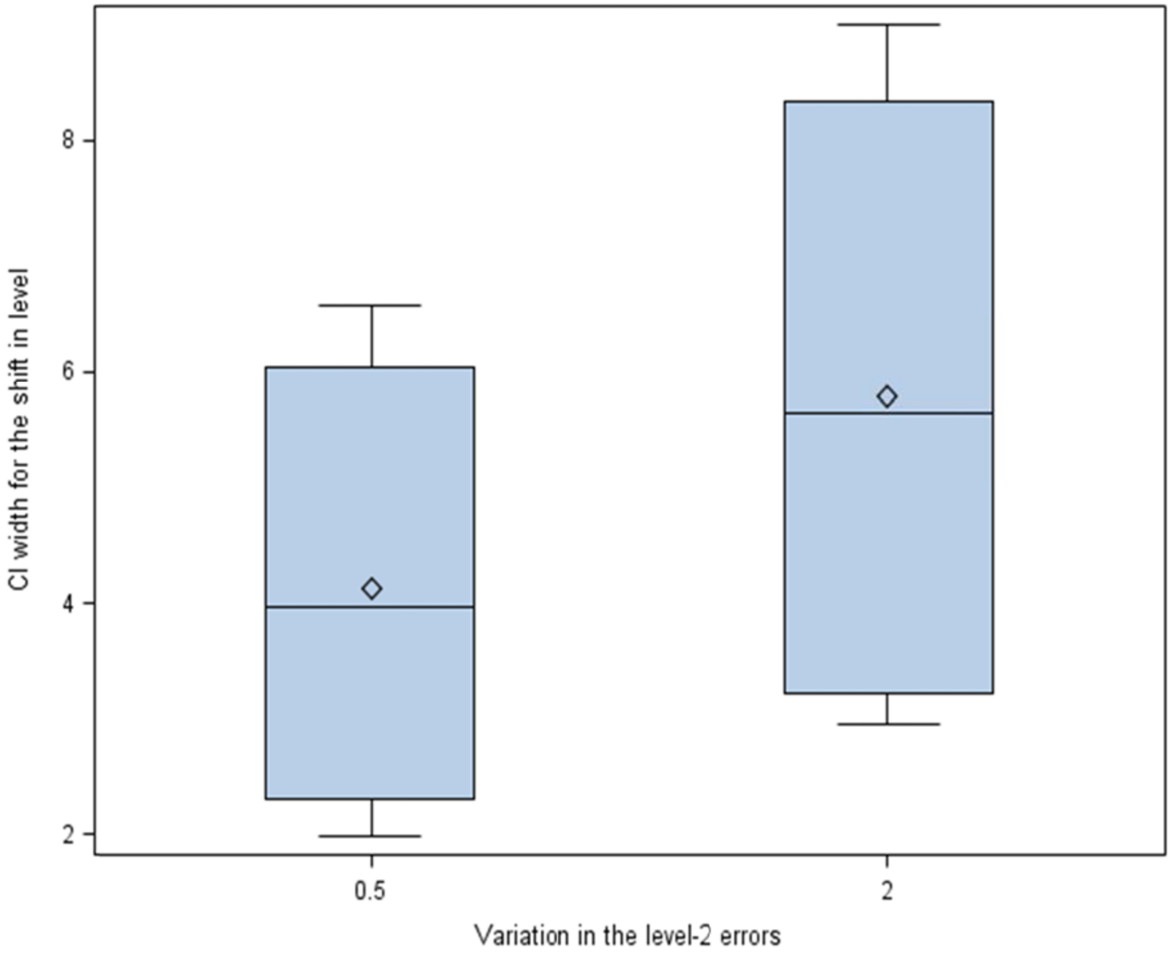


Figure 42. Box plots depicting the estimated CI width for the shift in level as a function of the variation in the level-2 errors.

Average treatment effect for interaction (shift in slope). The average credible interval (CI) width values for the treatment effect for interaction were very similar across the two models (Model 1 and Model 2) (Figure 43). The average CI width value for Model 1 was $M = 1.68$, $SD = 0.96$, and the average CI width value for Model 2 was $M = 1.68$, $SD = 0.95$. The type of model explained little of the variability ($\eta^2 = .00003$).

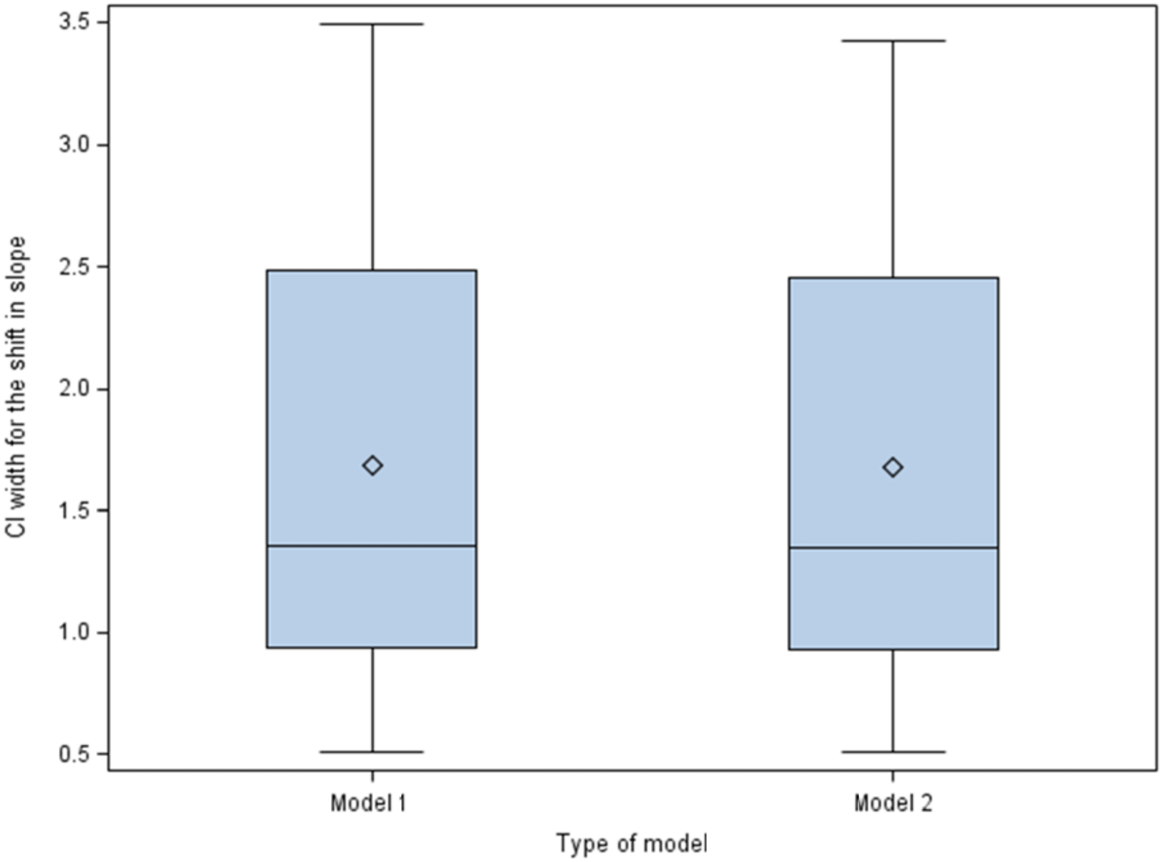


Figure 43. Box plots illustrating the distribution of the CI width values for the shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI width values of the treatment effect for interaction across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 44). The average CI width values were very similar across the two models within the three true level-1 error structures, and there were no or little differences across the three true level-1 error structures, with little of the variability explained by the different types of the true level-1 error structures (η^2 = .0005). The smallest average CI width difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2$-$M_1|$ = .001), and the biggest average CI width difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2$-$M_1| = .030$).

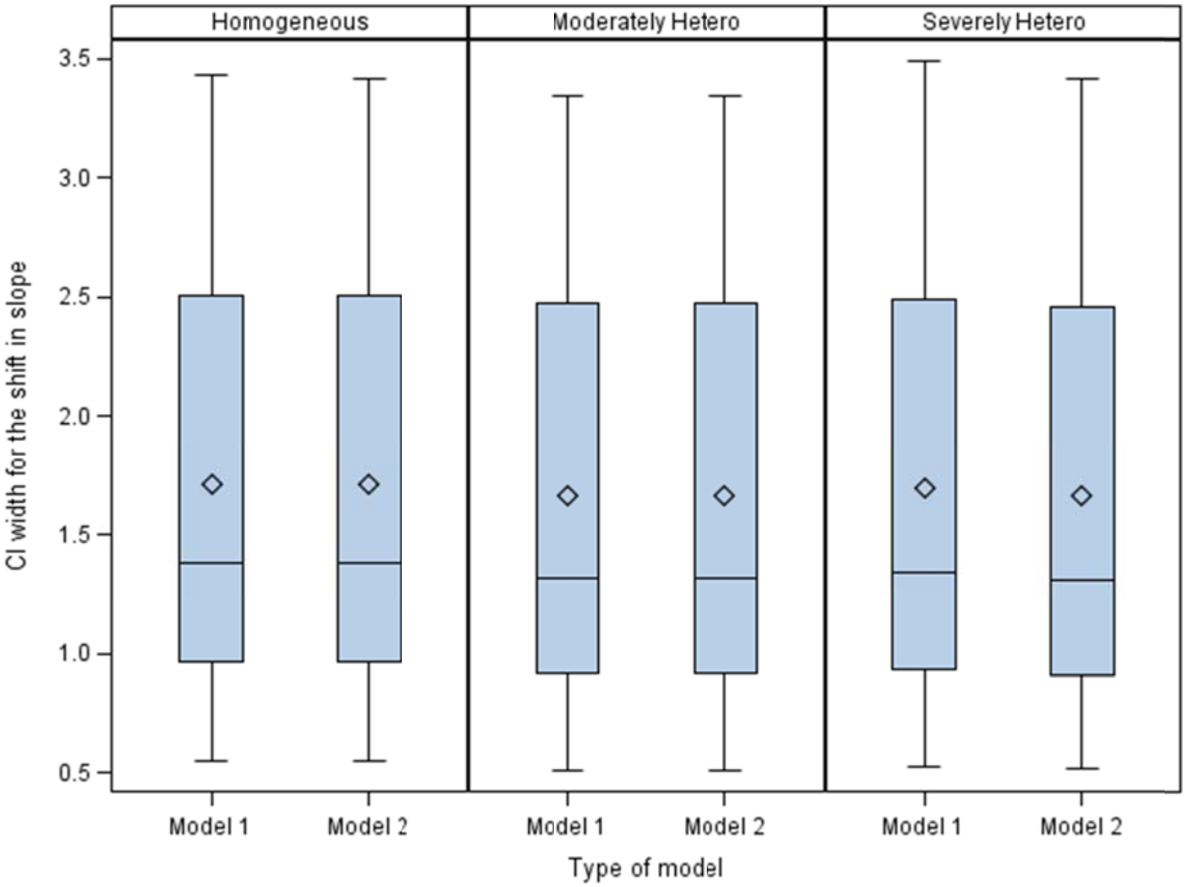


Figure 44. Box plots illustrating the distribution of the CI width values for the shift in slope across the two models within the three true level-1 error structures.

In order to further explore the variability in the CI width of the treatment effect for interaction, a GLM model was run. The main effects only model explained 94% of the variability. It was found that three of the design factors had a medium or large effect, including the number of cases (η^2 = .65), the series length per case (η^2 = .19), and the variation in the level-2 errors (η^2 = .10).

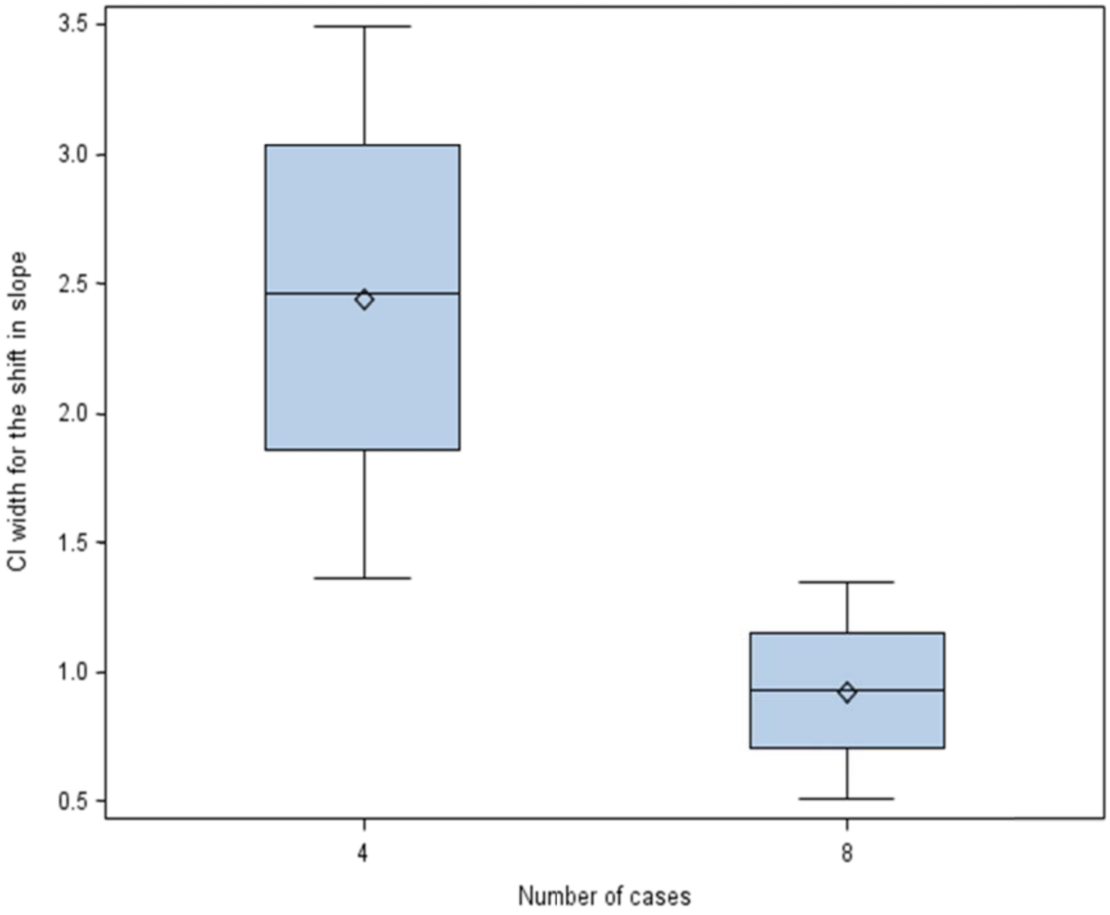


Figure 45. Box plots depicting the estimated CI width for the shift in slope as a function of the number of cases.

These main effects are illustrated in Figure 45, 46, and 47. As illustrated in Figure 45, as the number of cases increased from 4 to 8, the average CI width decreased from $M = 2.44$, $SD = 0.75$ to $M = 0.92$, $SD = 0.29$. Similarly, as the series length per case increased from 10 to 20, the average CI width decreased from $M = 2.09$, $SD = 1.01$ to $M = 1.27$, $SD = 0.68$.

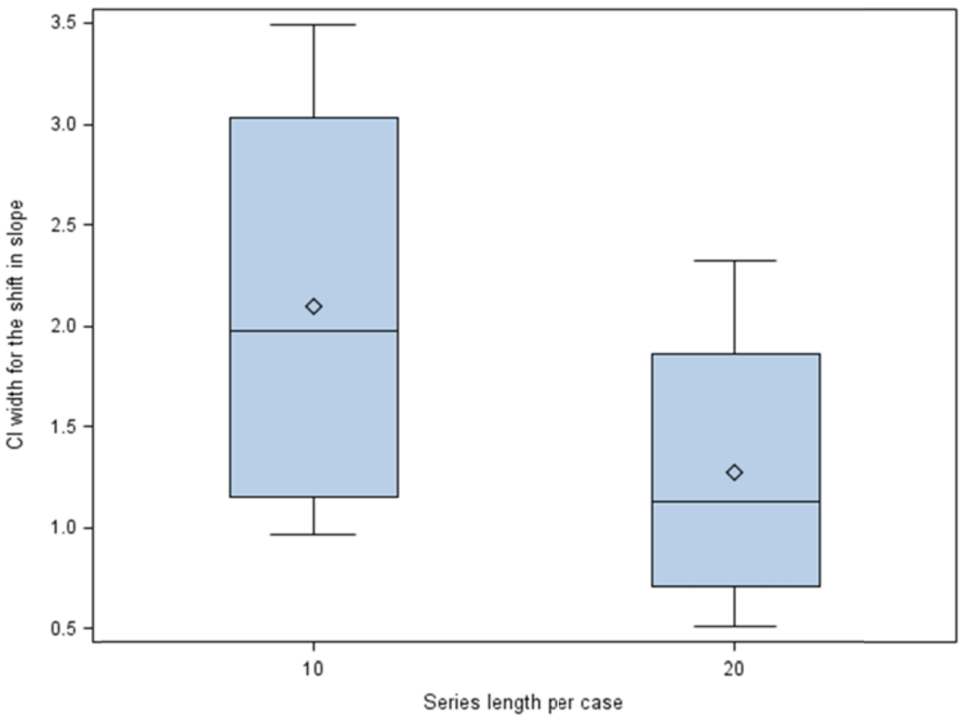


Figure 46. Box plot depicting the estimated CI width for the shift in slope as a function of the series length per case.

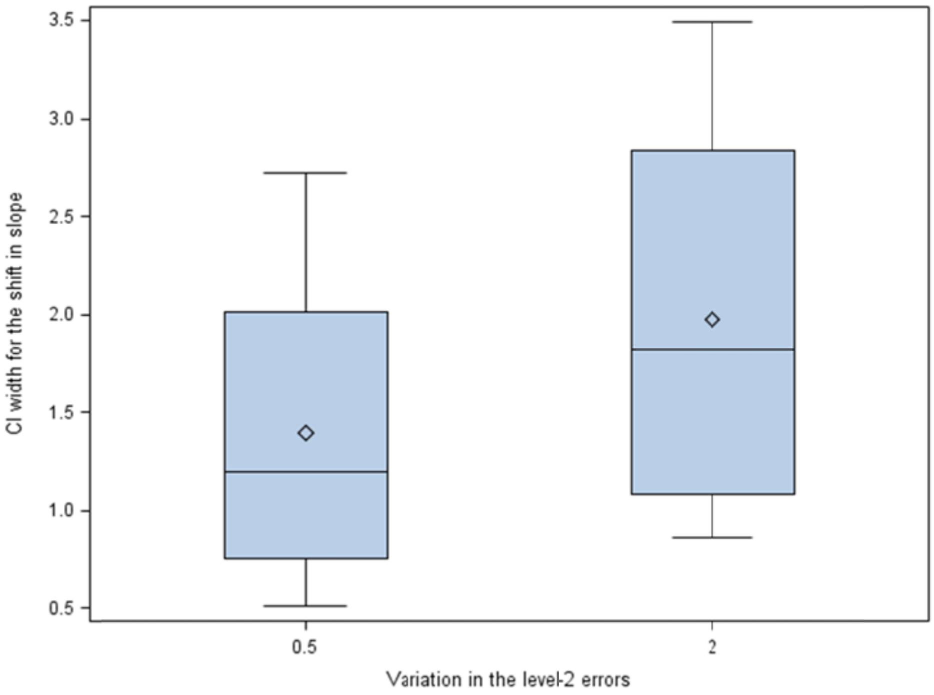


Figure 47. Box plot depicting the estimated CI width for the shift in slope as a function of the variation in the level-2 errors.

In addition, as the variation in the level-2 errors shifted from most of the variance at the level-1 error to most of the variance at the level-2 error, the average CI width increased from *M* = 1.39, *SD* = 0.81to *M* = 1.97, *SD* = 1.00.

In addition to the examination of the average fixed treatment effects, individual treatment effects were also examined in terms of the four outcome measures (Bias, RMSE, CI coverage and width). The results of the individual treatment effects were similar with the average fixed treatment effects across all outcome measures. Although the CI coverage and the widths of the individual treatment effects were closer to the nominal level, and narrower than the CI coverage and the widths of the average treatment effects, there was no substantial difference across the two models, which is consistent with the average treatment effects results. Since the interest of the current study is focused on the average treatment effects rather than the individual treatment effects, and the results of both the average and the individual treatment effects were very similar, the results of the average treatment effects were only provided in this section. However, the summary table and the figures of the individual treatment effects were provided in Appendix C for the researchers who are interested in the results of the individual treatment effects.

Variance Components

The second research question considers the estimates of the variance components and the consequences of modeling and not modeling between case variation in the level-1 error structure. More specifically, (1) the bias and the RMSE for the variance components as function of the design and data factors, and (2) the credible interval coverage and width for the variance components as function of the design and data factors. All variance components parameters results are displayed in standard deviation units, since the results of the variance components parameters were produced in the standard deviation units in all analyses.

Bias

The distribution of bias values of the level-2 error standard deviation of intervention effects (shift in level and shift in slope), the level-1 error standard deviation, and autocorrelation are illustrated in Figures 48 through 62. Relative bias values for the all parameters are provided in Appendix A. The full information about the η^2 values for the GLM models is also provided in Appendix B.

Level-2 error standard deviation for phase (shift in level). The average bias values of the level-2 error standard deviation (SD) for phase were similar and positively biased across the two models (Model 1 and Model 2) with little variability explained by the type of model (η^2 = .00005) (Figure 48). The average bias value for Model 1 and Model 2 was *M* = 0.86, *SD* = 0.64 and *M* = 0.85, *SD* = 0.63, respectively.

The average bias values were similar across the two models within the three true level-1 error structures. The different types of the true level-1 error structures explained little of the variability (η^2 = .0004) which indicates similarity of the average bias across the true level-1error structures. Specifically, the smallest average bias value was found when the true level-1 error structure was moderately heterogeneous and estimated by Model 2 (*M*= 0.83, *SD* = .66), and the largest average bias was found when the true level-1 error structure was severely heterogeneous and estimated by Model 1 (*M*= 0.87, *SD* = .69).

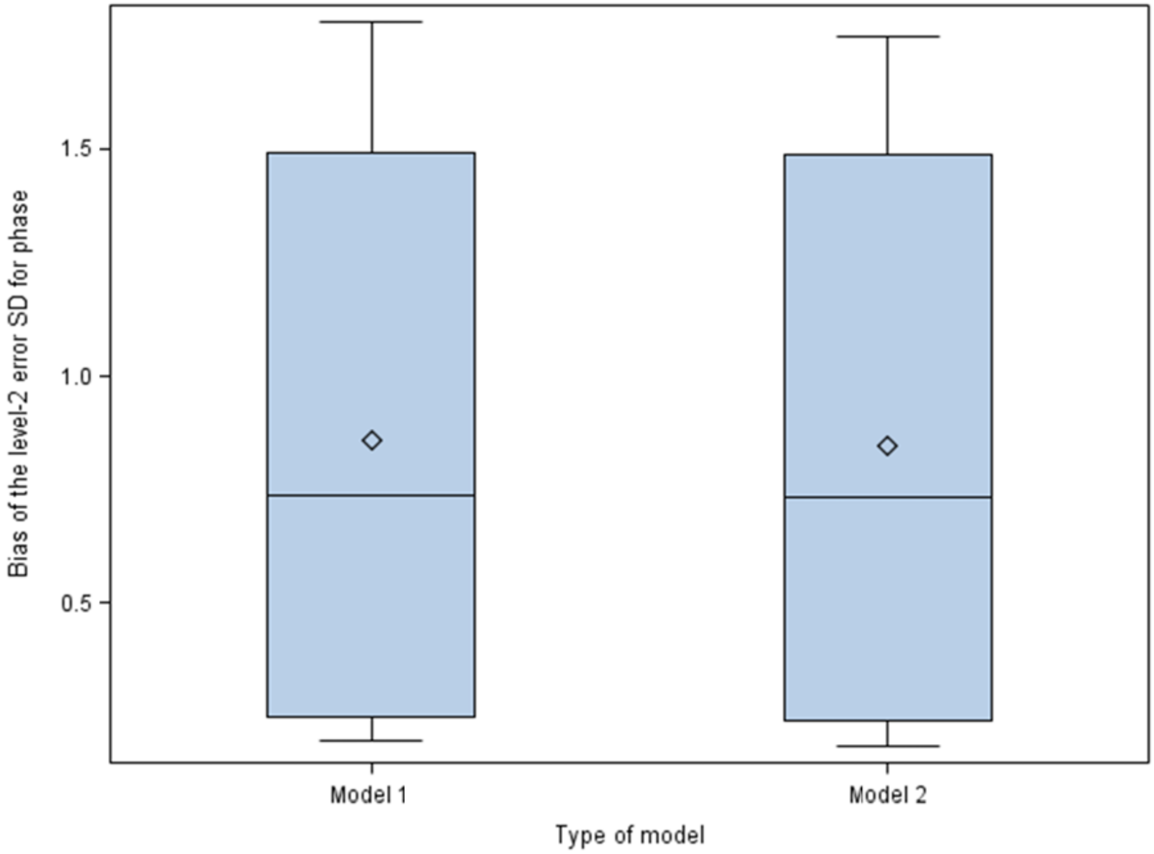


Figure 48. Box plots illustrating the distribution of the bias values for the level-2 error standard deviation of shift in level across Model 1 which did not model between case variation, and Model 2 which models between case variation.

In addition, there were very little differences across the two models within the three different types of the true level-1 error structures. The smallest average bias difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1|$ = 0.001), and the biggest average bias difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1|$ = 0.018).

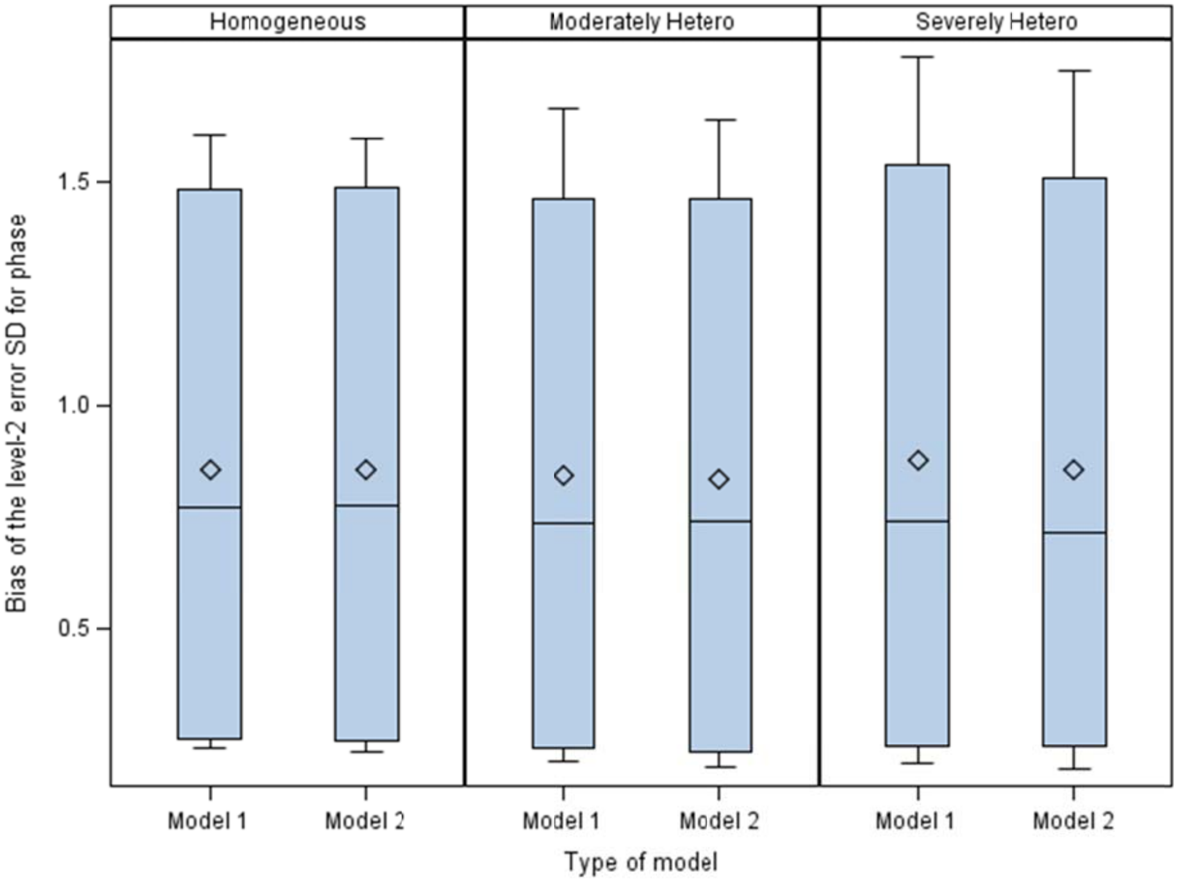


Figure 49. Box plots illustrating the distribution of the bias values for the level-2 error standard deviation of phase across the two models within the three true level-1 error structures.

In order to further explore if any design factor had a significant effect on the bias of the level-2 error standard deviation for the shift in level, a GLM model was run. The main effects only model explained 98% of the variability, and indicated one of the design factors, the number of cases, had a large effect (η^2 = .96). This main effect is illustrated using box plots in Figure 50.

As illustrated in Figure 50, as the number of cases increased from 4 to 8, the average bias value decreased from *M* = 1.46, *SD* = 0.18 to *M* = 0.25, *SD* = 0.04.

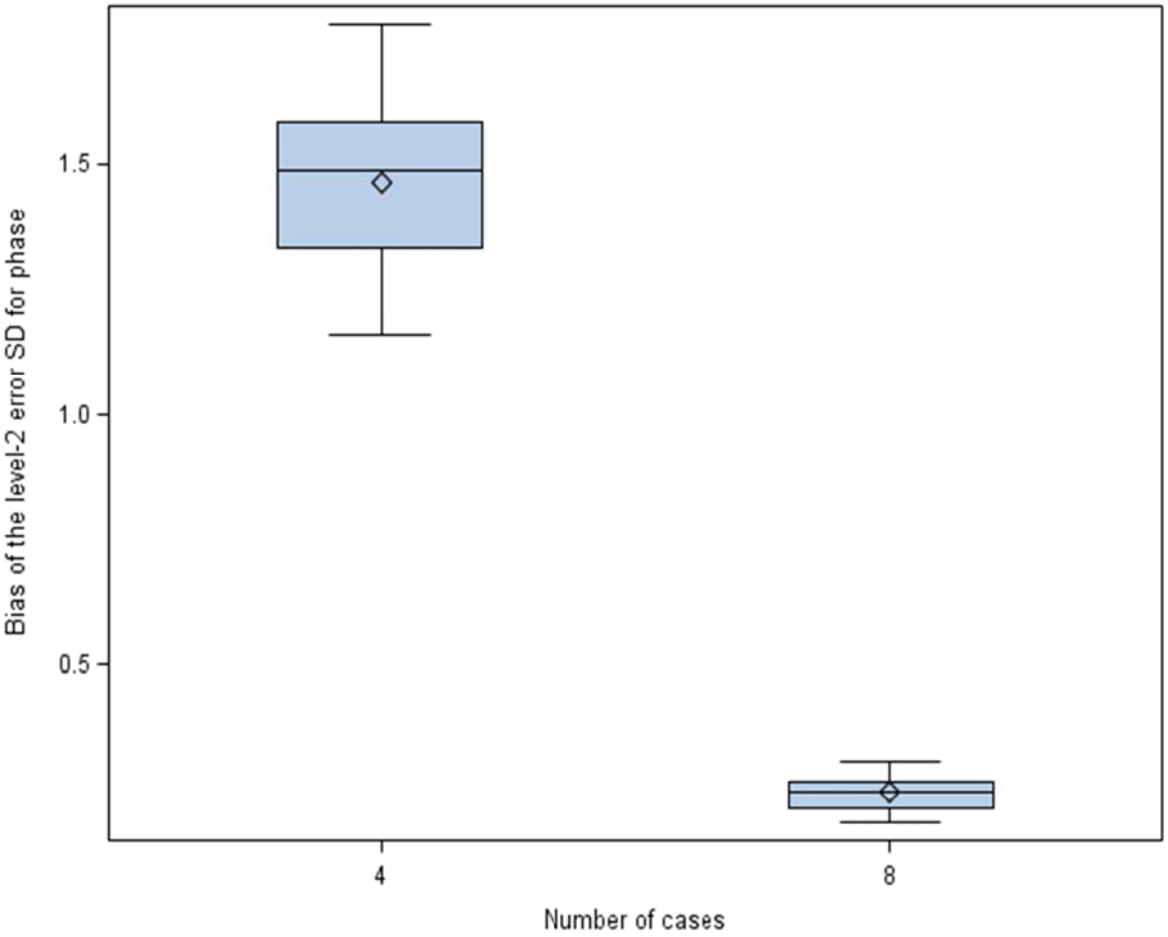


Figure 50. Box plots depicting the estimated bias of the level-2 error standard deviation for the shift in level as a function of the number of cases.

Level-2 error standard deviation for interaction (shift in slope). The average bias values of the level-2 error standard deviation for interaction were similar and positively biased across the two models (Model 1 and Model 2) with little variability explained by the type of model (η^2 = .0002) (Figure 51). The average bias value for Model 1 and Model 2 was *M* = 0.31, *SD* = 0.26 and *M* = 0.30, *SD* = 0.25, respectively.

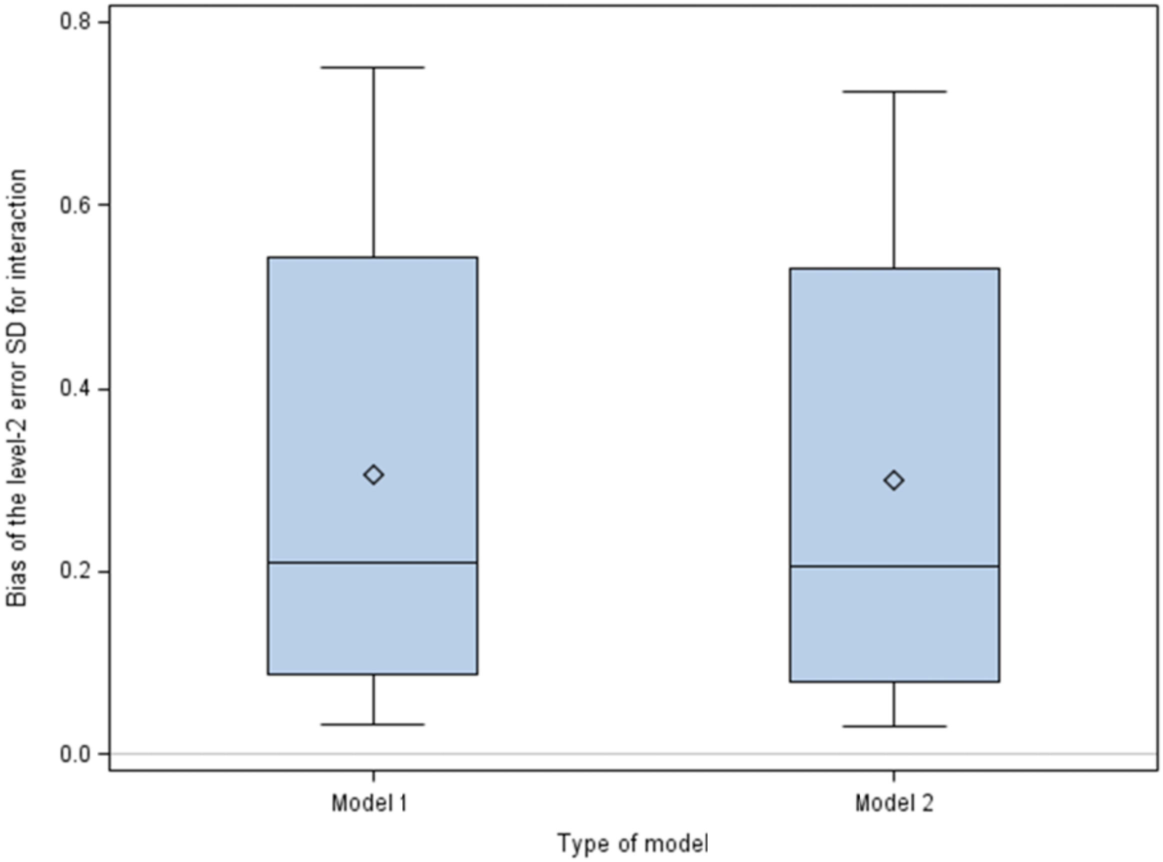


Figure 51. Box plots illustrating the distribution of the bias values for the level-2 error standard deviation of shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average bias values for the level-2 error standard deviation of the interaction effect across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 52). The average bias values were similar across the two models within the three true level-1 error structures. The different types of the true level-1 error structures explained little of the variability (η^2 = .0004) which indicates similarity of the average bias across the true level-1error structures. Specifically, the smallest average bias value was found when the true level-1 error structure was moderately heterogeneous and estimated by Model 2 (*M*= 0.29, *SD* = .26), and the largest average bias was found when the true level-1 error structure was severely heterogeneous and estimated by Model 1 (M= 0.31, SD = .27). In addition, there were very little differences across the two models within the three different types of true level-1 error structures. The smallest average bias difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1|$ = 0.003), and the biggest average bias difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1|$ = 0.012).

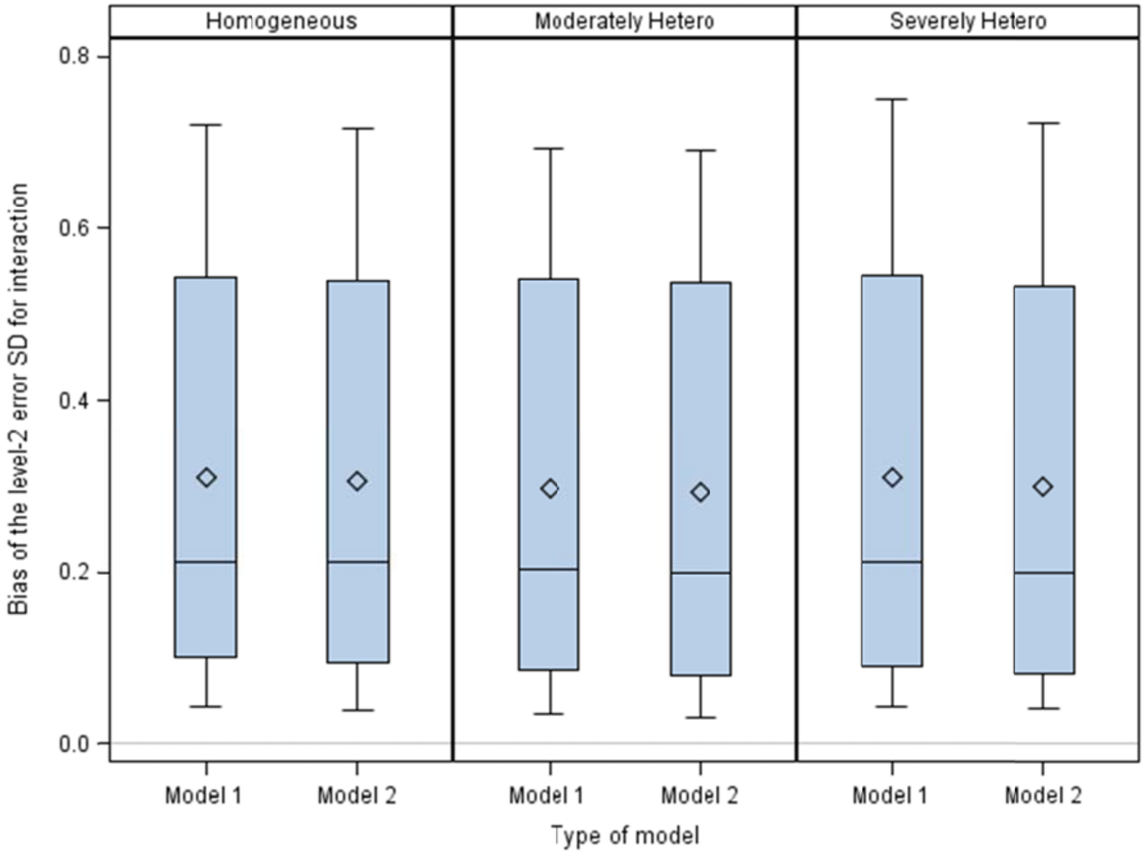


Figure 52. Box plots illustrating the distribution for the bias values of the level-2 error standard deviation for the shift in slope across the two models within the three true level-1 error structures.

In order to further explore if any design factor had a significant effect on the bias of the level-2 error standard deviation for the shift in slope, GLM models were run. The model explained over 99% of the variability after including 2-way interactions, and indicated a 2-way interaction between the number of cases and the series length per case had a medium effect (η^2 = .07). This interaction effect is illustrated using a line graph in Figure 53.

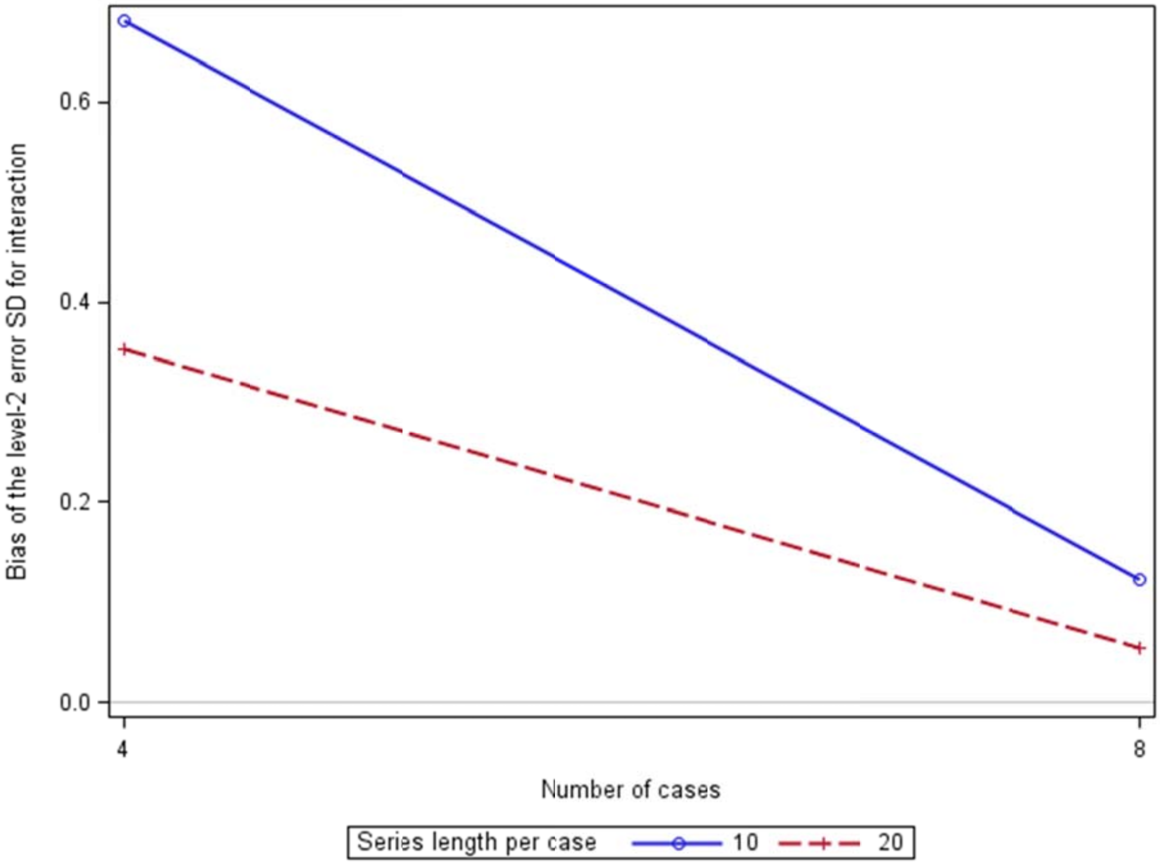


Figure 53. Line graph depicting average bias for the level-2 error standard deviation of shift in slope as a function of the two-way interaction effect between the number of cases and the series length per case.

The line graph shows that the effect of the number of cases (from 4 to 8) on the mean bias value was dependent on the series length per case. Specifically, when the series length per case was 10, mean bias value decreased greatly as the number of cases increased from 4 (*M* = 0.68, *SD* = 0.41) to 8(*M* = 0.12, *SD* =0.02). However, when the series length per case was 20, mean bias value decreased less as the number of cases increased from 4 (*M* = 0.35, *SD* = 0.08) to 8(*M* = 0.06, *SD* =0.02).

Level-1 error standard deviation. The average bias values of the level-1 error standard deviation were similar and positively biased across the two models (Model 1 and Model 2) with little variability explained by the type of model (η^2 = .005) (Figure 54). The average bias value for Model 1 and Model 2 was *M* = 0.05, *SD* = 0.03 and *M* = 0.04, *SD* = 0.02, respectively.

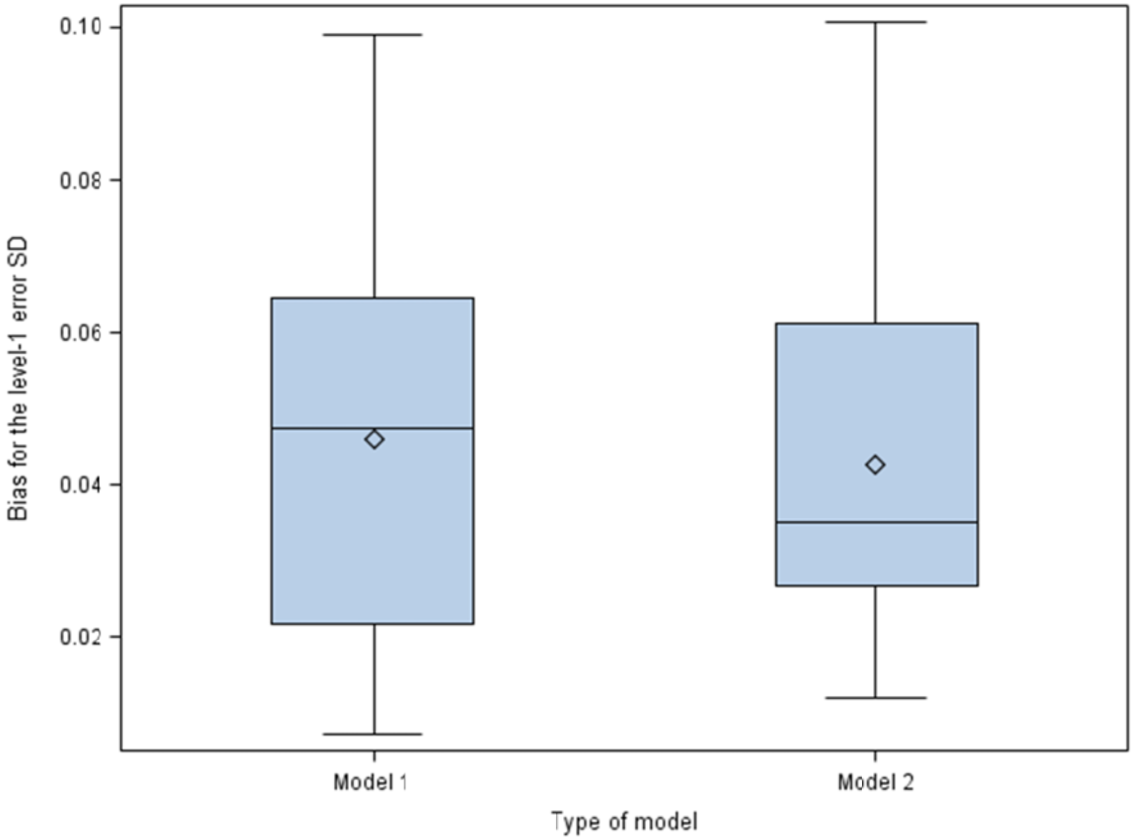


Figure 54. Box plots illustrating the distribution of the bias values for the level-1 error standard deviation across Model 1 which did not model between case variation, and Model 2 which models between case variation.

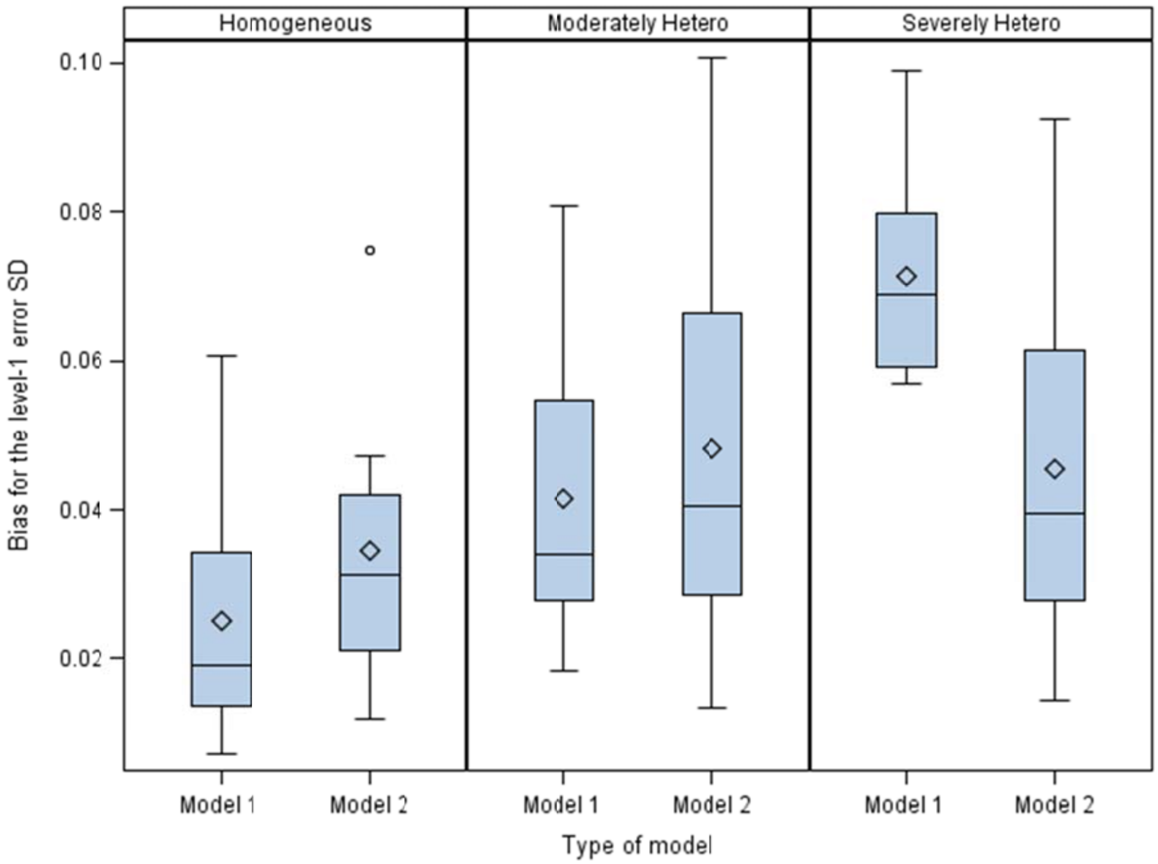


Figure 55. Box plots illustrating the distribution for the level-1 error standard deviation across the two models within the three true level-1 error structures.

The average bias values for the level-1 error standard deviation across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 55). The figure illustrated that the average bias values were different across the two models within the three true level-1 error structures with large variability explained by the different types of the true level-1 error structures (η^2 = .223). Specifically, the level-1 error standard deviation parameter tended to be more biased when estimated by Model 2 than Model 1 in the case that the true level-1 error structure was either homogeneous or moderately heterogeneous. However, the level-1 error standard deviation parameter tended to be more biased when estimated by Model 1 than Model 2 in the case that the true level-1 error structure was severely heterogeneous. In addition, the smallest average bias difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2\text{-}M_1| = 0.006$), and the biggest average bias difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1| = 0.025$).

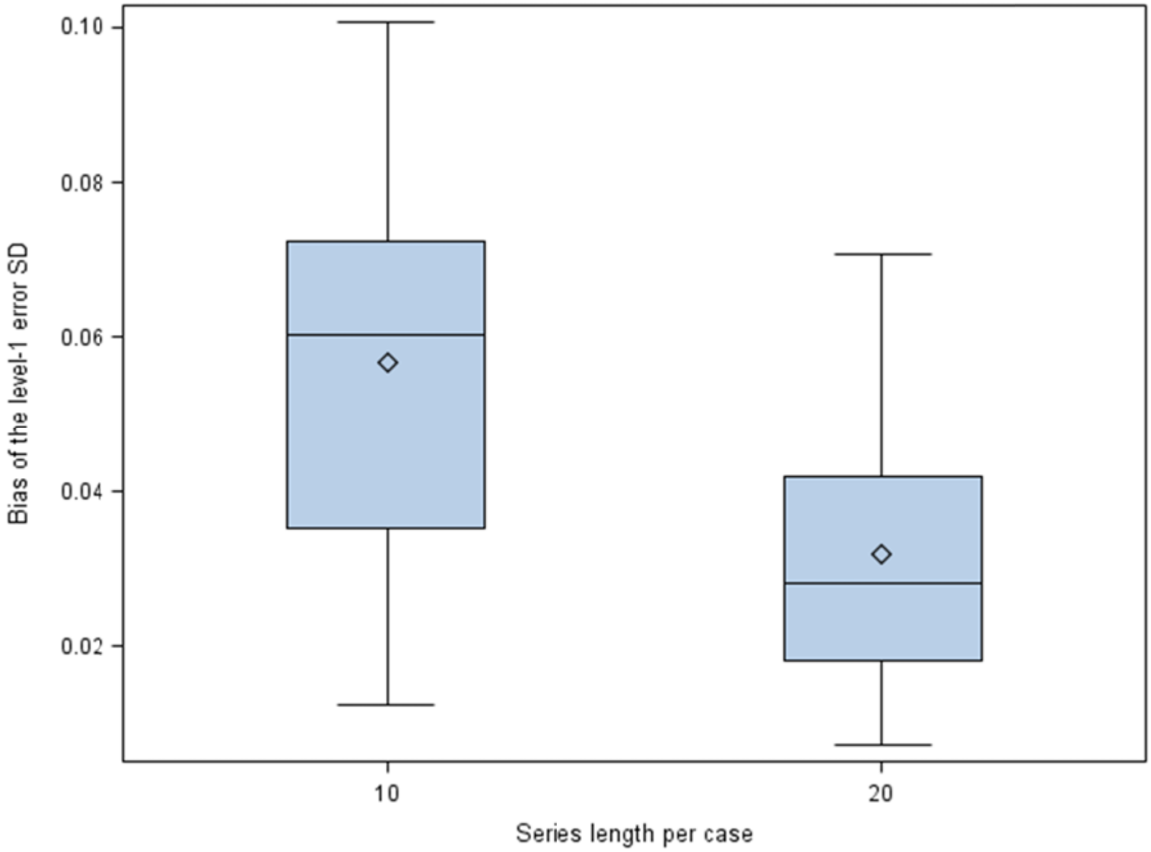


Figure 56. Box plots depicting the estimated bias of the level-1 error standard deviation as a function of the series length per case.

In order to further explore if any design factor had a significant effect on the bias for the level-1 error standard deviation, GLM models were run. The model explained 99% of the variability after including 2-way interactions, and indicated three main effects and one 2-way interaction had a medium or large effect, including the series length per case ($\eta^2 = .25$), the variation in the level-2 errors ($\eta^2 = .19$), the number of cases ($\eta^2 = .11$), and the 2-way interaction between the type of model and the true level-1 error structure ($\eta^2 = .10$). These main and interaction effects are illustrated in Figures 56 through 59.

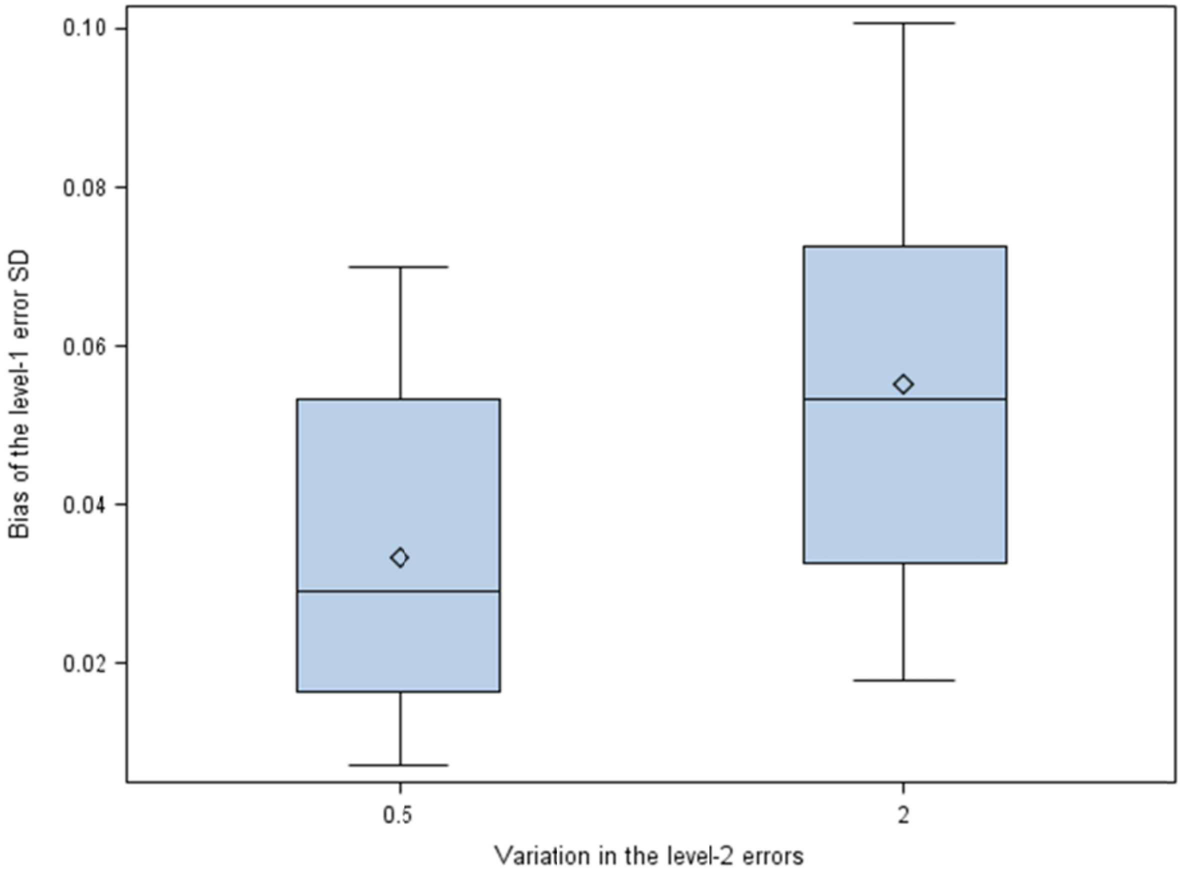


Figure 57. Box plots depicting the estimated bias of the level-1 error standard deviation as a function of the variation in the level-2 errors.

As illustrated in Figure 56, as the series length per case increased from 10 to 20, the average bias value decreased from *M* = 0.06, *SD* = 0.03 to *M* = 0.03, *SD* = 0.02. In addition, Figure 57 portrays that as the variation in the level-2 errors shifted from most of the variance at the level-1 error (.5) to most of the variance at the level-2 error (2), the average bias value increased from *M* = 0.033, *SD* = .02 to *M* = 0.055, *SD* = .03. Similarly, Figure 58 shows that as the number of cases increased from 4 to 8, the average bias value decreased from *M* = 0.053, *SD* = .03 to *M* = 0.036, *SD* = .02.

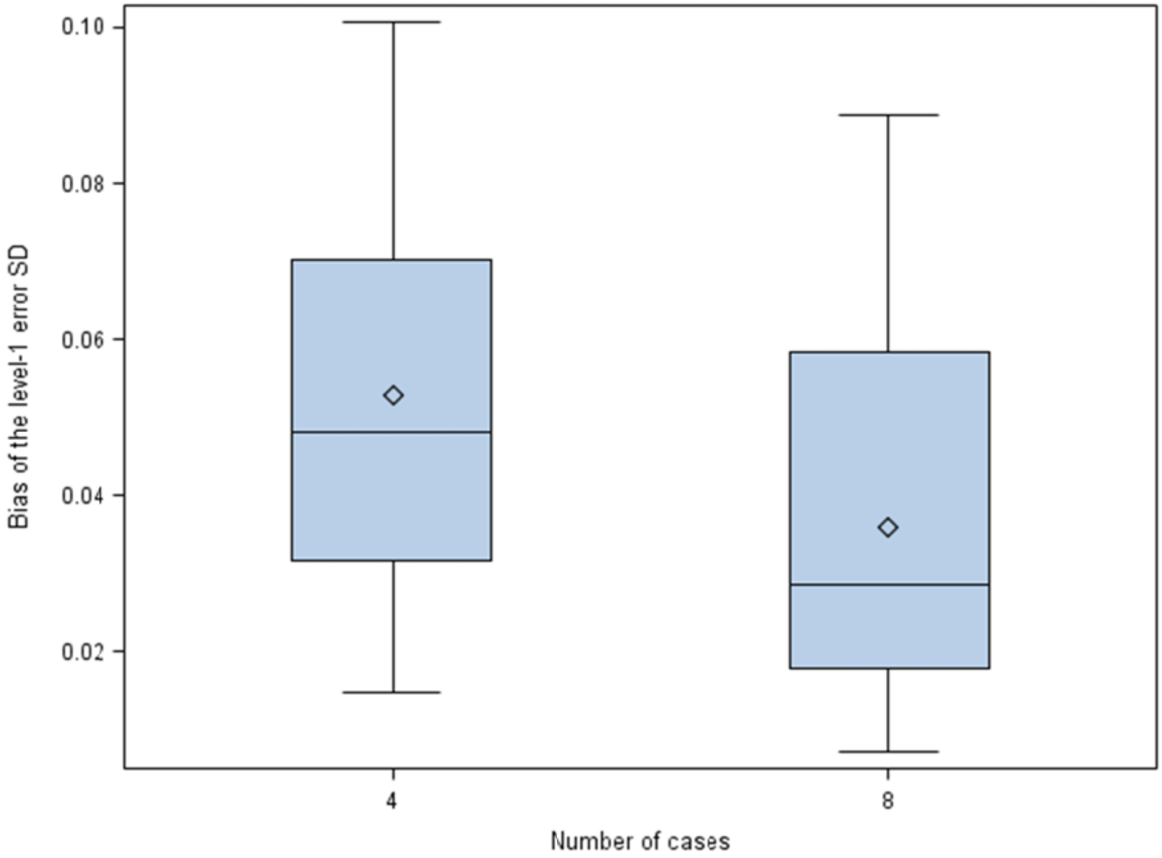


Figure 58. Box plots depicting the estimated bias of the level-1 error standard deviation as a function of the number of cases.

Figure 59 illustrated the interaction between the type of model and the true level-1 error structure on the average bias of the level-1 error standard deviation. The line graph shows that the effect of the true level-1 error structure on the mean bias was dependent on the type of model. Specifically, for Model 1, mean bias increased constantly as the true level-1 error structure shifts from homogeneous (*M* = 0.025, *SD* = 0.02) to moderately (*M* = 0.042, *SD* = 0.02) and severely (*M* = 0.071, *SD* = 0.02) heterogeneous error structure. However, for Model 2, mean bias increased as the true level-1 error structure shifts from homogeneous (*M* = 0.034, *SD* = 0.02) to moderately heterogeneous (*M* = 0.048, *SD* = 0.03) error structure, but decreased as the true level-1 error structure shifts from moderately heterogeneous to severely heterogeneous (*M* = 0.046, *SD* = 0.03).

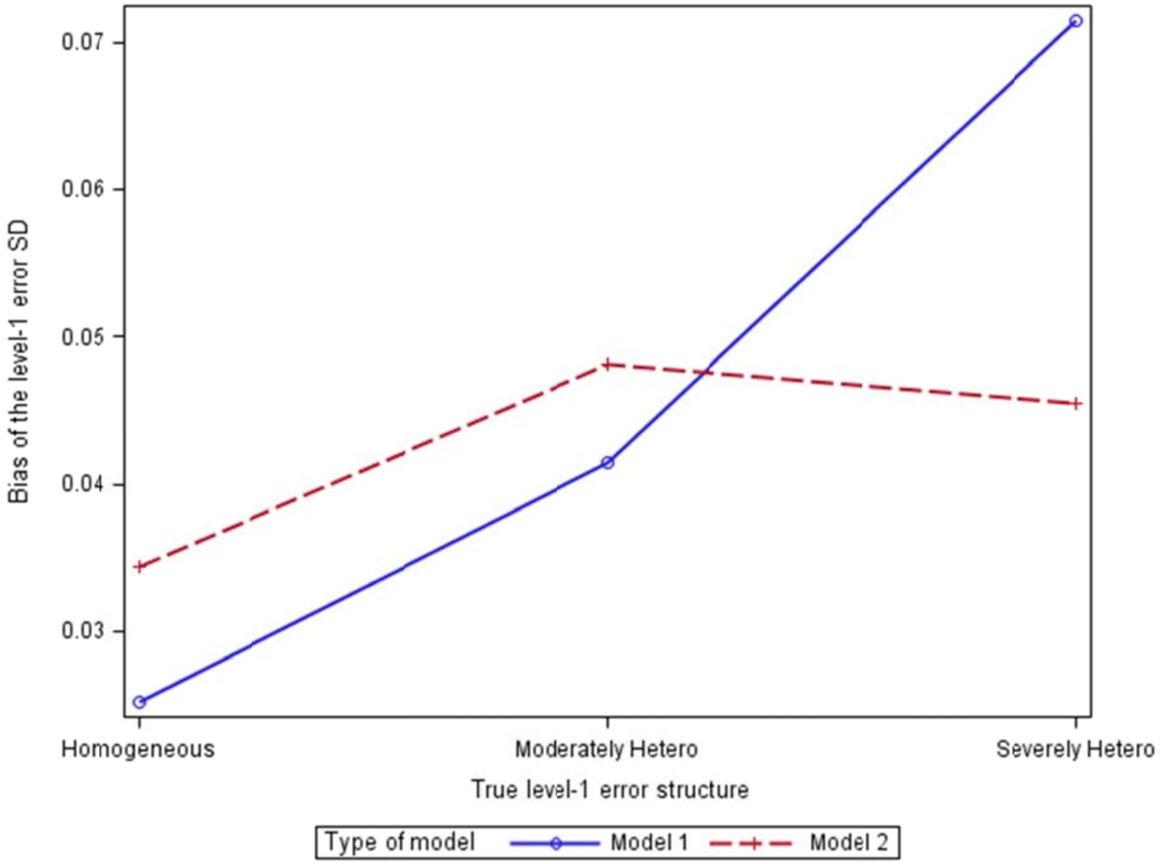


Figure 59. Line graph depicting average bias for the level-1 error standard deviation as a function of the two-way interaction effect between the type of model and the true level-1 error structure.

Autocorrelation. The average bias values of the autocorrelation were similar and negatively biased across the two models (Model 1 and Model 2) with little variability explained by the type of model ($\eta^2 = .004$) (Figure 60). The average bias value for Model 1 and Model 2 was *M* = -0.10, *SD* = 0.09 and *M* = -0.09, *SD* = 0.08, respectively.

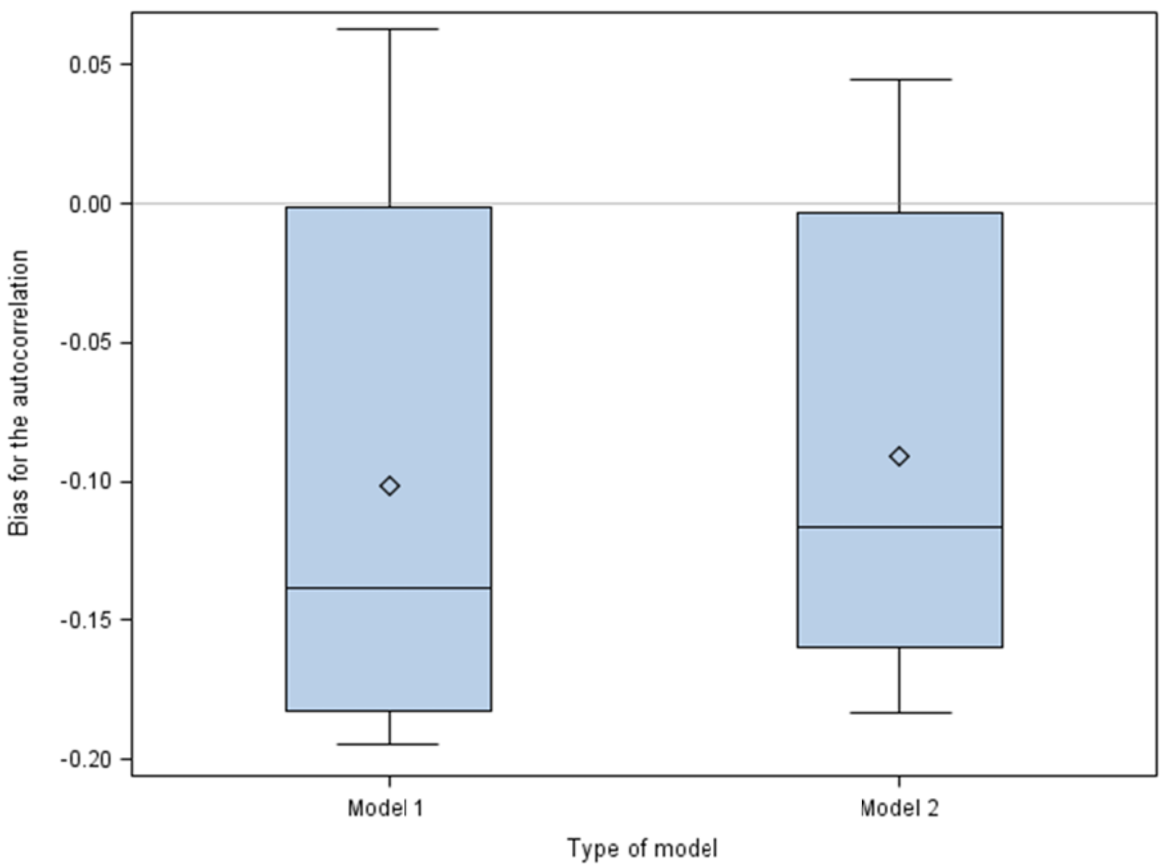


Figure 60. Box plots illustrating the distribution of the bias values for the autocorrelation across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average bias values for the autocorrelation across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 61). The figure illustrated that the average bias values were slightly different across the two models within the three true level-1 error structures. The smallest average bias difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1| = 0.008$), and the biggest average bias difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1| = 0.024$). In addition, the autocorrelation parameter tended to be more biased when estimated by Model 1 than Model 2 for all three types of the level-1 error structures.

The box plots also portrays that there were substantial differences across three different types of the true level-1 error structures, with large variability explained by the type of the true level-1 error structure ($\eta^2 = .88$). Specifically, the autocorrelation parameter tended to be more biased when the true level-1 error structure was the one of the heterogeneous error structures than homogeneous error structure, regardless of the type of model.

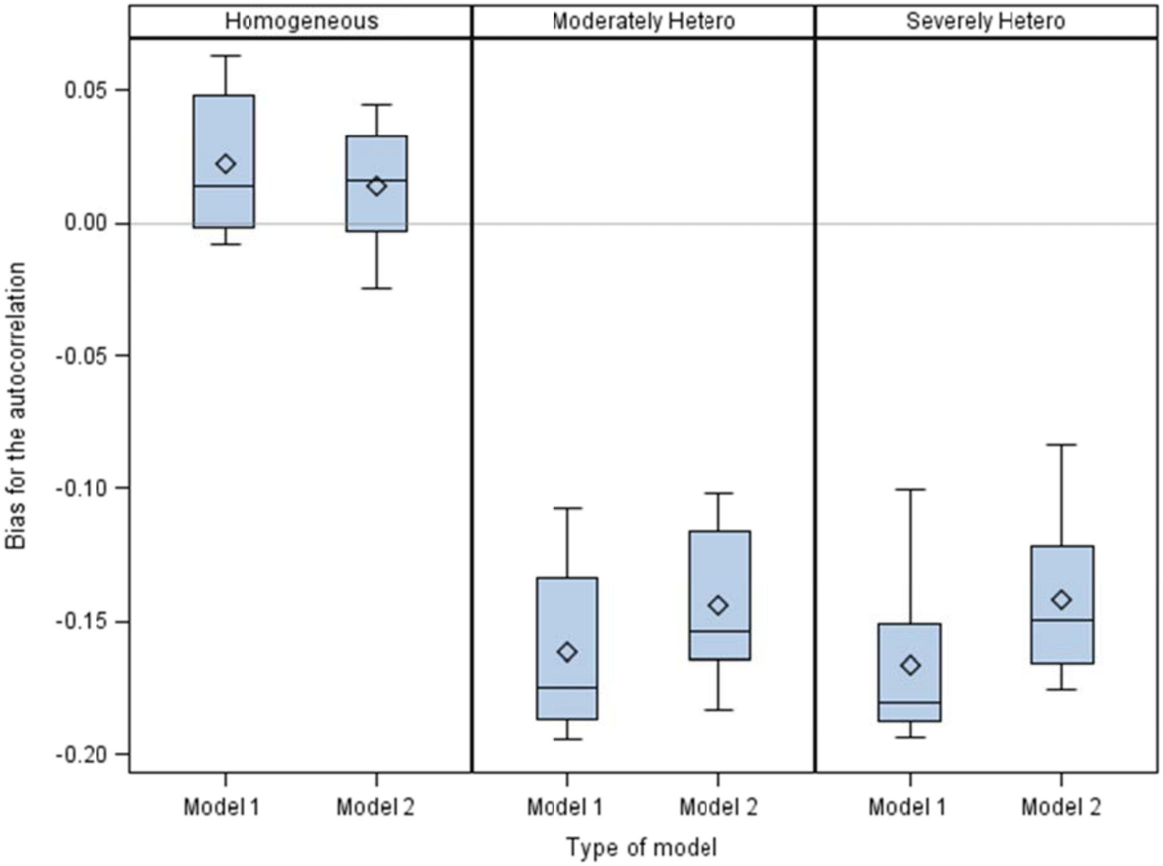


Figure 61. Box plots illustrating the distribution of the bias values for the autocorrelation across the two models within the three true level-1 error structures.

In order to further explore if any design factor had a significant effect on the bias for the autocorrelation, a GLM model was run. The main effects only model explained 96% of the variability, and indicated one of the design factors, the true level-1 error structure ($\eta^2 = .88$), had a large effect. This main effect is illustrated using box plots in Figure 62.

As illustrated in Figure 62, the autocorrelation parameter tended to be more biased when the true level-1 error structure was one of the heterogeneous error structures than homogeneous error structure. The average bias value for the homogeneous error structure was *M* = 0.018, *SD* = .03, and the average bias value for the moderately and severely heterogeneous error structures was *M* = -0.153, *SD* = .03, and *M* = -0.154, *SD* = .03, respectively.

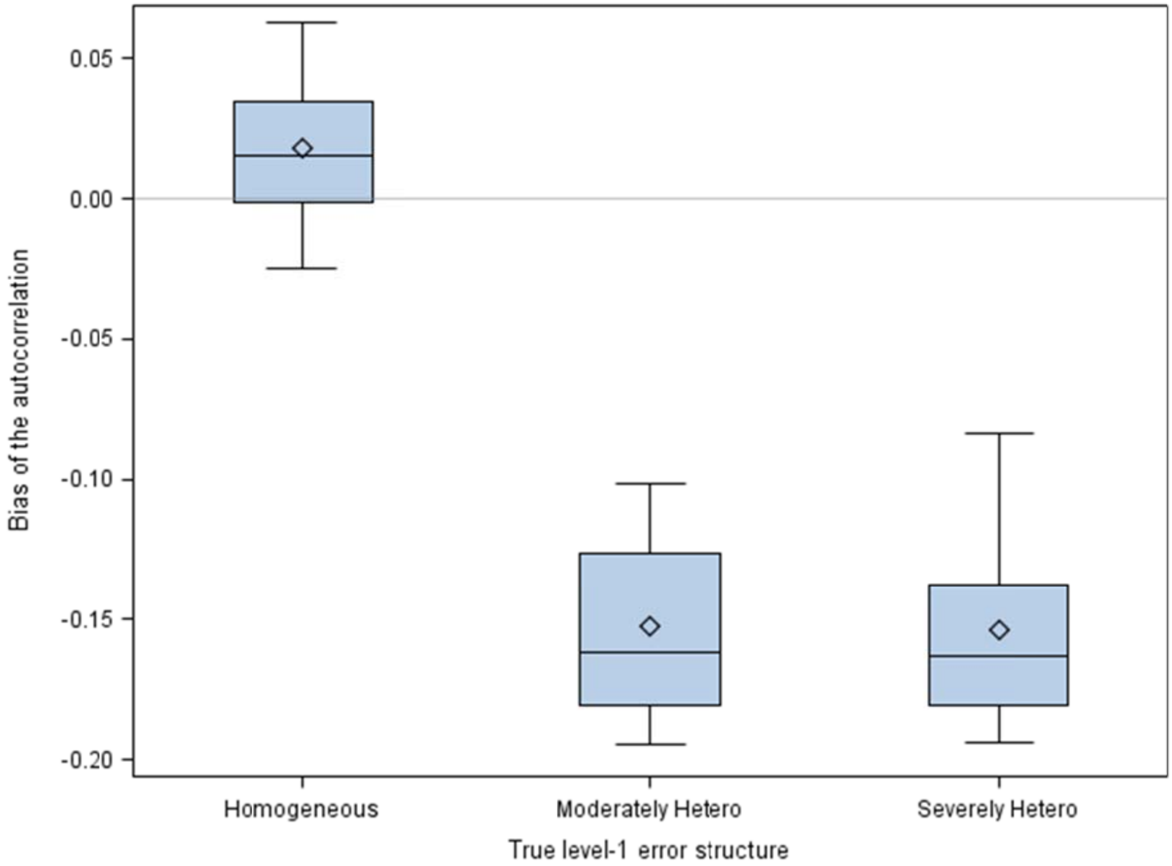


Figure 62. Box plots depicting the estimated bias of the autocorrelation as a function of the true level-1 error structure.

Root Mean Squared Error (RMSE)

The distribution of RMSE values of the level-2 error standard deviation for intervention effects (shift in level and shift in slope), the level-1 error standard deviation, and the autocorrelation are illustrated in Figures 63 through 78. The full information about the η^2 values for the GLM models is provided in Appendix B.

Level-2 error standard deviation for phase (shift in level). The average RMSE values of the level-2 error standard deviation for phase were very similar across the two models (Model 1 and Model 2) with little variability explained by the type of model (η^2 = .0002) (Figure 62). The average RMSE value for Model 1 was *M* = 1.23, *SD* = 0.68, and the average RMSE value for Model 2 was *M* = 1.21, *SD* = 0.67.

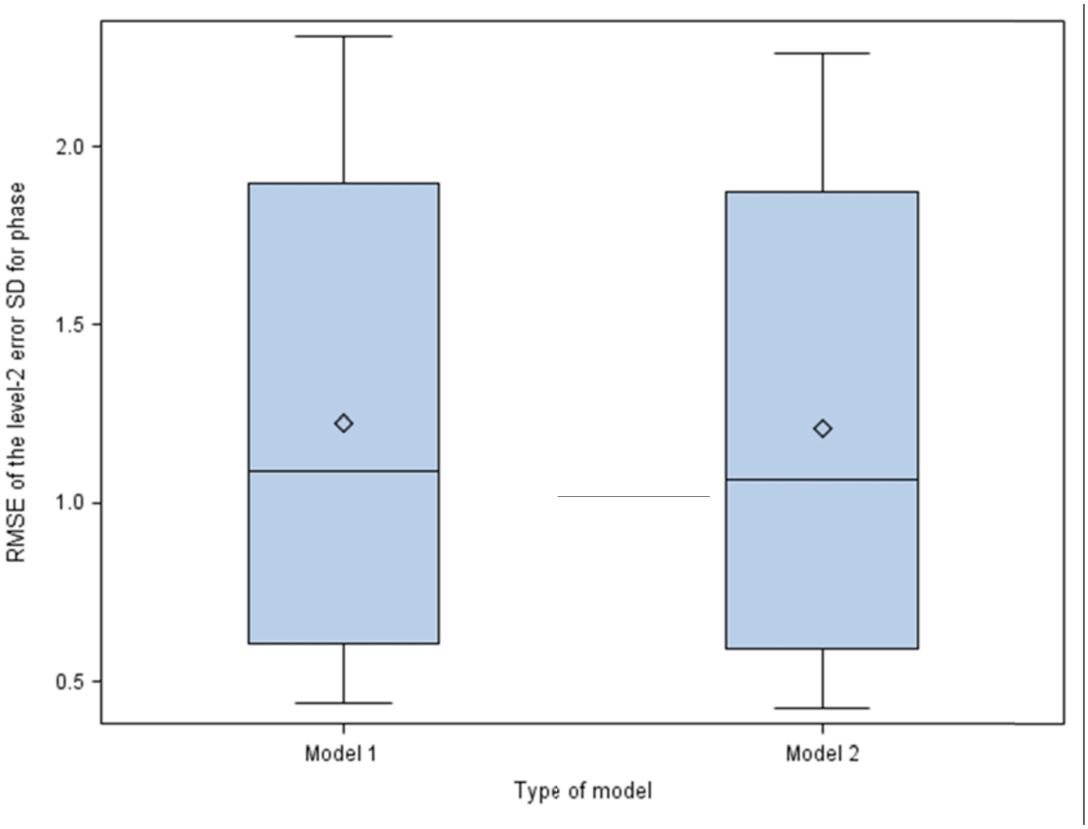


Figure 63. Box plots illustrating the distribution of the RMSE values for the level-2 error standard deviation of the shift in level across Model 1 which did not model between case variation, and Model 2 which models between case variation.

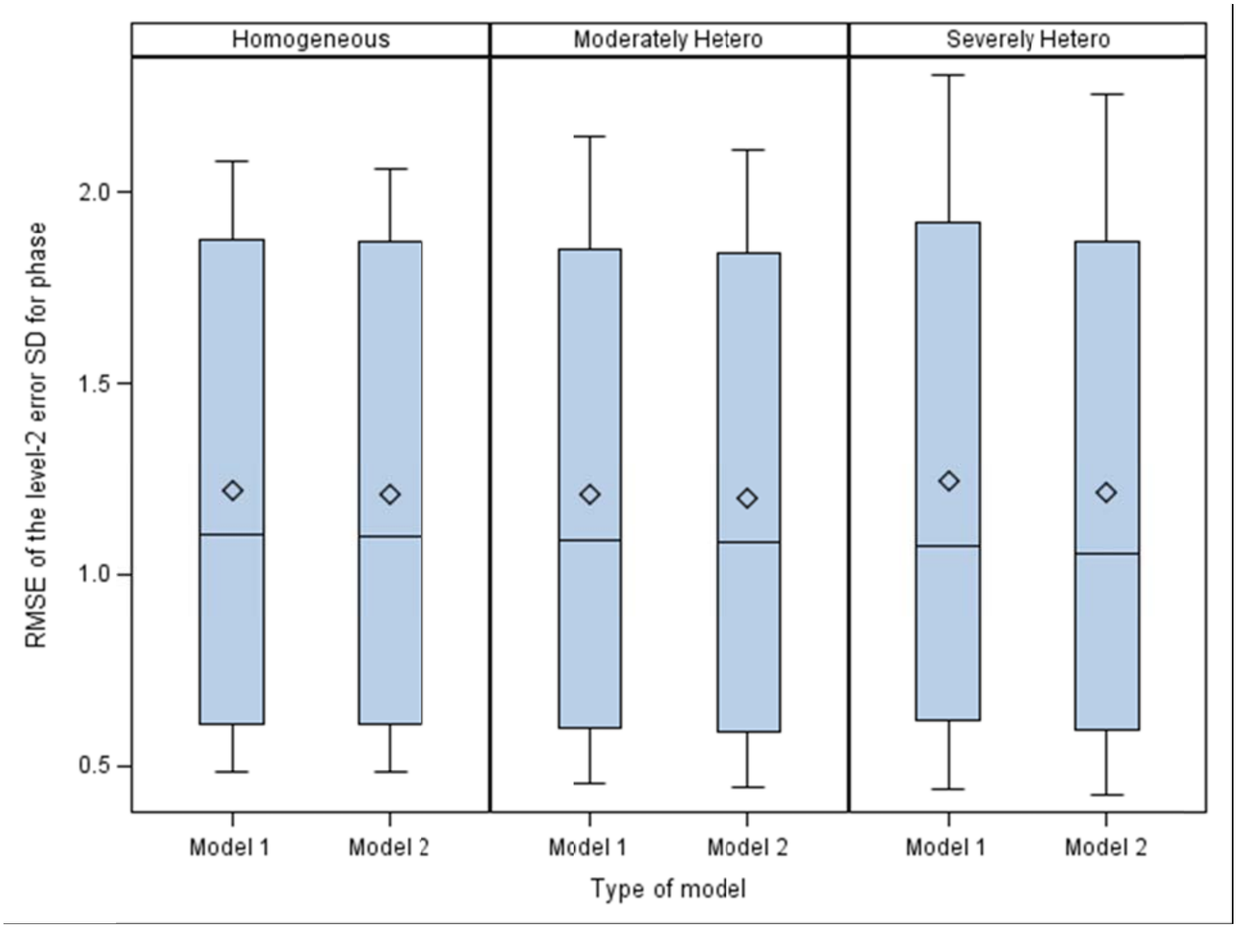


Figure 64. Box plots illustrating the distribution of the RMSE values for the level-2 error standard deviation of shift in level across the two models within the three true level-1 error structures.

The average RMSE values of the level-2 error variance for phase across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 64). The average RMSE values were very similar across the two models within the three true level-1 error structures, and there were little differences across the true level-1 error structures, with very little of the variability explained by the different types of the true level-1 error structures (η^2 = .0003). However, the average RMSE value for Model 1 was larger than the average RMSE value for Model 2 for all three true level-1 error structures. The smallest average RMSE difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2$-$M_1| = 0.005$), and the biggest average RMSE difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2$-$M_1| = 0.032$).

In order to further explore if any design factor had a significant effect on the RMSE of the level-2 error standard deviation for the shift in level, a GLM model was run. The main effects only model explained 98% of the variability and found that two of the design factors had a medium or large effect, including the number of cases ($\eta^2 = .89$), and the variation in the level-2 errors ($\eta^2 = .08$). These two main effects are illustrated in Figures 65 and 66.

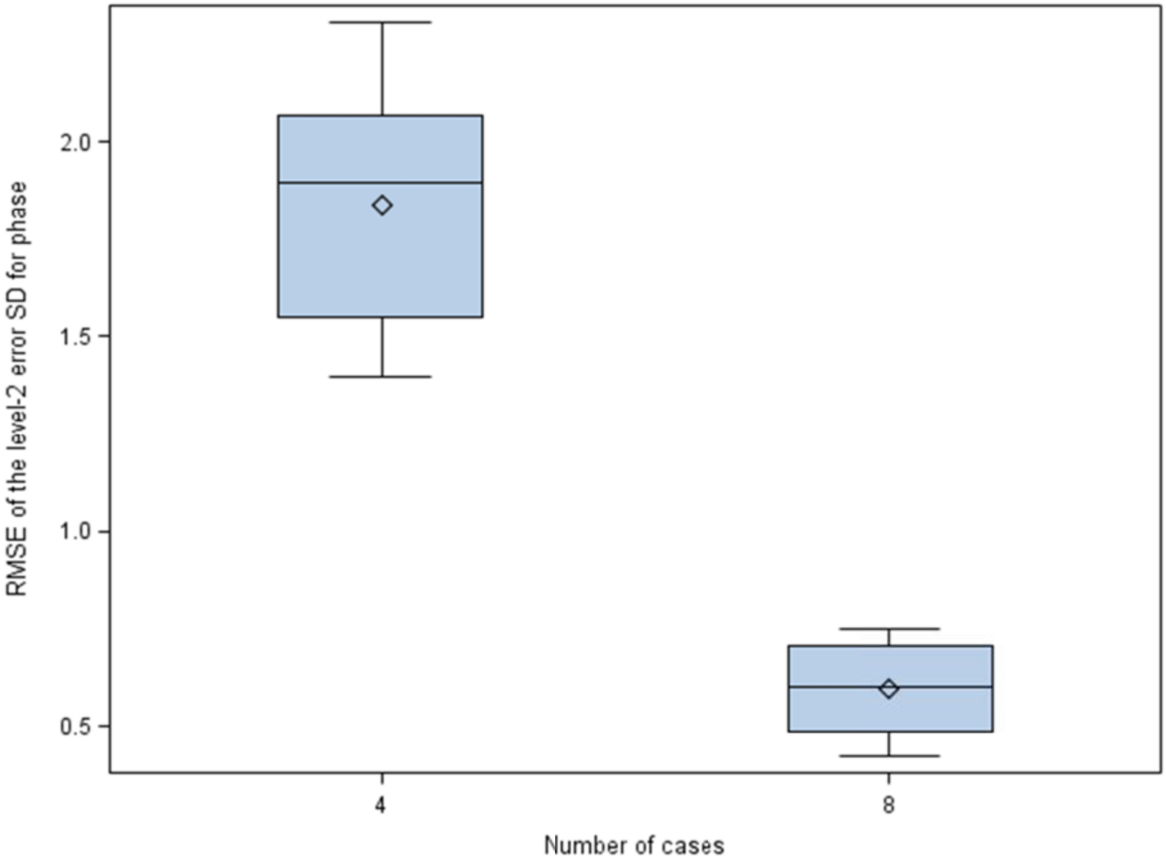


Figure 65. Box plots depicting the estimated RMSE values of the level-2 error standard deviation for the shift in level as a function of the number of cases.

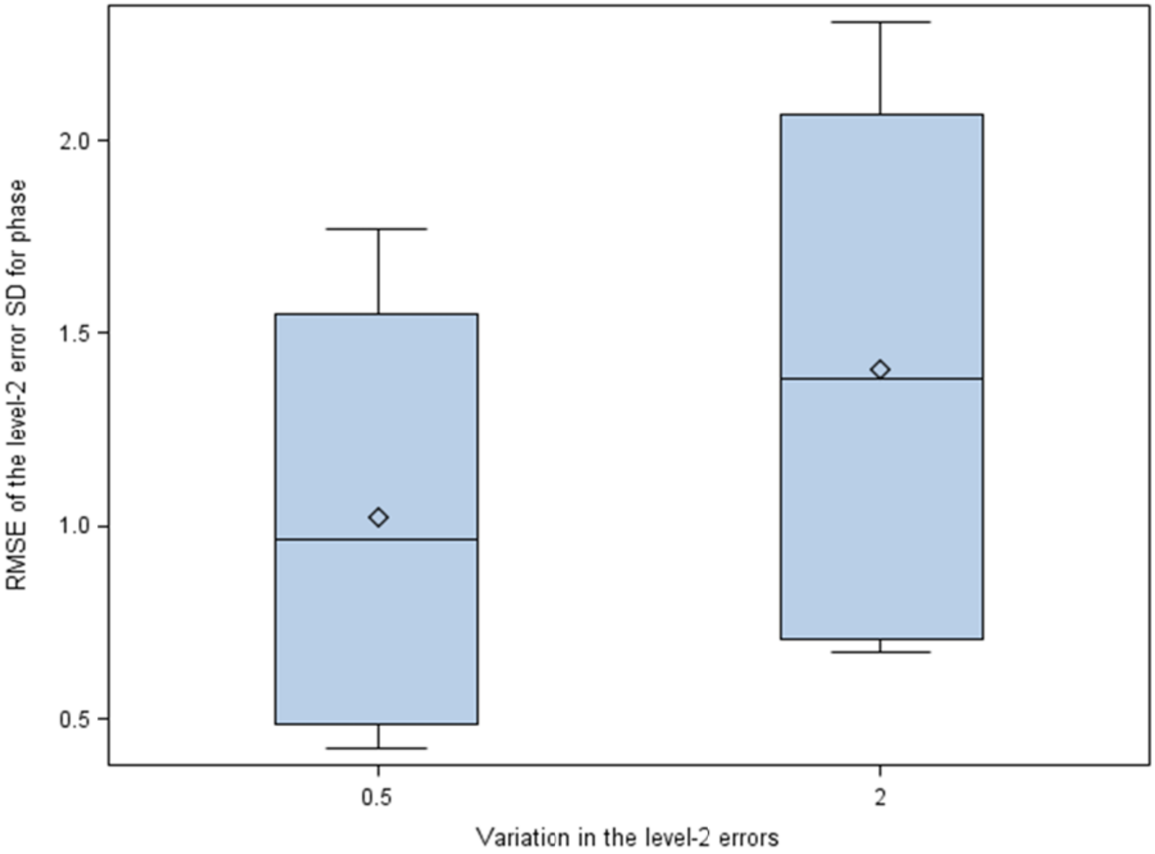


Figure 66. Box plots depicting the estimated RMSE values of the level-2 error standard deviation of the shift in level as a function of the variation in the level-2 errors.

As illustrated in Figure 65, as the number of cases increased from 4 to 8, the average RMSE values decreased from *M* = 1.84, *SD* = .30 to *M* = 0.60, *SD* = .12. Similarly, Figure 65 shows that as the variation in the level-2 errors shifted from most of the variance at the level-1 error (.5) to most of the variance at the level-2 error (2), the average RMSE values increased from *M* = 1.03, *SD* = .56 to *M* = 1.41, *SD* = .71.

Level-2 error standard deviation for interaction (shift in slope). The average RMSE values of the level-2 error standard deviation for the interaction effect were very similar across the two models (Model 1 and Model 2) with little variability explained by the type of model (η^2

= .0003) (Figure 67). The average RMSE value for Model 1 was *M* = 0.42, *SD* = .27, and the average RMSE for Model 2 was *M* = 0.41, *SD* = .27.

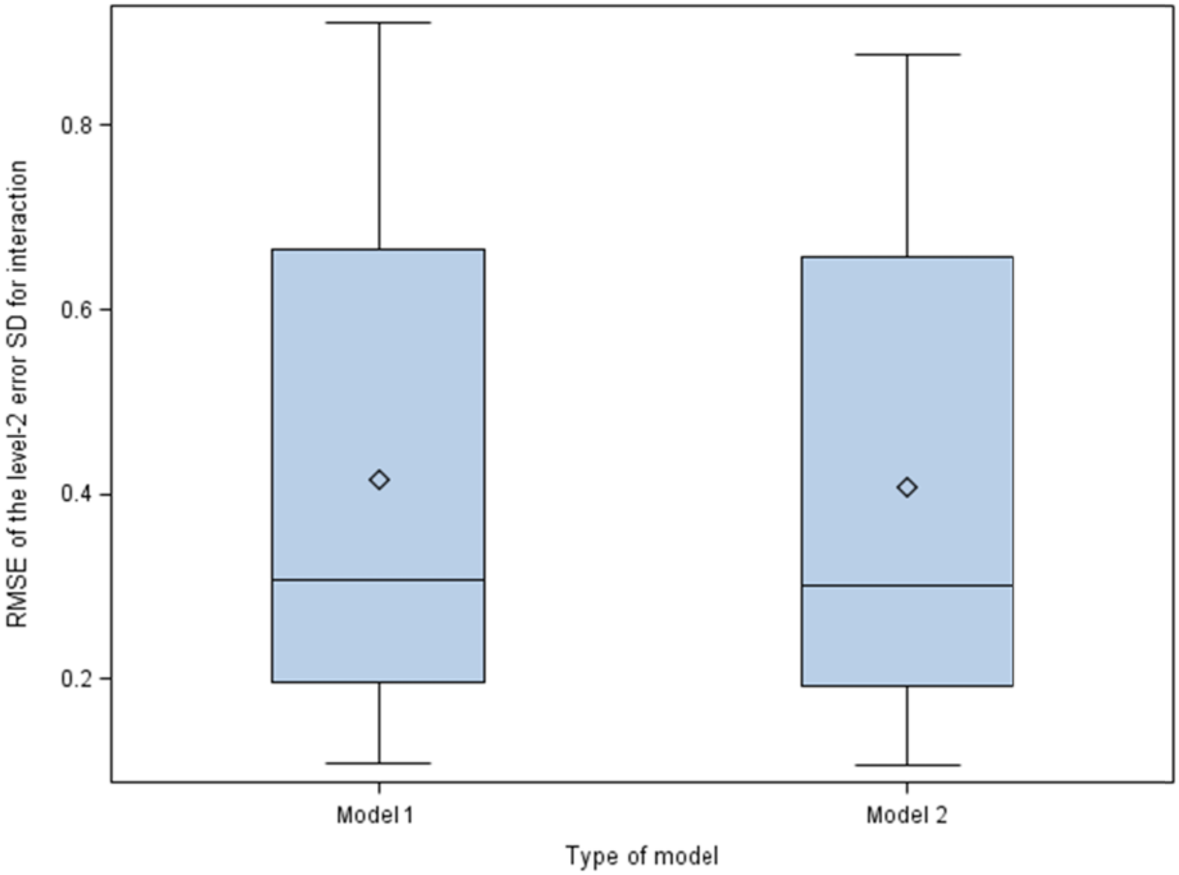


Figure 67. Box plots illustrating the distribution of the RMSE values for the level-2 error standard deviation for shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average RMSE values of the level-2 error standard deviation for the interaction effect across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 68). The average RMSE values were very similar across the two models within the three true level-1 error structures, and there was little difference across the true level-1 error structures, with very little of the variability explained by the different types of the true level-1 error structures (η^2 = .0002). Similar to the result of the level-2 error standard deviation for phase, the average RMSE value for Model 1 was larger than the average RMSE value for Model 2 for the all three true level-1 error structures. The smallest average RMSE difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1|$ = 0.004), and the biggest average RMSE difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1|$ = 0.016).

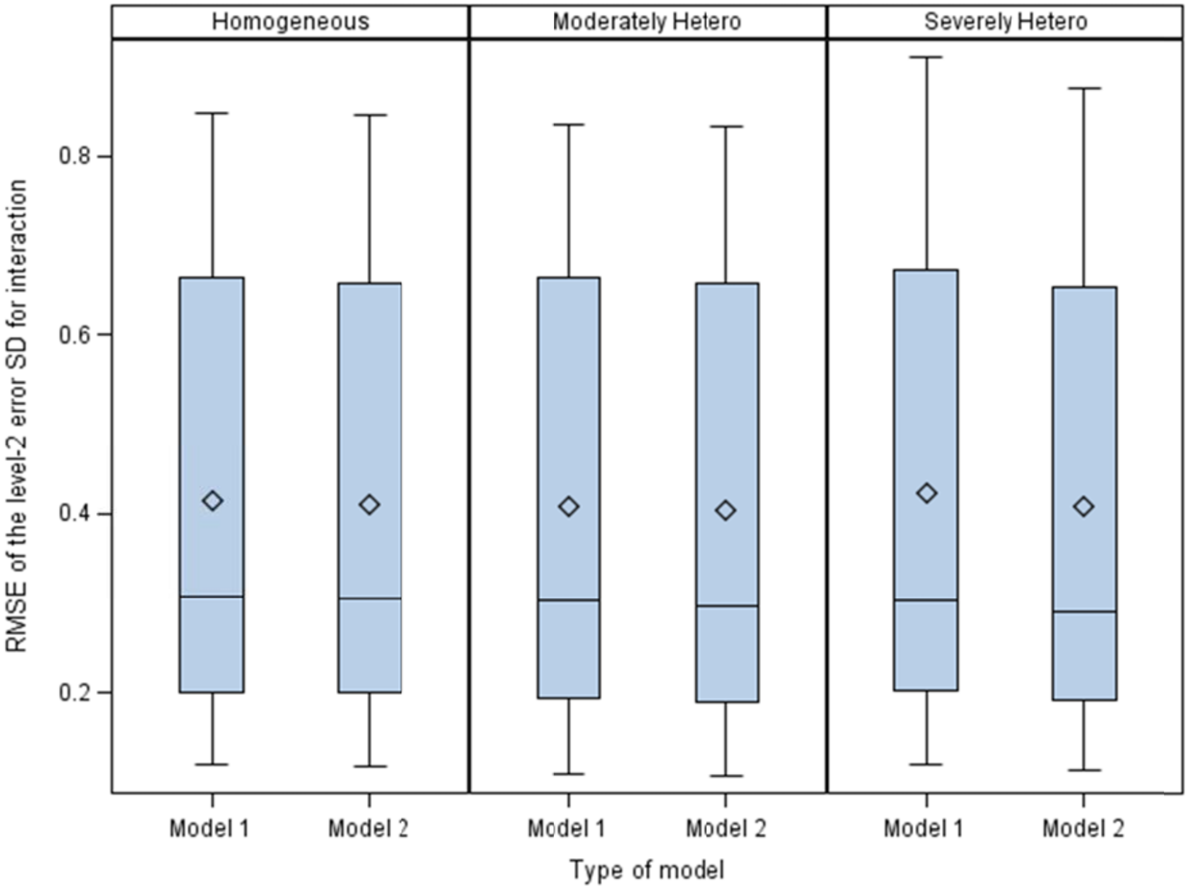


Figure 68. Box plots illustrating the distribution of the RMSE values for the level-2 error standard deviation for shift in slope across the two models within the three true level-1 error structures.

In order to further explore if any design factor had a significant effect on the RMSE values of the level-2 error standard deviation for shift in slope, GLM models were run. The model explained over 99% of the variability after including 2-way interactions. The model indicated that two of the design factors had a medium or large effect, including the number of cases ($\eta^2 = .73$) and the series length per case ($\eta^2 = .13$). These two main effects are illustrated in Figures 69 and 70.

As illustrated in Figure 69, as the number of cases increased from 4 to 8, the average RMSE values decreased from *M* = 0.64, *SD* = .19 to *M* = 0.19, *SD* = .05. Similarly, Figure 70 shows that as the series length per case increased from 10 to 20, the average RMSE values decreased from *M* = 0.51, *SD* = .30 to *M* = 0.32, *SD* = .19.

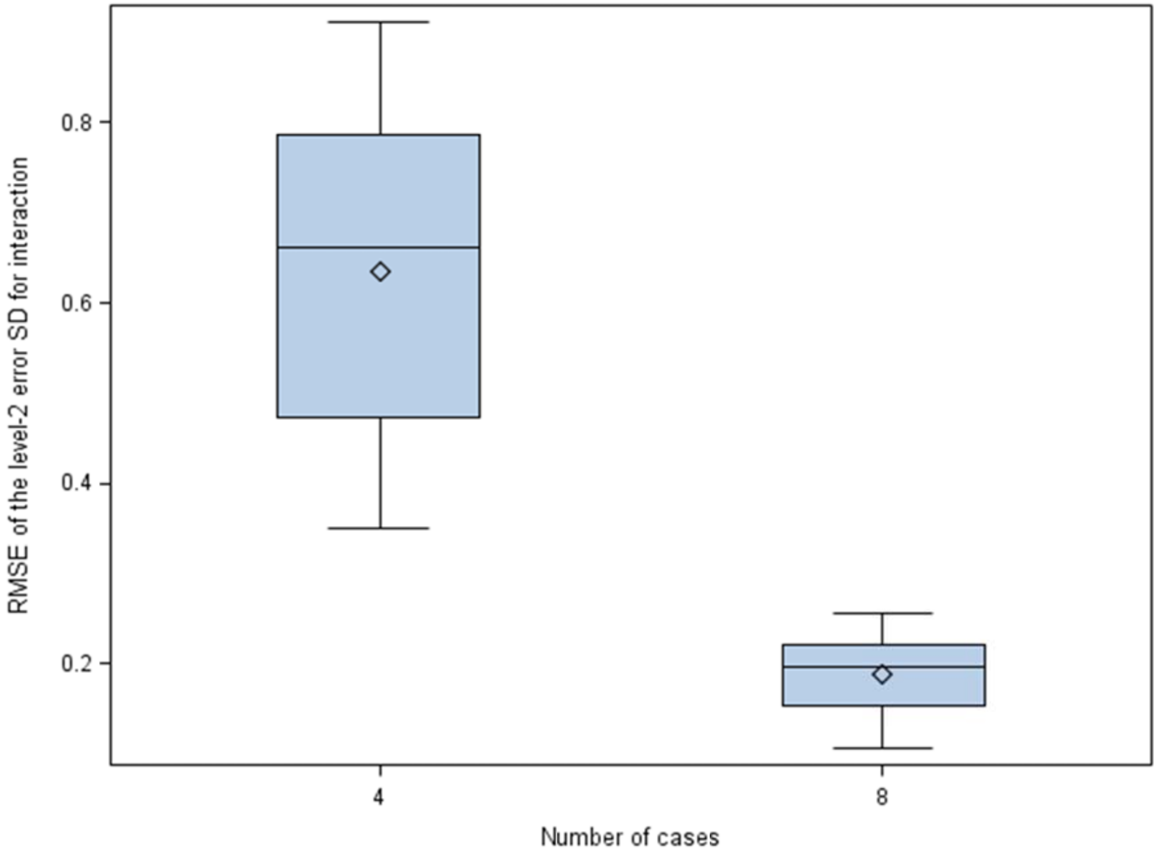


Figure 69. Box plots depicting the estimated RMSE values of the level-2 error standard deviation for shift in slope as a function of the number of cases.

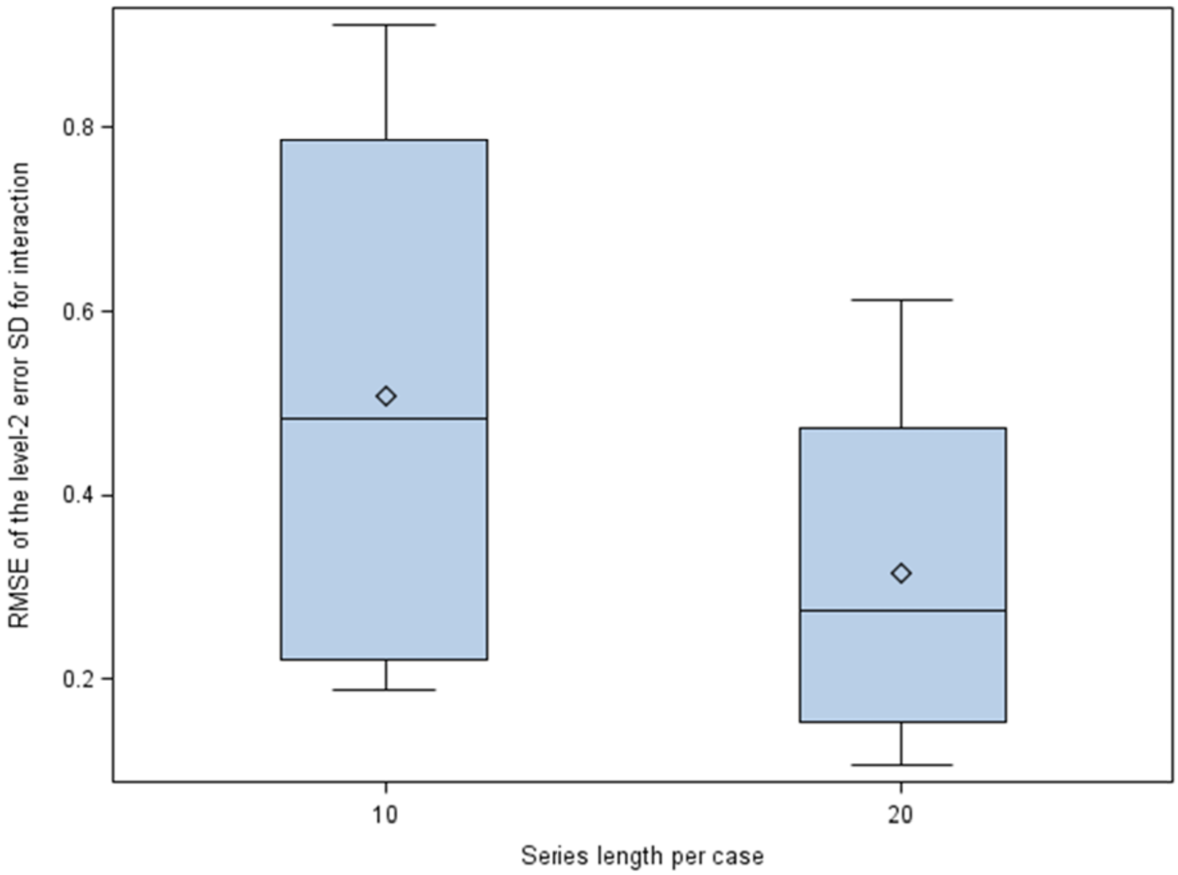


Figure 70. Box plots depicting the estimated RMSE values of the level-2 error standard deviation for shift in slope as a function of the series length per case.

Level-1 error standard deviation. The average RMSE values of the level-1 error standard deviation were different across the two models (Model 1 and Model 2) with little variability explained by the type of model ($\eta^2 = .05$) (Figure 71). The average RMSE value for Model 1 was bigger than the average RMSE for Model 2, and Model 1 had more variability of the RMSE values than Model 2. The average RMSE value for Model 1 and Model 2 was *M* = 0.22, *SD* = .10 and *M* = 0.18, *SD* = 0.05, respectively.

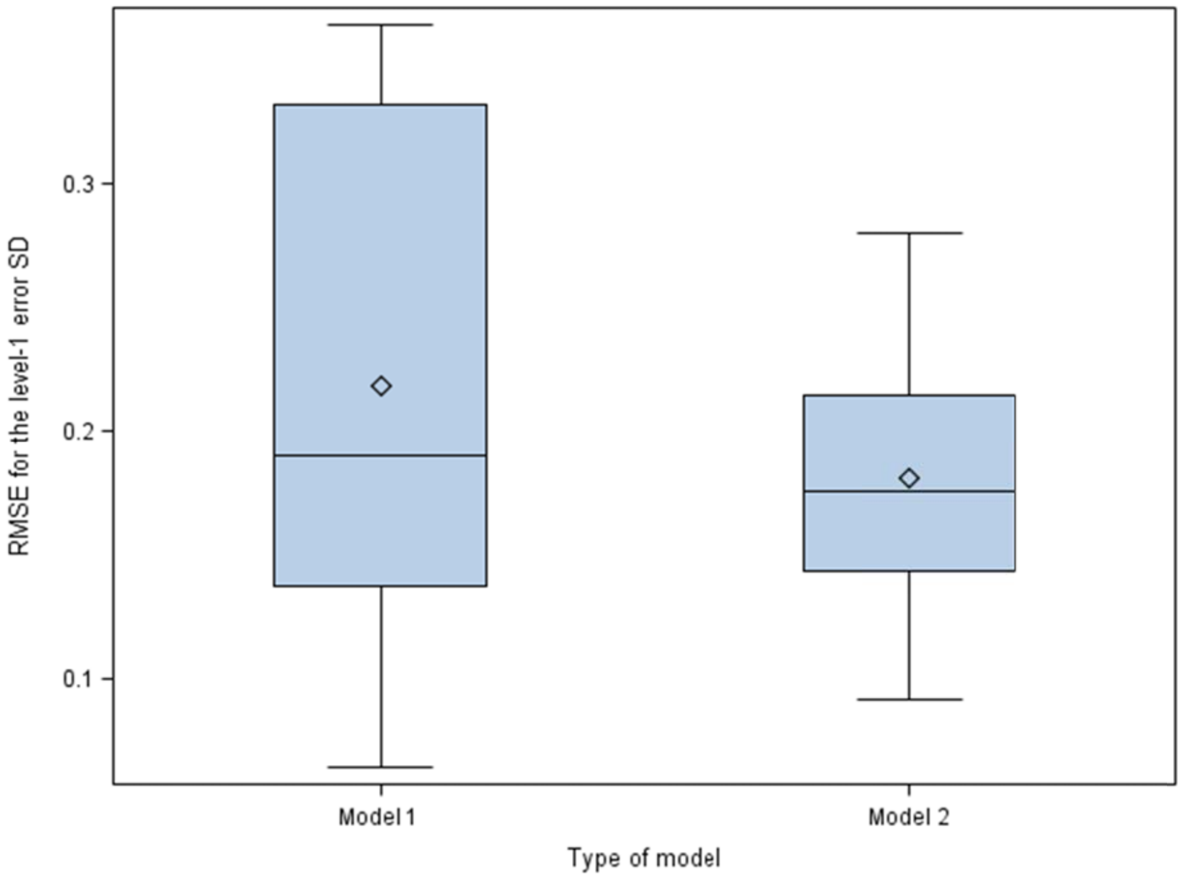


Figure 71. Box plots illustrating the distribution of the RMSE values for the level-1 error standard deviation across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average RMSE values for the level-1 error standard deviation across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 72). The figure illustrated that the average RMSE values were different across the two models within the three true level-1 error structures, with large variability explained by the different types of the true level-1 error structures ($\eta^2 = .62$). Specifically, the average RMSE value of the level-1 error standard deviation tended to be larger when estimated by Model 2 than Model 1, and when the true level-1 error structure was homogeneous. However, the average RMSE value of the level-1 error standard deviation tended to be smaller when estimated by Model 2 than Model 1, and when the true level-1 error structure was one of the heterogeneous error structures. In addition, the smallest average RMSE difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2-M_1| = 0.02$), and the biggest average RMSE difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2-M_1| = 0.13$).

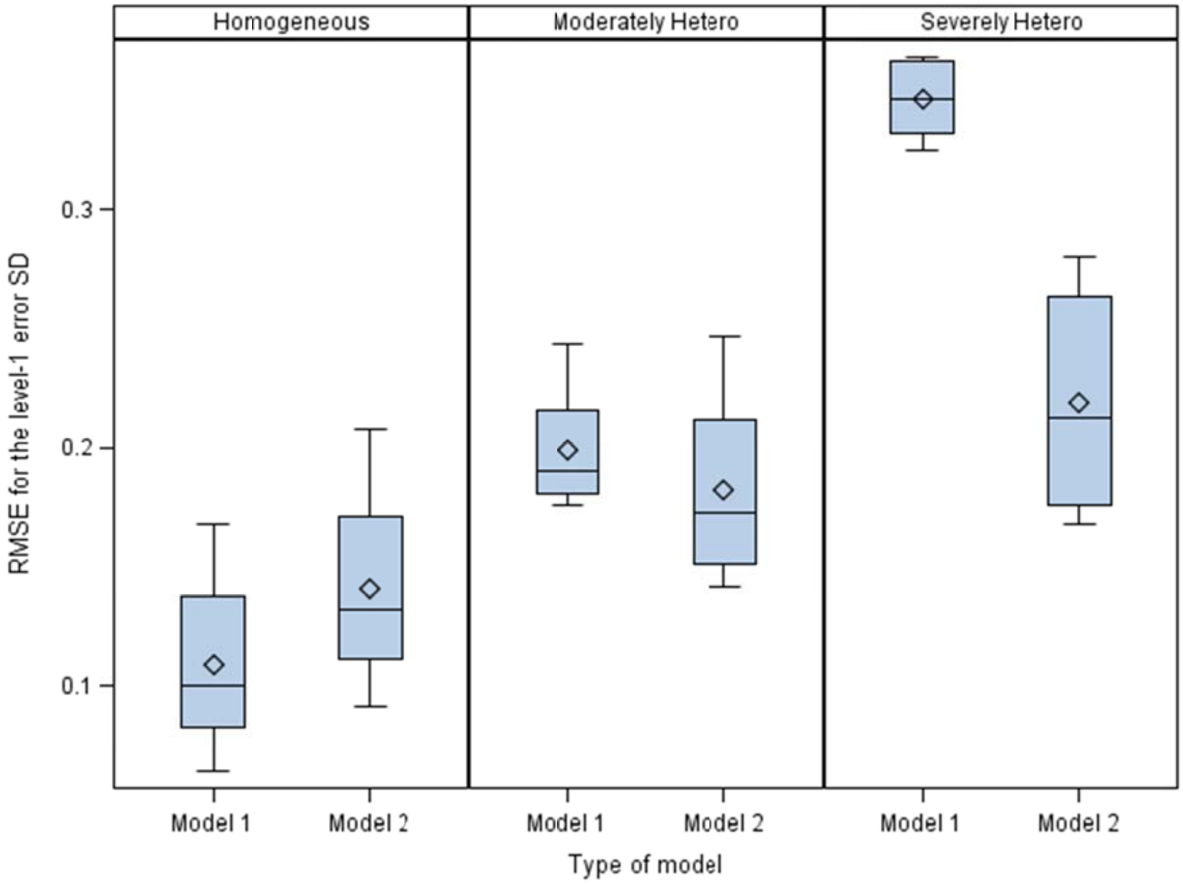


Figure 72. Box plots illustrating the distribution of RMSE of the level-1 error standard deviation across the two models within the three true level-1 error structures.

In order to further explore if any design factor had a significant effect on the RMSE value of the level-1 error standard deviation, GLM models were run. The model explained 99% of the variability after including 2-way interactions, and indicated one main effect and one interaction effect had a medium or large effect, including the series length per case ($\eta^2 = .11$) and the 2-way interaction between the type of model and the true level-1 error structure ($\eta^2 = .16$). These main and interaction effects are illustrated in Figures 73 and 74.

As illustrated in Figure 73, as the series length per case increased from 10 to 20, the average RMSE value decreased from *M* = 0.23, *SD* = 0.08 to *M* = 0.17, *SD* = 0.08.

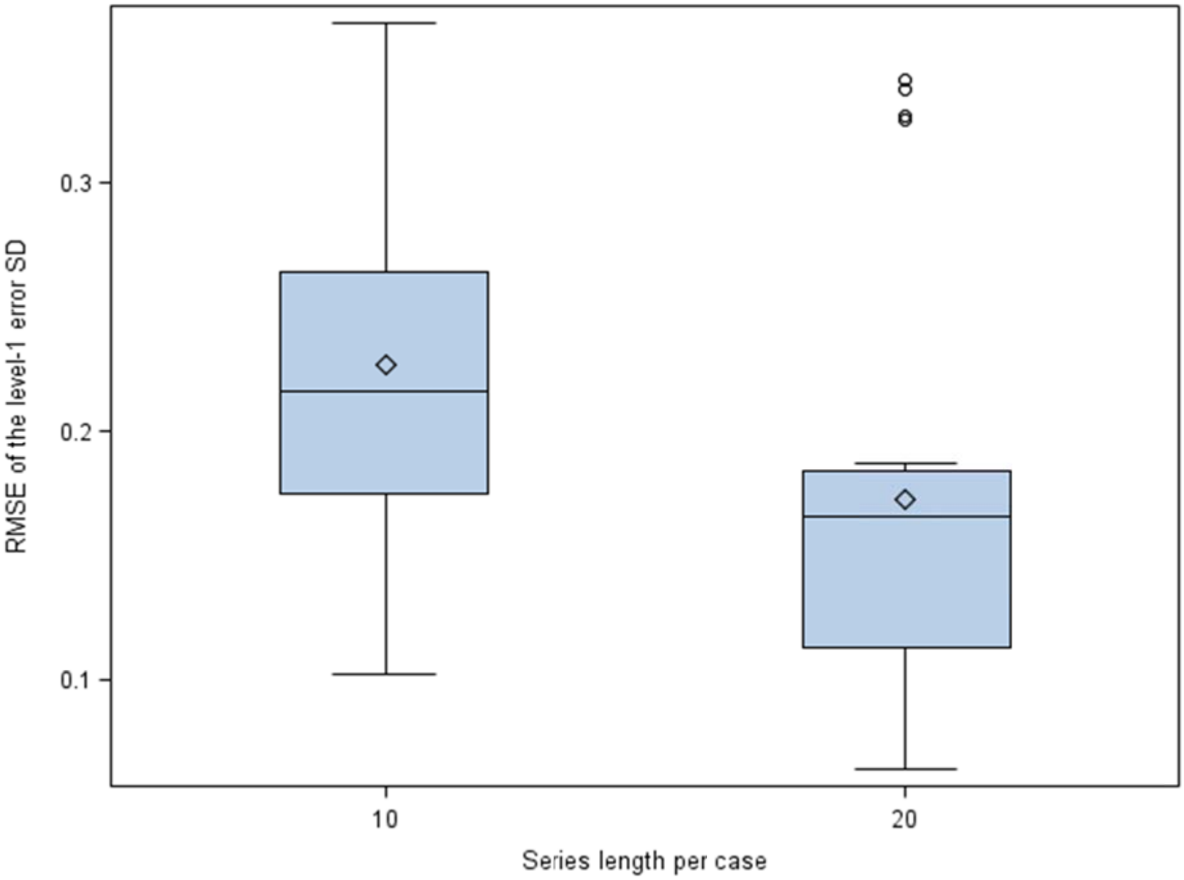


Figure 73. Box plots depicting the estimated RMSE values of the level-1 error standard deviation as a function of the series length per case.

Figure 74 illustrates the interaction between the type of model and the type of true level-1 error structure on the average RMSE value of the level-1 error standard deviation. The line graph shows that the effect of the true level-1 error structure on the mean RMSE value was dependent on the type of model. Specifically, mean RMSE increased constantly as the true level-1 error structure shifts from homogeneous (*M* = 0.025, *SD* = 0.02) to moderately (*M* = 0.042, *SD* = 0.02) and severely (*M* = 0.071, *SD* = 0.02) heterogeneous error structure for both Model 1 and Model 2. However, mean RMSE increased more rapidly for Model 1 than Model 2. Model 1 started with a smaller mean RMSE value than Model 2 when the true level-1 error structure was homogeneous (*M* = 0.11, *SD* = 0.04; *M* = 0.14, *SD* = 0.04, respectively). However, the mean RMSE value for Model 1 became larger than the mean RMSE value for Model 2 as the true level-1 error structure shifts to moderately (*M* = 0.20, *SD* = 0.02; *M* = 0.18, *SD* = 0.04, respectively) and severely (*M* = 0.35, *SD* = 0.02; *M* = 0.22, *SD* = 0.05, respectively) heterogeneous error structure.

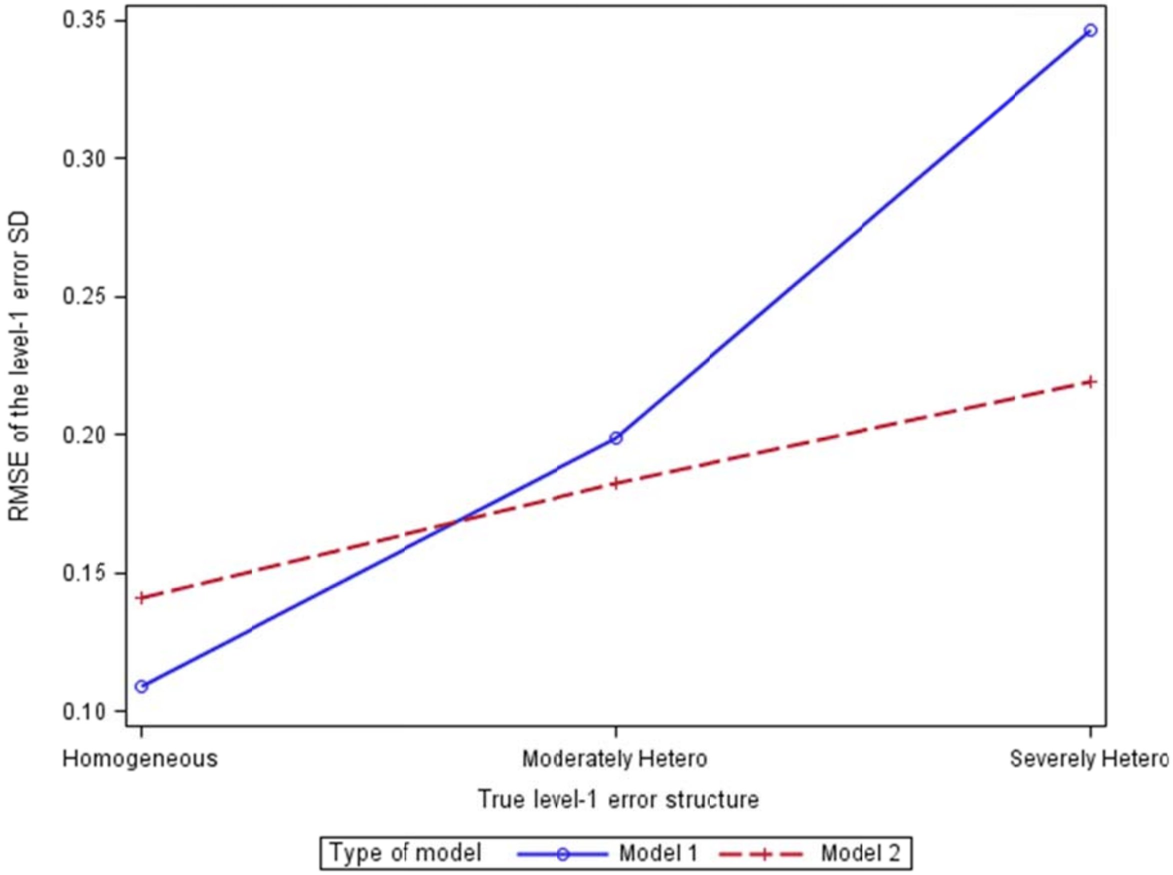


Figure 74. Line graph depicting average RMSE for the level-1 error variance as a function of the two-way interaction effect between the type of model and the true level-1 error structure.

Autocorrelation. The average RMSE values of the autocorrelation were similar across the two models (Model 1 and Model 2) with little variability explained by the type of model ($\eta^2 = .002$) (Figure 75). The average RMSE value for Model 1 and Model 2 was *M* = 0.26, *SD* = 0.07 and *M* = 0.25, *SD* = 0.06, respectively.

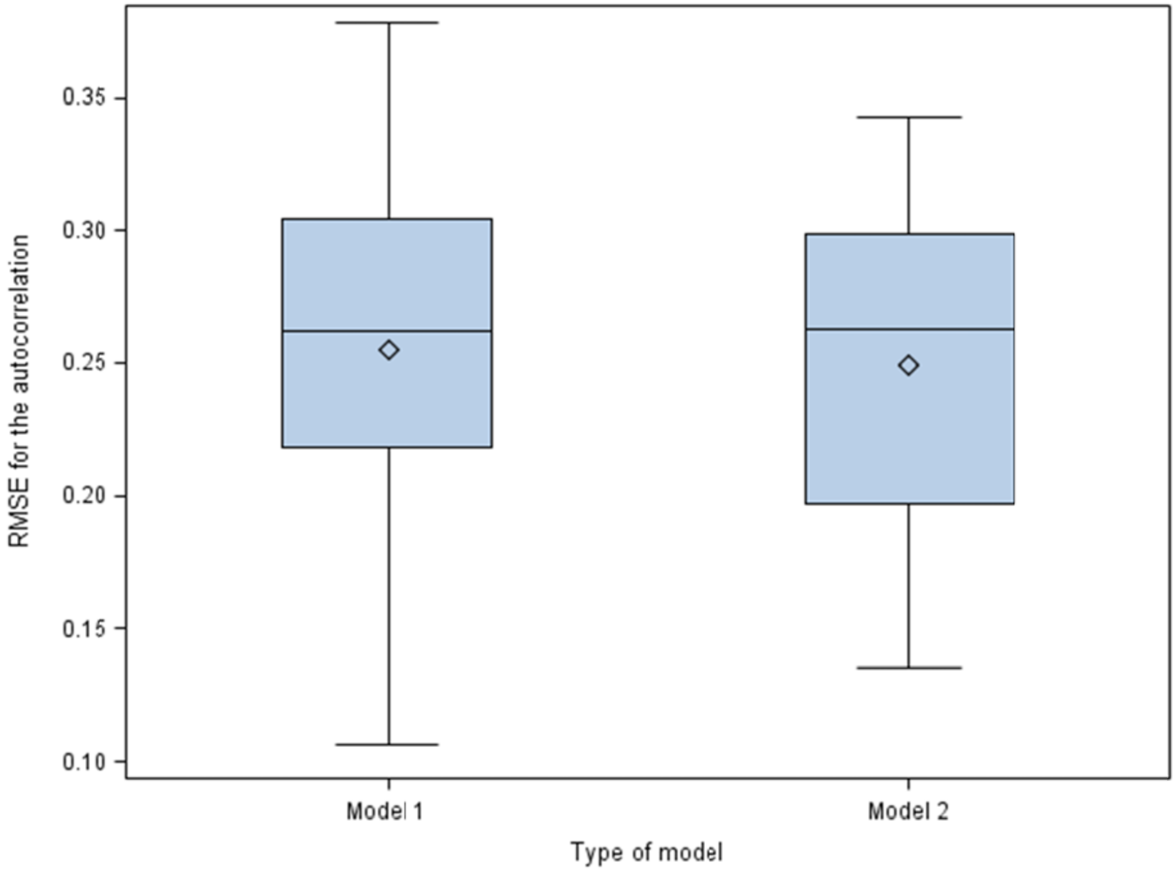


Figure 75. Box plots illustrating the distribution for the RMSE values for the autocorrelation across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average RMSE values for the autocorrelation across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 76). The figure illustrated that the average RMSE values were slightly different across the two models within the three true level-1 error structures. The smallest average RMSE difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1| = 0.004$), and the biggest average RMSE difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1| = 0.016$). In addition, the average RMSE value of the autocorrelation tended to be larger when estimated by Model 2 than Model 1, and when the true level-1 error structure was homogeneous. On the contrary, the average RMSE value of the autocorrelation tended to be smaller when estimated by Model 2 than Model 1, and when the true level-1 error structure was one of the heterogeneous error structures. Moreover, Model 1 had more variability of the RMSE values than Model 2.

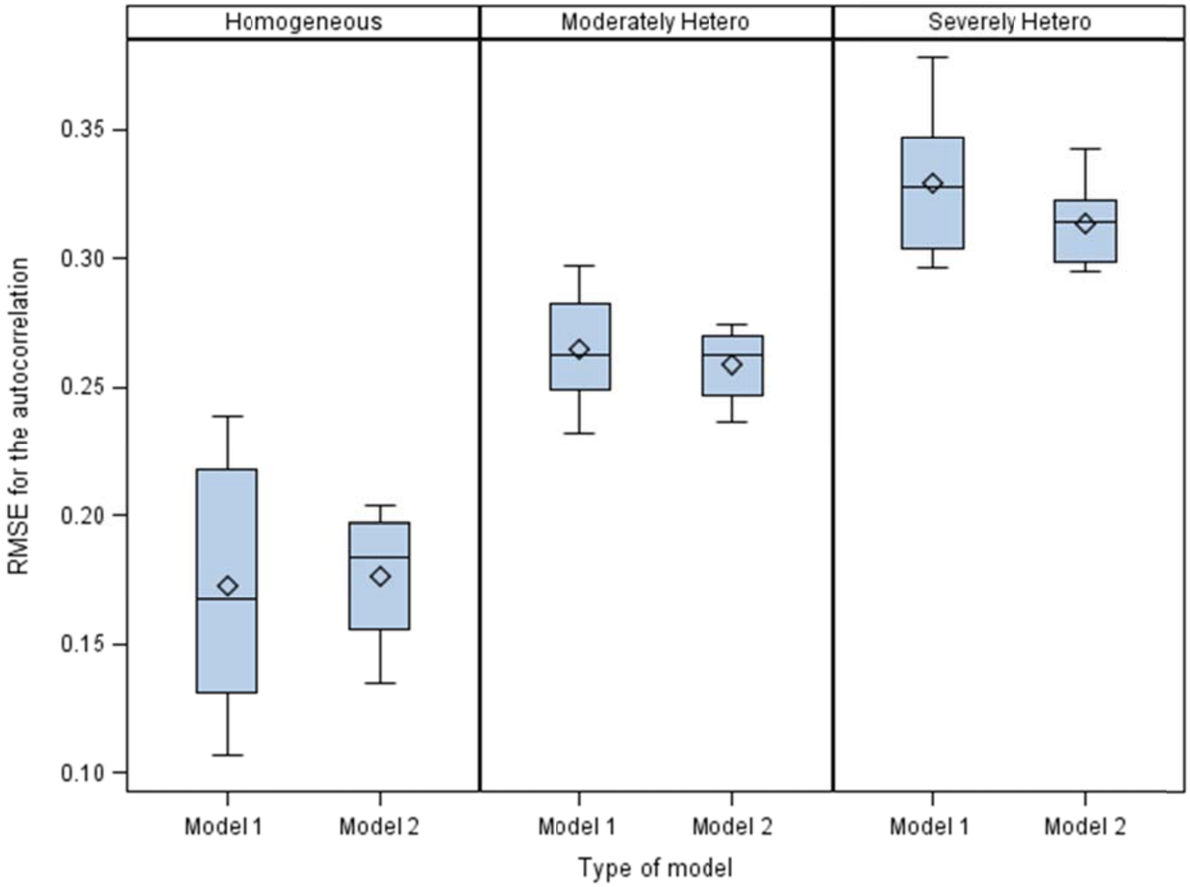


Figure 76. Box plots illustrating the distribution of the RMSE values for the autocorrelation across the two models within the three true level-1 error structures.

Figure 76 also portrays that there were substantial differences across three different types of the true level-1 error structures ($\eta^2 = .62$). Specifically, average RMSE values of the autocorrelation parameter tended to be larger when the true level-1 error structure was one of the heterogeneous error structures than when it was homogeneous error structure, regardless of the type of model.

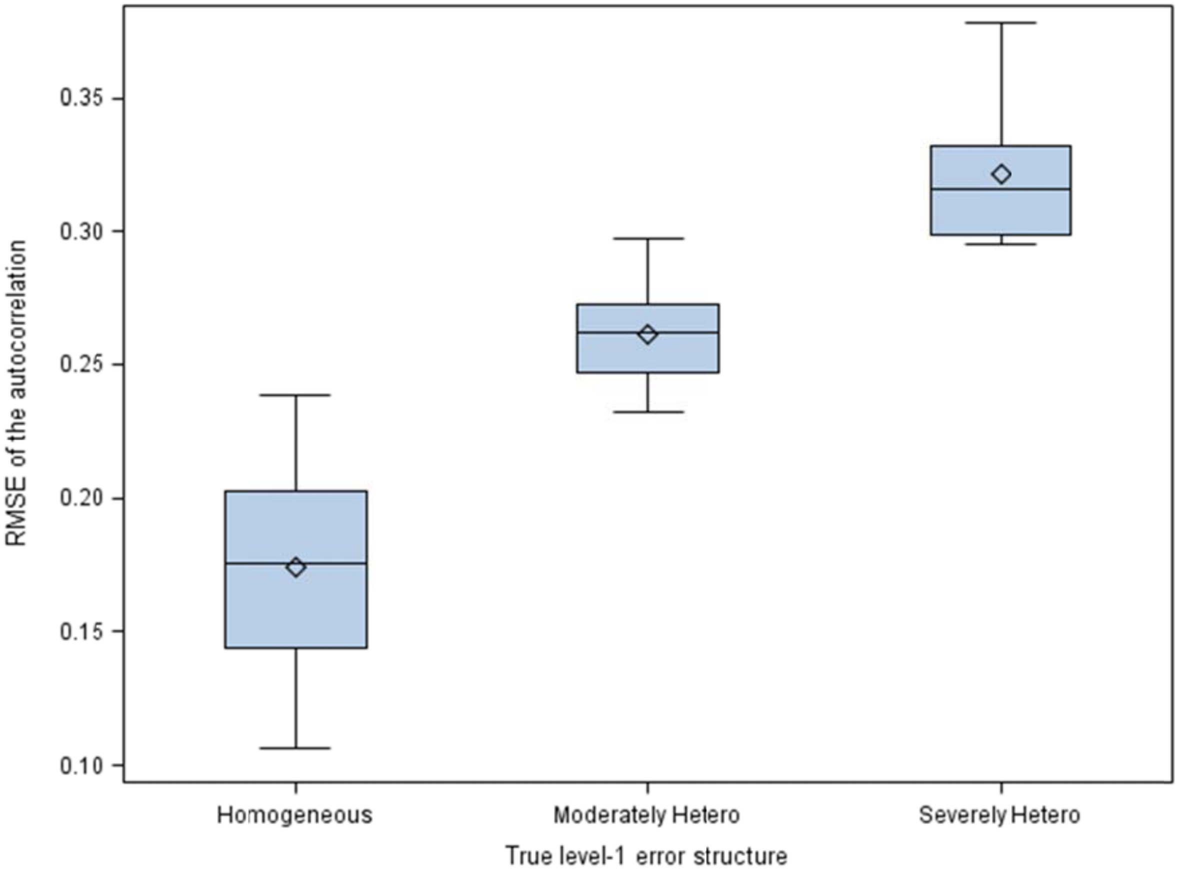


Figure 77. Box plots depicting the estimated RMSE value of the autocorrelation as a function of the true level-1 error structure.

In order to further explore if any design factor had a significant effect on the RMSE value for the autocorrelation, a GLM model was run. The main effects only model explained 94% of the variability, and indicated two of the design factors had a medium or large effect, including the true level-1 error structure ($\eta^2 = .83$) and the series length per case ($\eta^2 = .06$). These main effects are illustrated in Figures 77 and 78.

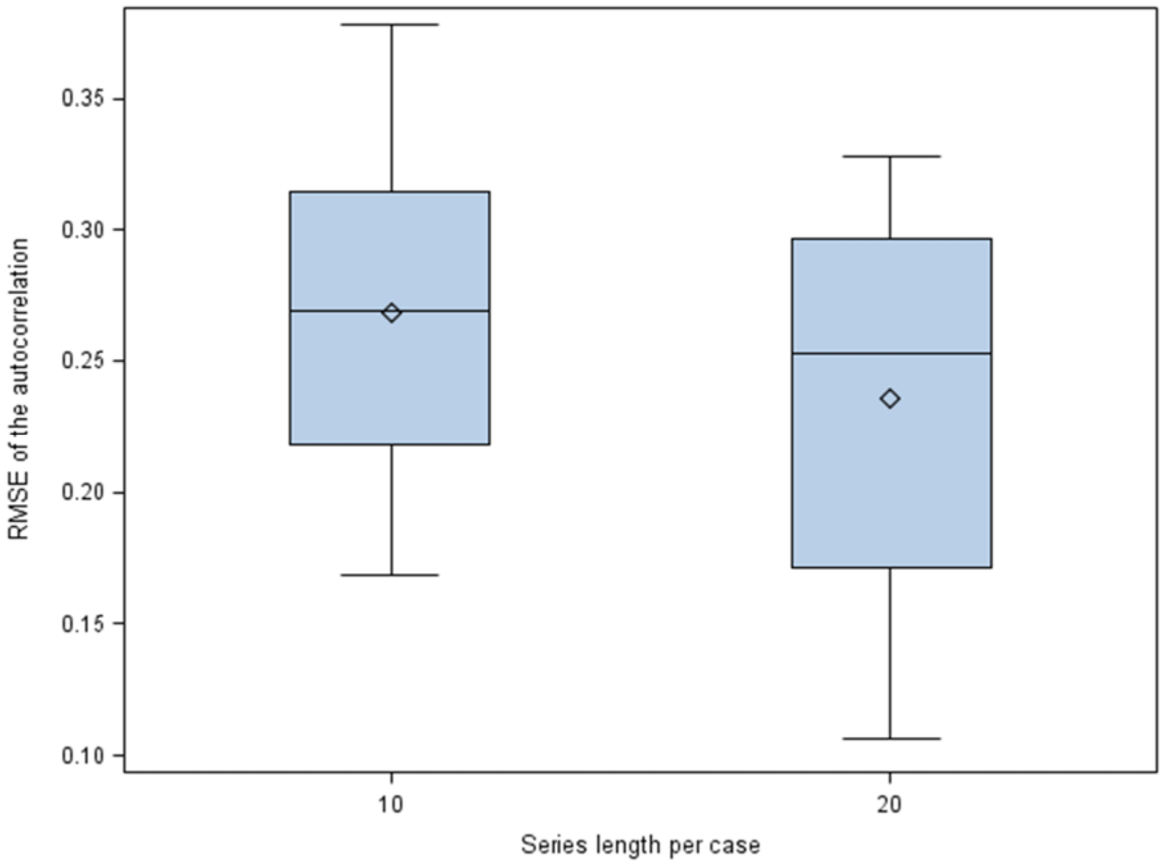


Figure 78. Box plots depicting the estimated RMSE of the autocorrelation as a function of the series length per case.

As illustrated in Figure 77, the average RMSE value of the autocorrelation tended to be increased as the true level-1 error structure shifted from homogeneous to heterogeneous. The average RMSE value for the homogeneous error structure was *M* = 0.17, *SD* = .04, and the average RMSE value for the moderately and severely heterogeneous error structures was *M* = 0.26, *SD* = .02, and *M* = 0.32, *SD* = .02, respectively. In addition, Figure 78 portrays that the average RMSE value was decreased from *M* = 0.27, *SD* = .06 to *M* = 0.24, *SD* = .07, as the series length per case increased from 10 to 20.

Credible Interval Coverage

The distribution of credible interval coverage values of the level-2 error standard deviation for intervention effects (shift in level and shift in slope), the level-1 error standard deviation, and the autocorrelation are illustrated in Figures 79 through 96. The full information about the η^2values for the GLM models is provided in Appendix B.

Level-2 error standard deviation for phase (shift in level). The average credible interval (CI) coverage value of the level-2 error standard deviation for phase were over the nominal value (.95) across the two models with some of the variability explained by the type of model (η^2 = .07) (Figure 79). The average CI coverage value for Model 1 was *M* = 0.968, *SD* = 0.01, and the average CI coverage for Model 2 was *M* = 0.973, *SD* = 0.01.

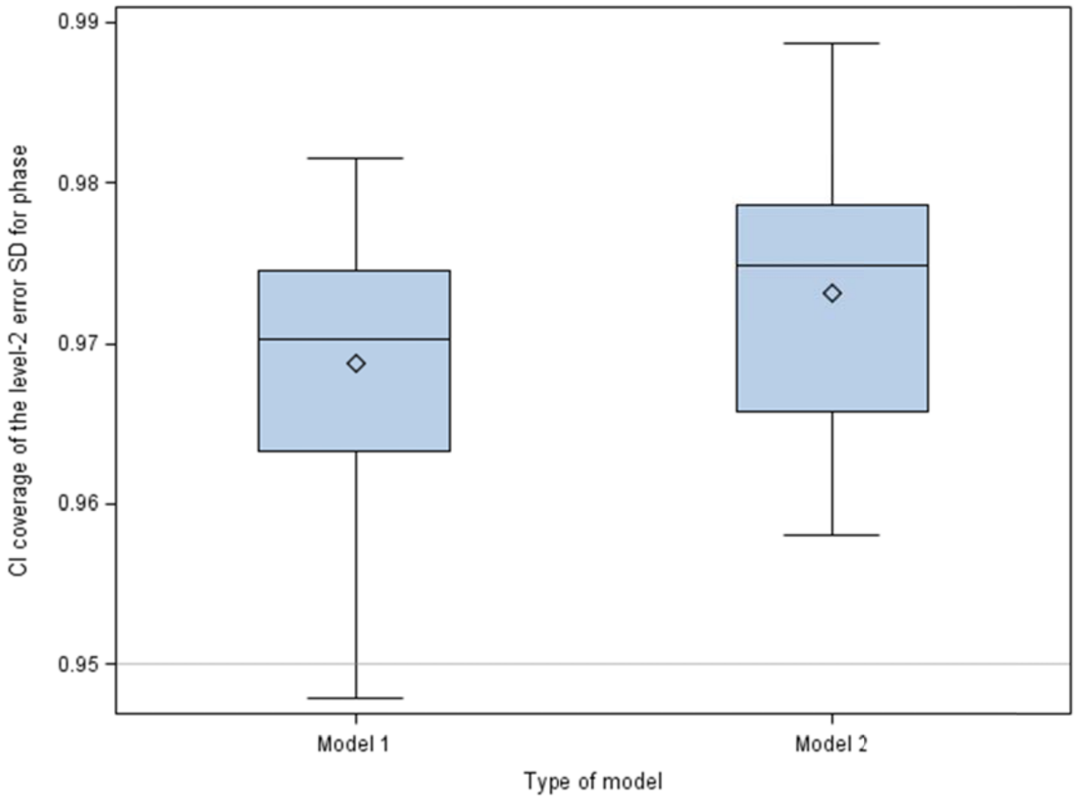


Figure 79. Box plots illustrating the distribution of the CI coverage values of the level-2 error standard deviation for shift in level across Model 1 which did not model between case variation, and Model 2 which models between case variation.

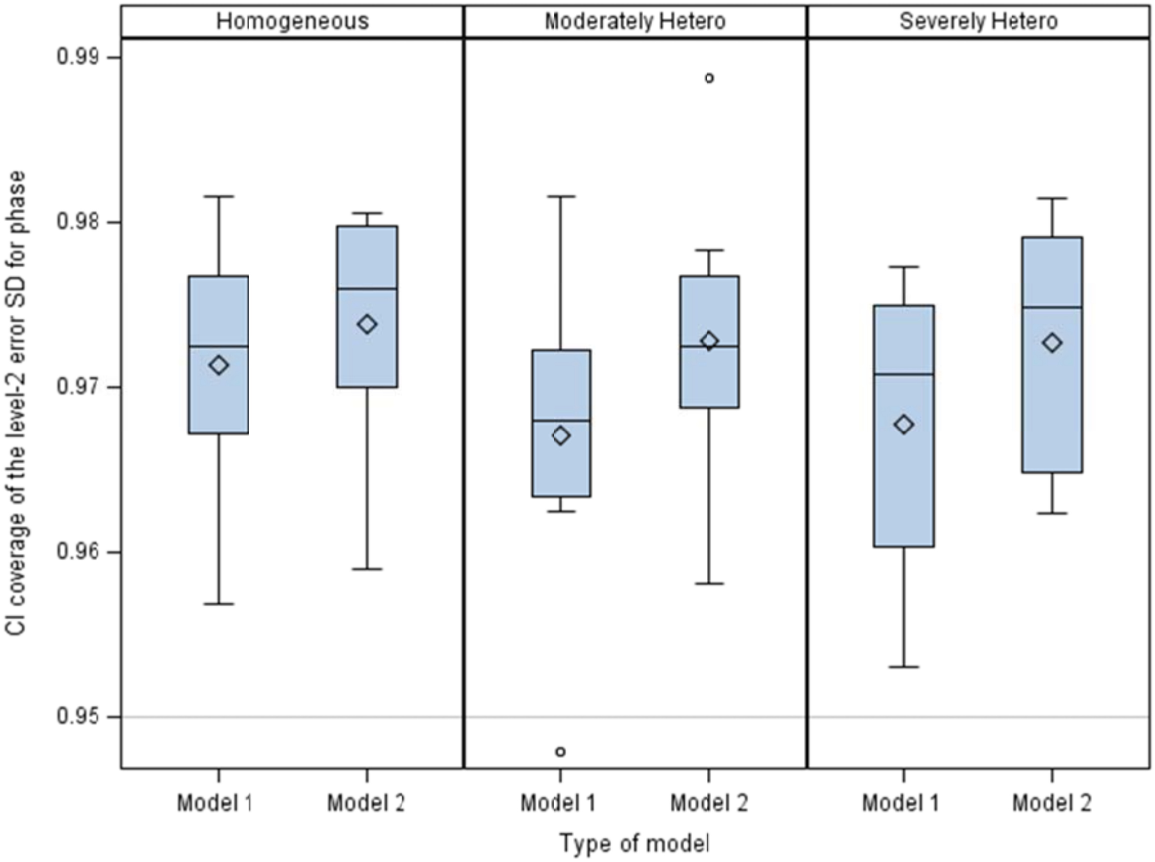


Figure 80. Box plots illustrating the distribution of the CI coverage values of the level-2 error standard deviation for shift in level across the two models within the three true level-1 error structures.

The average CI coverage values of the level-2 error standard deviation for phase across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 80). The average CI coverage values were slightly different across the two models within the three true level-1 error structures, and there were little differences across the three true level-1 error structures, with little of the variability explained by the different types of the true level-1 error structures (η^2 = .02). The smallest average CI coverage difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1|$ = 0.003), and the biggest average CI coverage difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2\text{-}M_1| = 0.006$). Generally, CI coverage for Model 2 tended to be more overly conservative than CI coverage for Model 1.

In order to further explore the variability in the CI coverage values for phase effect, GLM models were run. The model explained 99% of variability after including 4-way interactions, and indicated two main effects and one interaction effect had a medium or large effect, including the variation in the level-2 errors ($\eta^2 = .29$), the type of model ($\eta^2 = .07$), the 3-way interaction among the series length per case, the number of cases, and the true level-1 error structure ($\eta^2 = .11$). These main and interaction effects are illustrated in Figures 81 through 83.

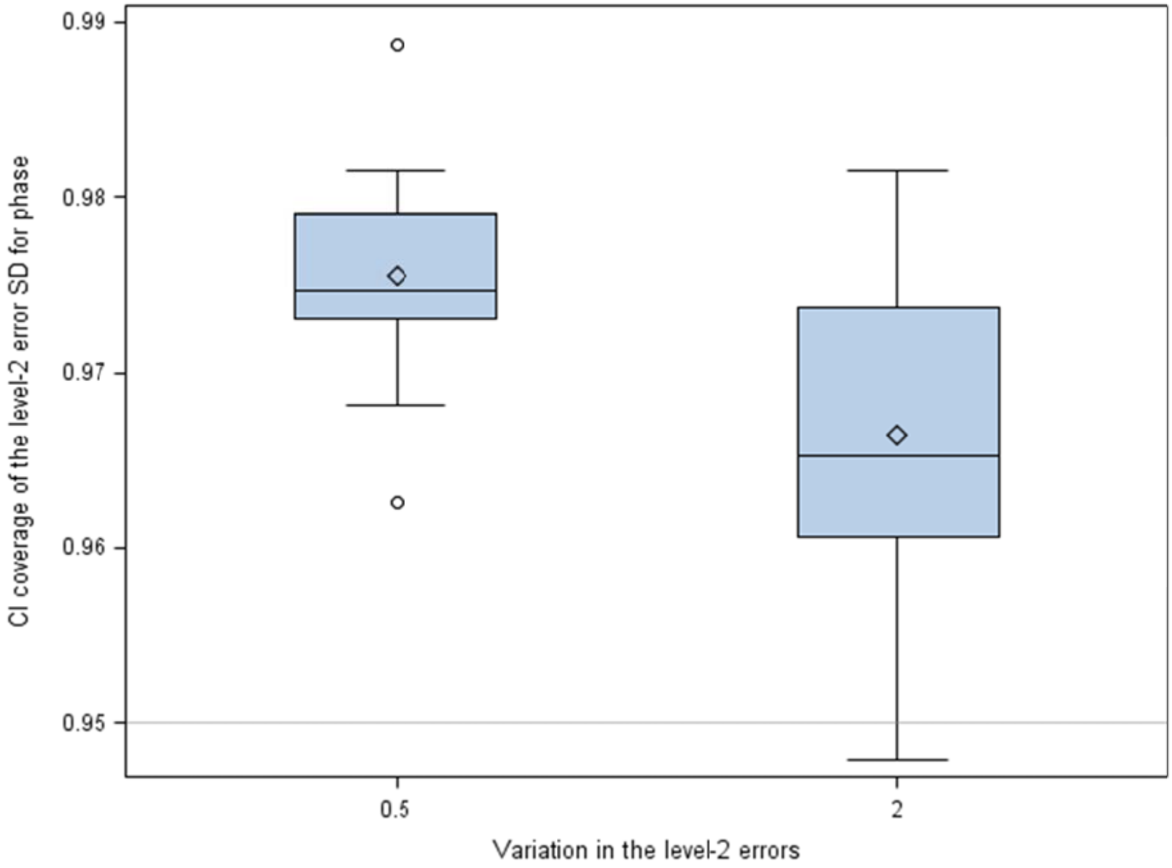


Figure 81. Box plots depicting the estimated CI coverage of the level-2 error standard deviation for shift in level as a function of the variation in the level-2 errors.

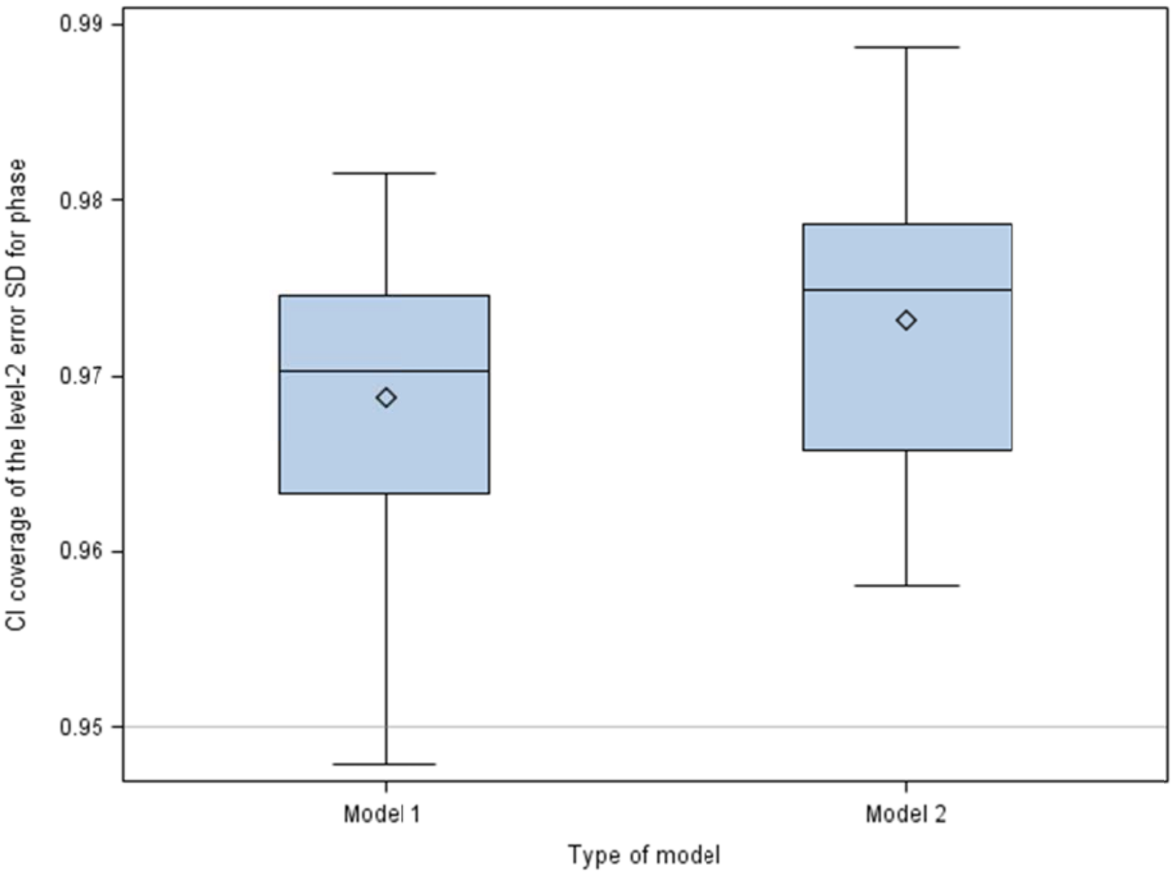


Figure 82. Box plots depicting the estimated CI coverage of the level-2 error standard deviation for shift in level as a function of the type of model.

As illustrated in Figure 81, as the variation in the level-2 errors shifted from most of the variance at the level-1 error (.5) to most of the variance at the level-2 error (2), the average CI coverage approached the nominal level, .95 (from *M* = 0.975, *SD* = .005 to *M* = 0.966, *SD* = .009). Similarly, Figure 82 shows that the CI coverage of Model 2 (*M* = 0.973, *SD* = .008) was more overly conservative than the CI coverage for Model 1 (*M* = 0.969, *SD* = .009).

Figure 83 indicated that when the true level-1 error structure was either homogeneous or moderately heterogeneous, the CI coverage decreased to the nominal level, .95, as the series length per case increased from 10 to 20, regardless of the number of cases. However, when the true level-1 error structure was severely heterogeneous, the relationship between CI coverage and the series length per case was dependent on the number of cases. When the number of case was 8, the CI coverage approached the nominal level, .95, as the series length per case increased from 10 to 20. However, when the number of cases was 4, the CI coverage increased to a more conservative level as the series length per case increased from 10 to 20.

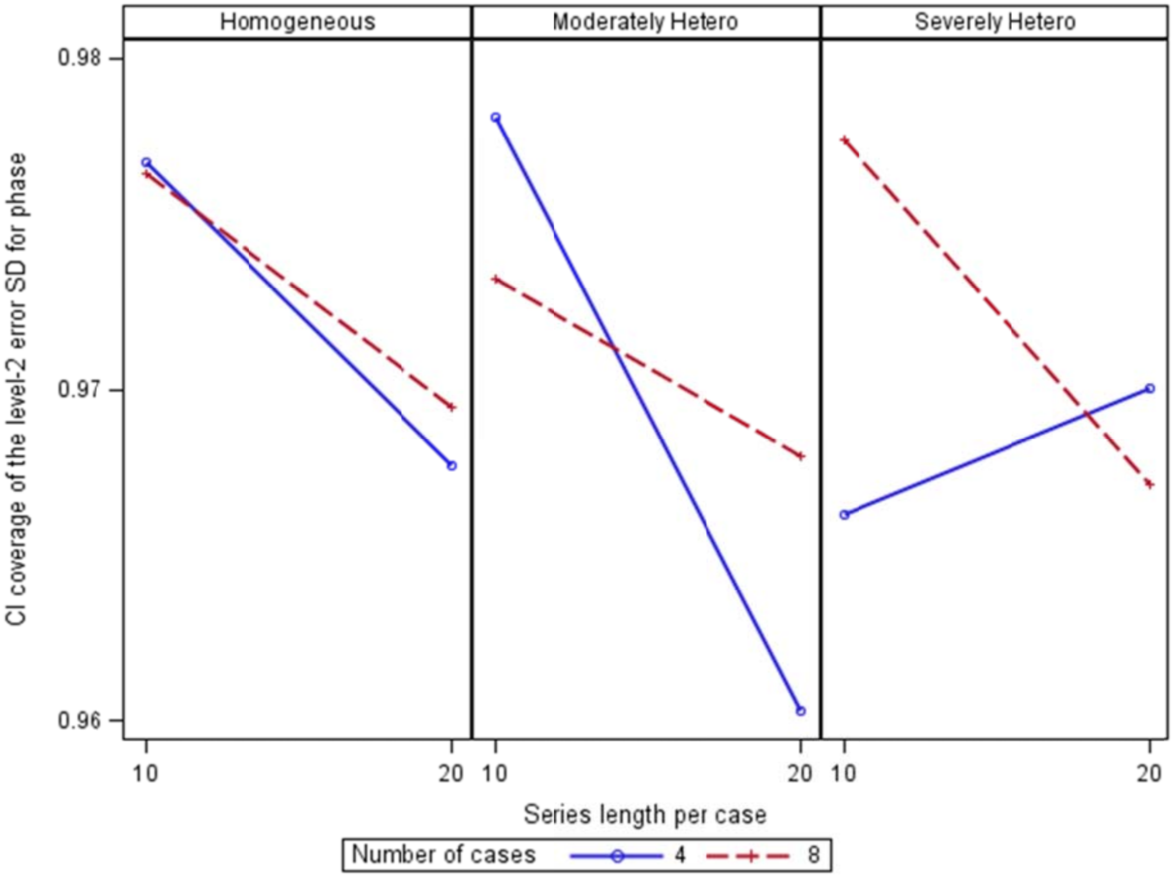


Figure 83. Line graph depicting average CI coverage of the level-2 error standard deviation for phase as a function of the three-way interaction effect among the number of cases, series length per case, and the true level-1 error structure.

Level-2 error standard deviation for interaction (shift in slope). The average credible interval (CI) coverage values of the level-2 error standard deviation for interaction were over the nominal value across the two models (Model 1 and Model 2) (Figure 84). The average CI coverage value for Model 1 was $M = 0.974$, $SD = 0.01$, and the average CI coverage for Model 2 was $M = 0.979$, $SD = 0.01$. The type of model explained some of the variability ($\eta^2 = .06$).

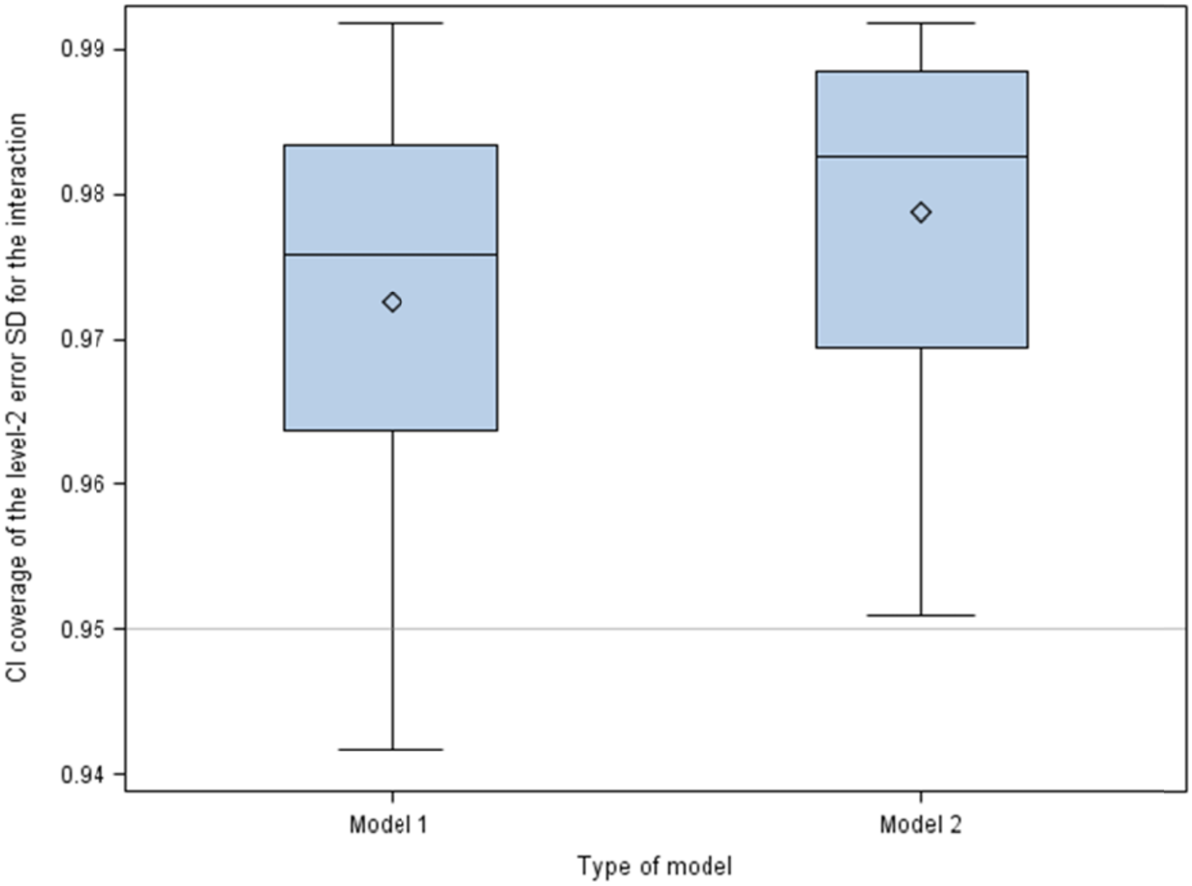


Figure 84. Box plots illustrating the distribution of the CI coverage of the level-2 error standard deviation for shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI coverage values of the treatment effect for interaction across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 85). The average CI coverage values were slightly different across the two models within the three true level-1 error structures, and there were some differences across the true level-1 error structures, with a small amount of the variability explained by the different types of the true level-1 error structures

(η^2 = .05). The smallest average CI coverage difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2$-$M_1|$ = 0.005), and the biggest average CI coverage difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2$-$M_1|$ = 0.008). Generally, the CI coverage for Model 2 tended to be more overly conservative than the CI coverage for Model 1.

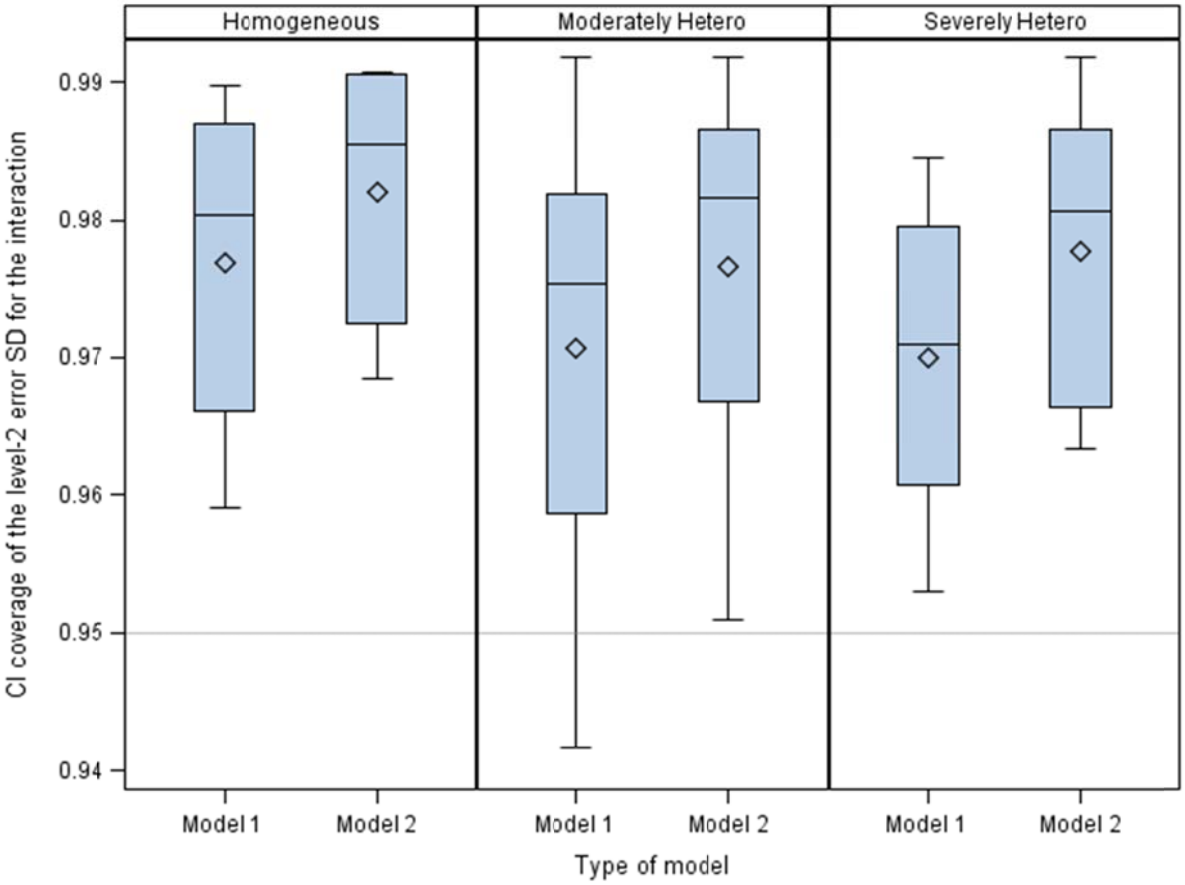


Figure 85. Box plots illustrating the distribution of the CI coverage for the level-2 error standard deviation for shift in slope across the two models within the three true level-1 error structures.

In order to further explore the variability in the CI coverage of the level-2 error standard deviation for interaction, GLM models were run. The model explained 95% of variability after including 2-way interactions. The GLM model found two of the design factors and one interaction effect that had a medium or large effect, including the series length per case (η^2 =

.65), the type of model ($\eta^2 = .06$), and the 2-way interaction between the variation in the level-2 errors and the true level-1 error structure ($\eta^2 = .07$). These main and interaction effects were illustrated in Figures 86 through 88.

Figure 86 portrays that as the series length per case increased from 10 to 20, the average CI coverage approached the nominal level, .95 (from *M* = 0.986, *SD* = .005 to *M* = 0.965, *SD* = .01). Similarly, Figure 87 depicts that the CI coverage for Model 2 *(M* = 0.97, *SD* = .01) tended to be more overly conservative than the CI coverage for Model 1 *(M* = 0.98, *SD* = .01).

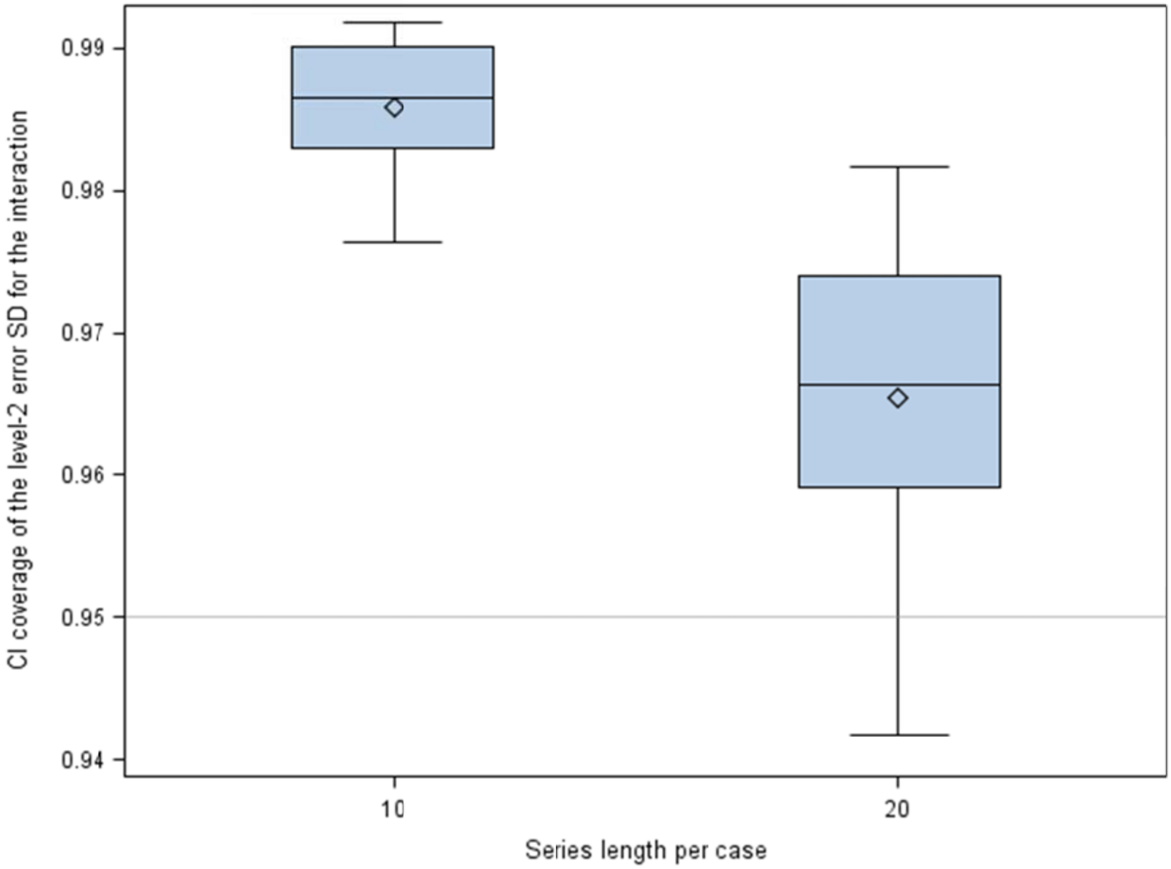


Figure 86. Box plots depicting the estimated CI coverage of the level-2 error standard deviation for shift in slope as a function of the series length per case.

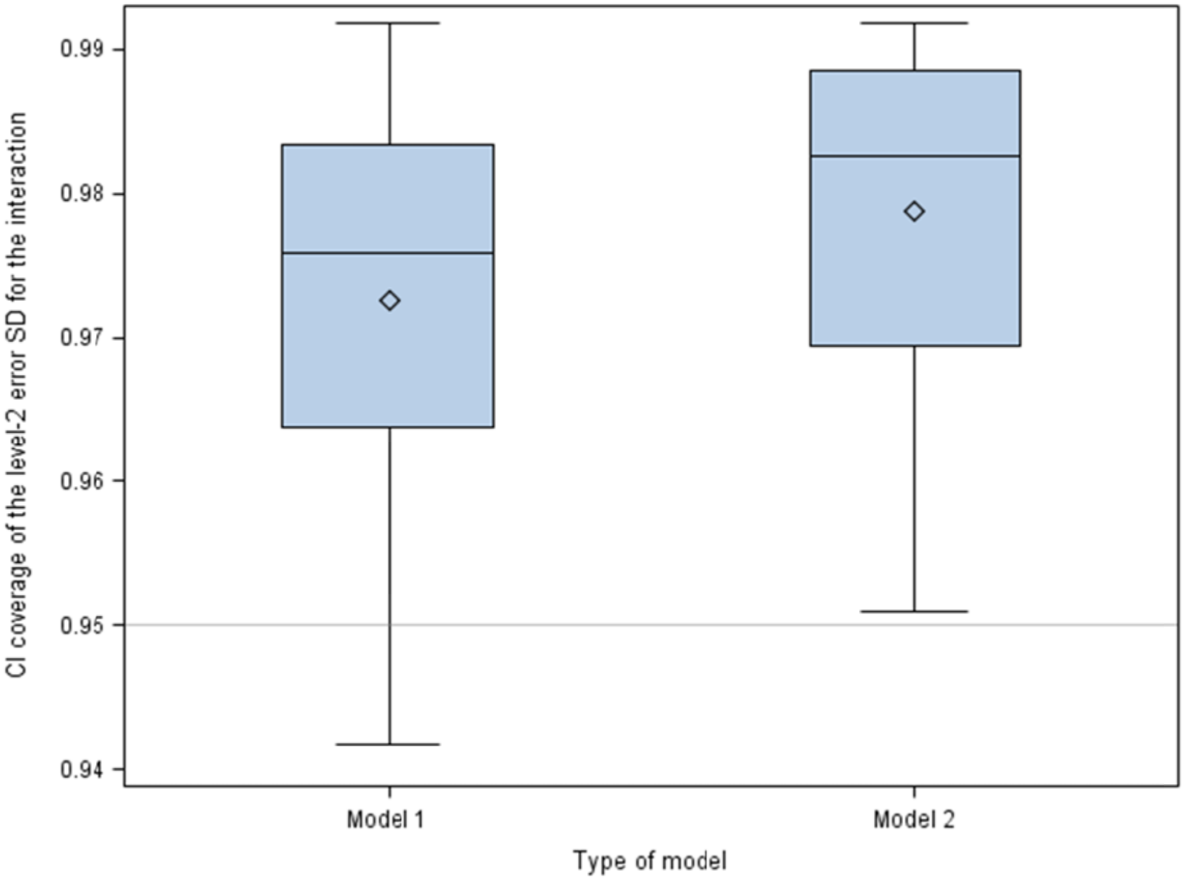


Figure 87. Box plots depicting the estimated CI coverage of the level-2 error standard deviation for shift in slope as a function of the type of model.

A line graph, Figure 87, depicts that the interaction between variation in the level-2 errors and CI coverage was dependent on the true level-1 error structure. When the true level-1 error structure was either homogeneous or moderately heterogeneous, as the variation in the level-2 errors shifted from most of the variance at the level-1 error (.5) to most of the variance at the level-2 error (2), the average CI coverage approached the nominal level, .95. However, when the true level-1 error structure was severely heterogeneous, as the variation in the level-2 errors shifted from most of the variance at the level-1 error (.5) to most of the variance at the level-2 error (2), the average CI coverage slightly increased, indicating moving in a more conservative direction.

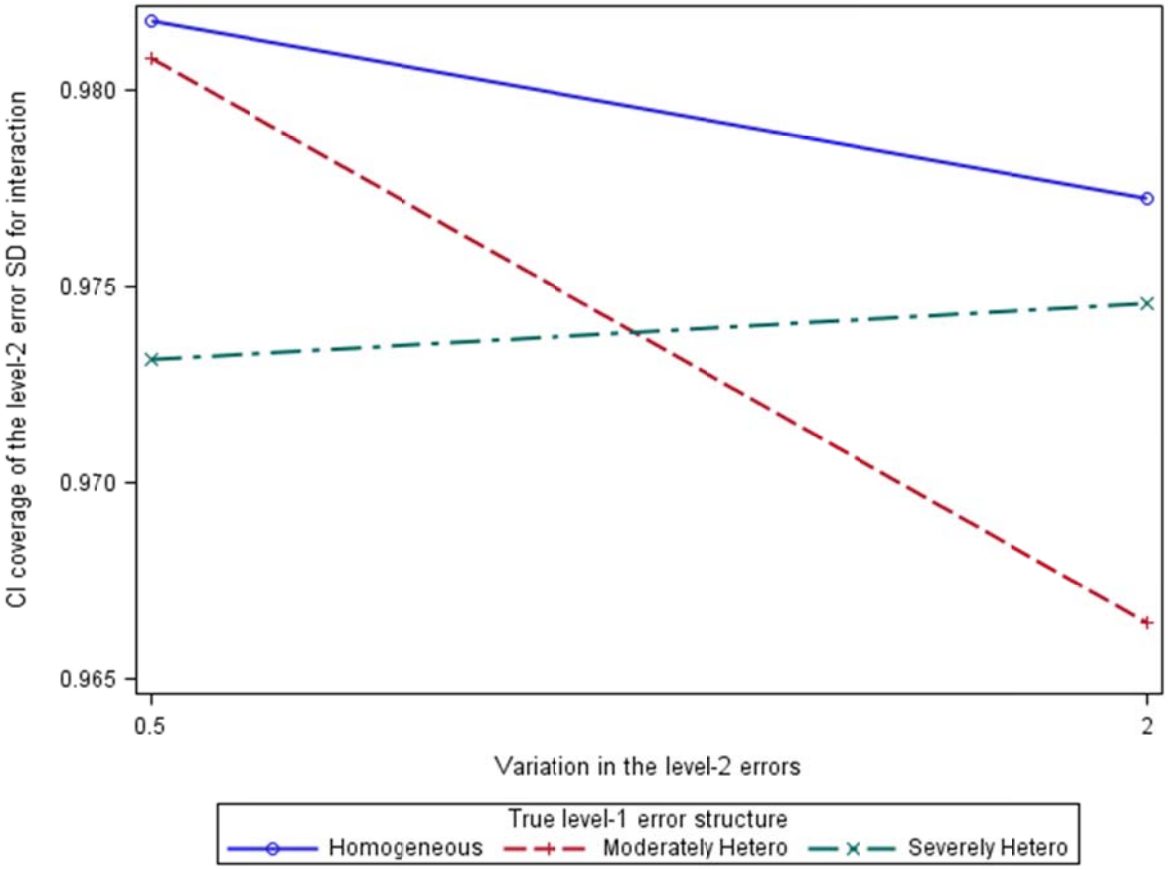


Figure 88. Line graph depicting average CI coverage of the level-2 error standard deviation for the shift in slope as a function of the two-way interaction effect between the variation in the level-2 errors and the true level-1 error structure.

Level-1 error standard deviation. The average credible interval (CI) coverage values of the level-1 error standard deviation were different across the two models (Model 1 and Model 2) with substantial variability explained by the type of model (η^2 = .3) (Figure 89). Specifically, the average CI coverage for Model 1 was *M* = 0.85, *SD* = 0.13, and the average CI coverage for Model 2 was *M* = 0.97, *SD* = 0.02. In addition, the CI coverage for Model 1 had more variability than the CI coverage for Model 2.

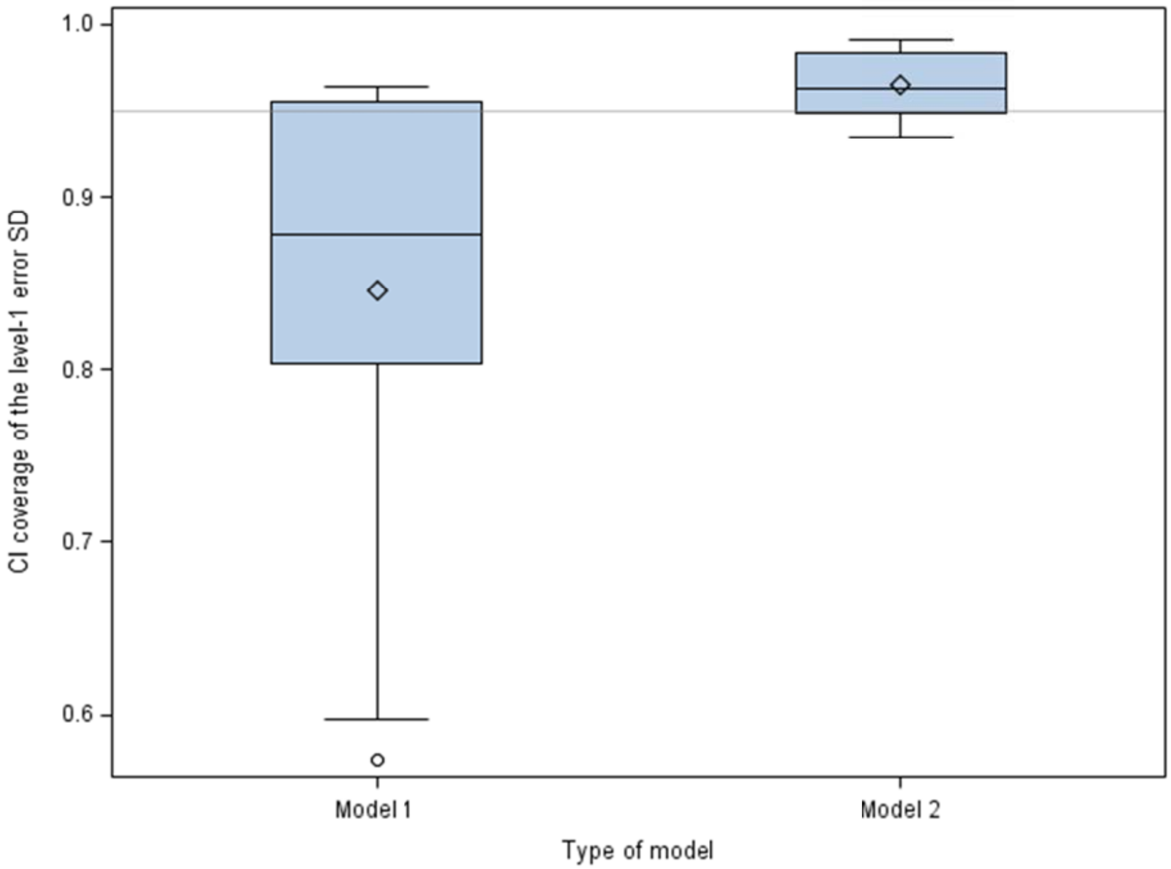


Figure 89. Box plots illustrating the distribution of the CI coverage for the level-1 error standard deviation across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI coverage values for the level-1 error standard deviation across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 90). The average CI coverage values were different across the two models within the three true level-1 error structures, and there were differences across the three true level-1 error structures, with substantial variability explained by the different types of true level-1 error structures ($\eta^2 = .3$). When the true level-1 error structure was homogeneous, both the CI coverage for Model 1 and Model 2 tended to be above the nominal level, $M = 0.96$, $SD = 0.01$ for Model 1, $M = 0.99$, $SD <$

0.01 for Model 2. When the true level-1 error structure was one of the heterogeneous error structures, the CI coverage for Model 1 tended to be under the nominal level (moderately hetero: *M* = 0.88, *SD* = 0.05; severely hetero: *M* = 0.70, *SD* = 0.11) and the CI coverage for Model 2 tended to be either at the nominal level, .95 or slightly above (moderately hetero: *M* = 0.96, *SD* = 0.01; severely hetero: *M* = 0.95, *SD* = 0.01). The CI coverage for Model 1 generally had more variability than the CI coverage for Model 2.

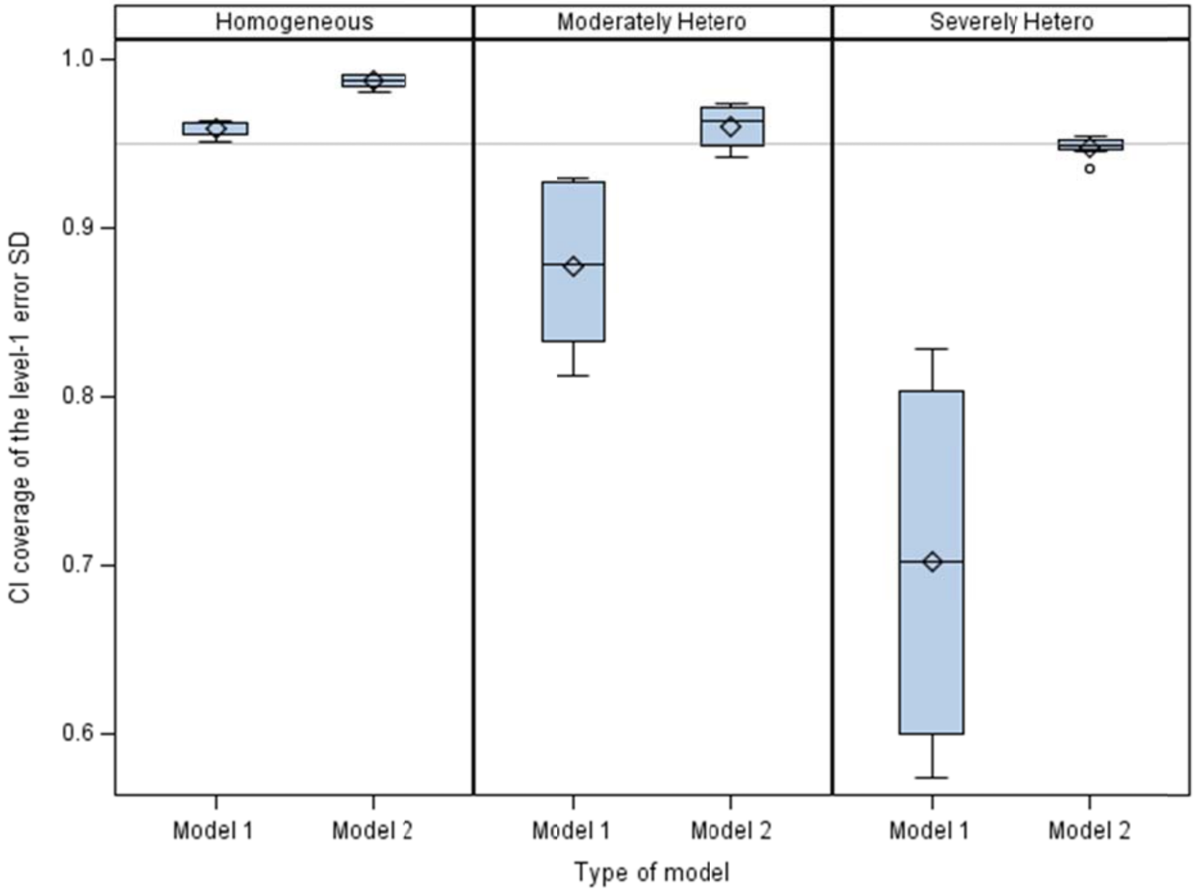


Figure 90. Box plots illustrating the distribution for the CI coverage of the level-1 error standard deivation across the two models within the three true level-1 error structures.

In addition, the smallest average CI coverage difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1|$ = 0.027), and the biggest average CI coverage difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1| = 0.246$).

In order to further explore the variability in the CI coverage of the level-1 error standard deviation, GLM models were run. The model explained 96% of the variability after including 2-way interactions, and indicated one main effect and one interaction effect had a medium or large effect, including the series length per case ($\eta^2 = .06$), and the 2-way interaction between the type of model and the true level-1 error structure ($\eta^2 = .19$). These main and interaction effects were illustrated in Figures 91 through 92.

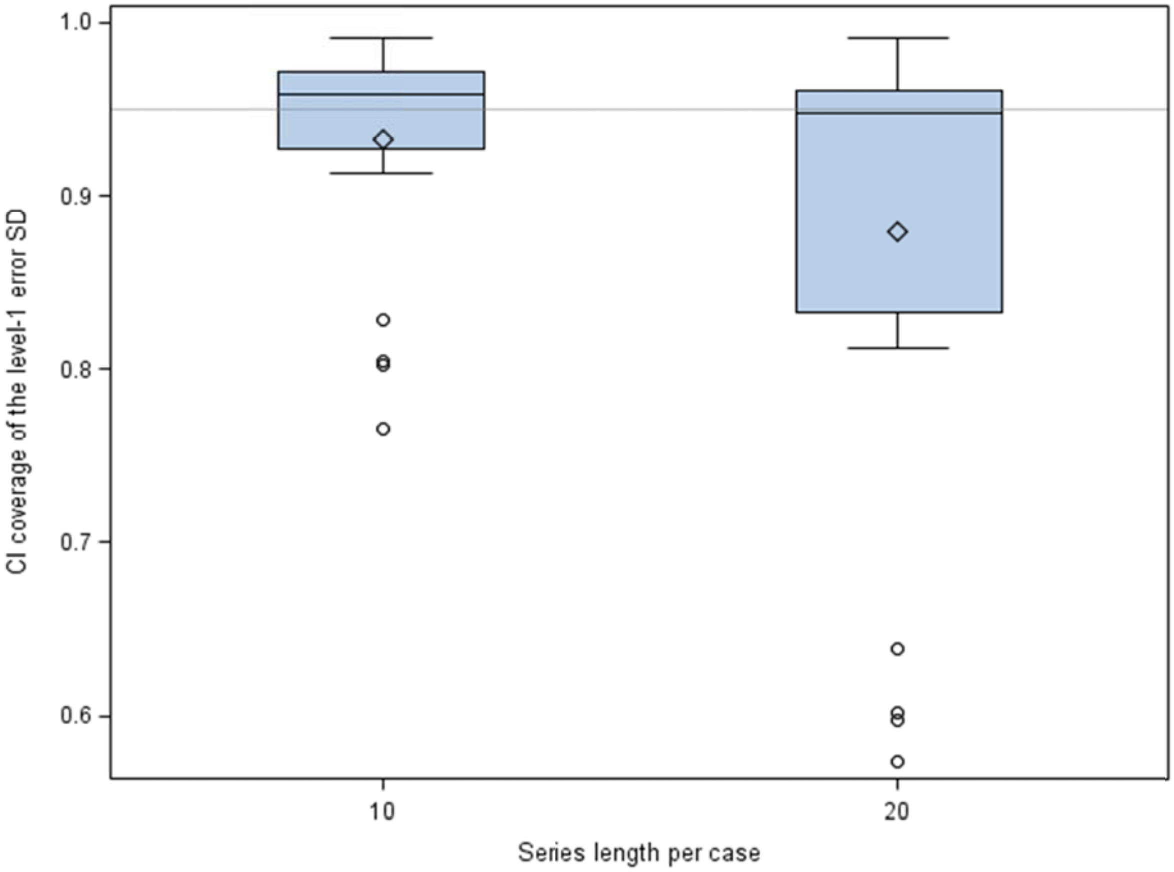


Figure 91. Box plots depicting the estimated CI coverage of the level-1 error standard deviation as a function of the series length per case.

As illustrated in Figure 91, as the series length per case increased from 10 to 20, the average CI coverage decreased from *M* = 0.93, *SD* = .06 to *M* = 0.88, *SD* = .14. In addition, more variability of the CI coverage shows when the series length per case was 20.

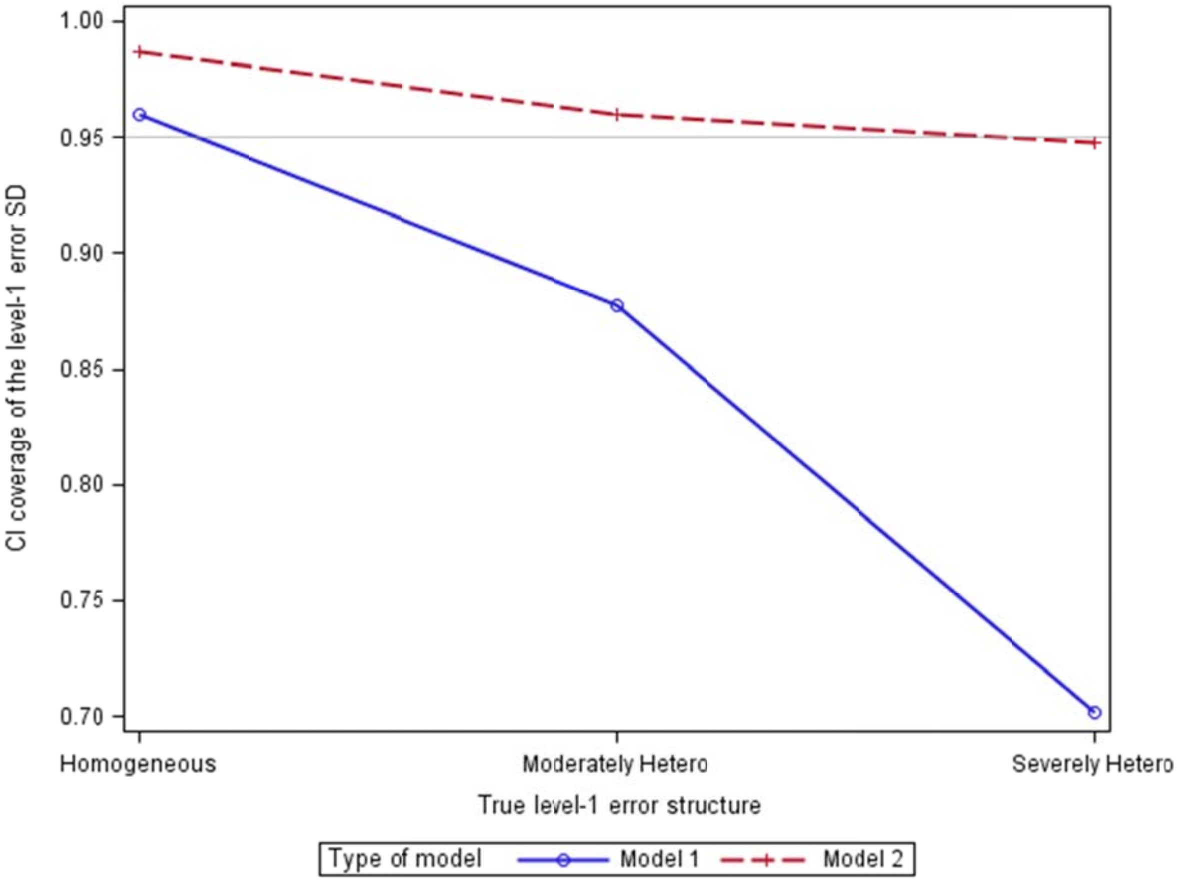


Figure 92. Line graph depicting average CI coverage of the level-1 error standard deviation as a function of the two-way interaction effect between the type of model and the true level-1 error structure.

Figure 92 indicated that when the true level-1 error structure was homogeneous, the CI coverage of the level-1 error standard deviation was above the nominal level for both Model 1 and Model 2. However, when the true level-1 error structure was one of the heterogeneous error structures, CI coverage of the level-1 error standard deviation decreased to the nominal level, .95 for Model 2, while decreased below the nominal level (under covered) for Model 1. Therefore, the difference of the CI coverage in the level-1 error standard deviation between Model 1 and Model 2 was smaller when the true level-1 error structure was homogeneous than one of the heterogeneous error structures, and as the severity of the heterogeneous in the error structure increased from moderately heterogeneous to severely heterogeneous, the difference of the CI coverage in the level-1 error standard deviation between Model 1 and Model 2 greatly increased.

Autocorrelation. The average credible interval (CI) coverage values of the autocorrelation were different across the two models (Model 1 and Model 2) (Figure 93). The average CI coverage for Model 1 was under the nominal value, *M* = 0.81, *SD* = 0.17, and the average CI coverage for Model 2 was close to the nominal level, *M* = 0.94, *SD* = 0.06.

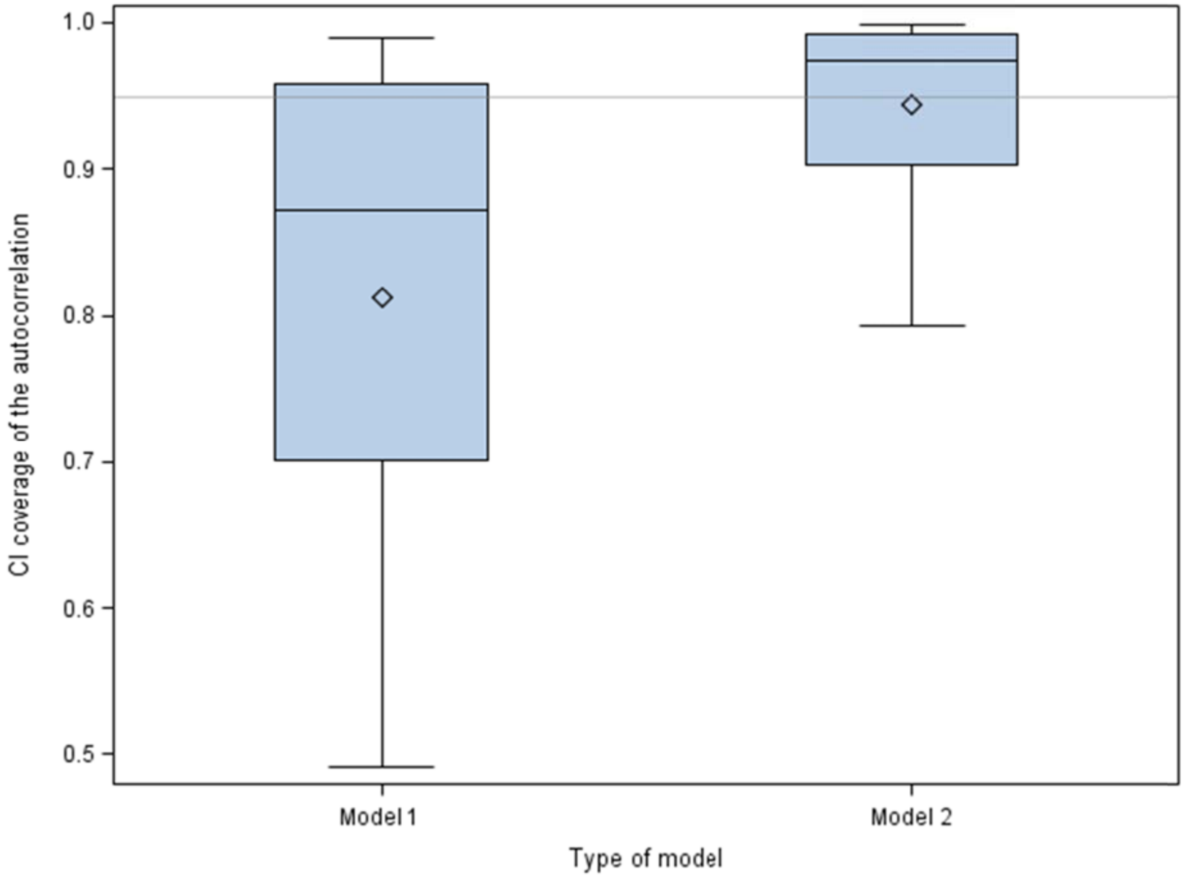


Figure 93. Box plots illustrating the distribution of the CI coverage for the autocorrelation across Model 1 which did not model between case variation, and Model 2 which models between case variation.

In addition, the CI coverage of the autocorrelation for Model 1 had more variability than Model 2. The type of model explained substantial variability ($\eta^2 = .2$).

The average CI coverage values for the autocorrelation across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 94). The average CI coverages were different across the two models within the three true level-1 error structures, and there were differences across the true level-1 error structures, with substantial variability explained by the different types of true level-1 error structures ($\eta^2 = .30$). When the true level-1 error structure was homogeneous, the average CI coverage was over the nominal value for both Model 1 and Model 2 (Model 1: $M = 0.97$, $SD = 0.01$; Model 2: $M = 0.99$, $SD < 0.01$). However, when the true level-1 error structure was one of the heterogeneous error structures, the average CI coverage for Model 1 was severely under the nominal value (moderately hetero: $M = 0.77$, $SD = 0.15$; severely hetero: $M = 0.70$, $SD = 0.15$), while the average CI coverage for Model 2 was either close to the nominal level or slightly under the nominal level (moderately hetero: $M = 0.94$, $SD = 0.05$; severely hetero: $M = 0.90$, $SD = 0.07$). The smallest average CI coverage difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1| = 0.028$), and the biggest average CI coverage difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1| = 0.204$). Generally, the CI coverage for Model 1 tended to have more variability than the CI coverage for Model 2.

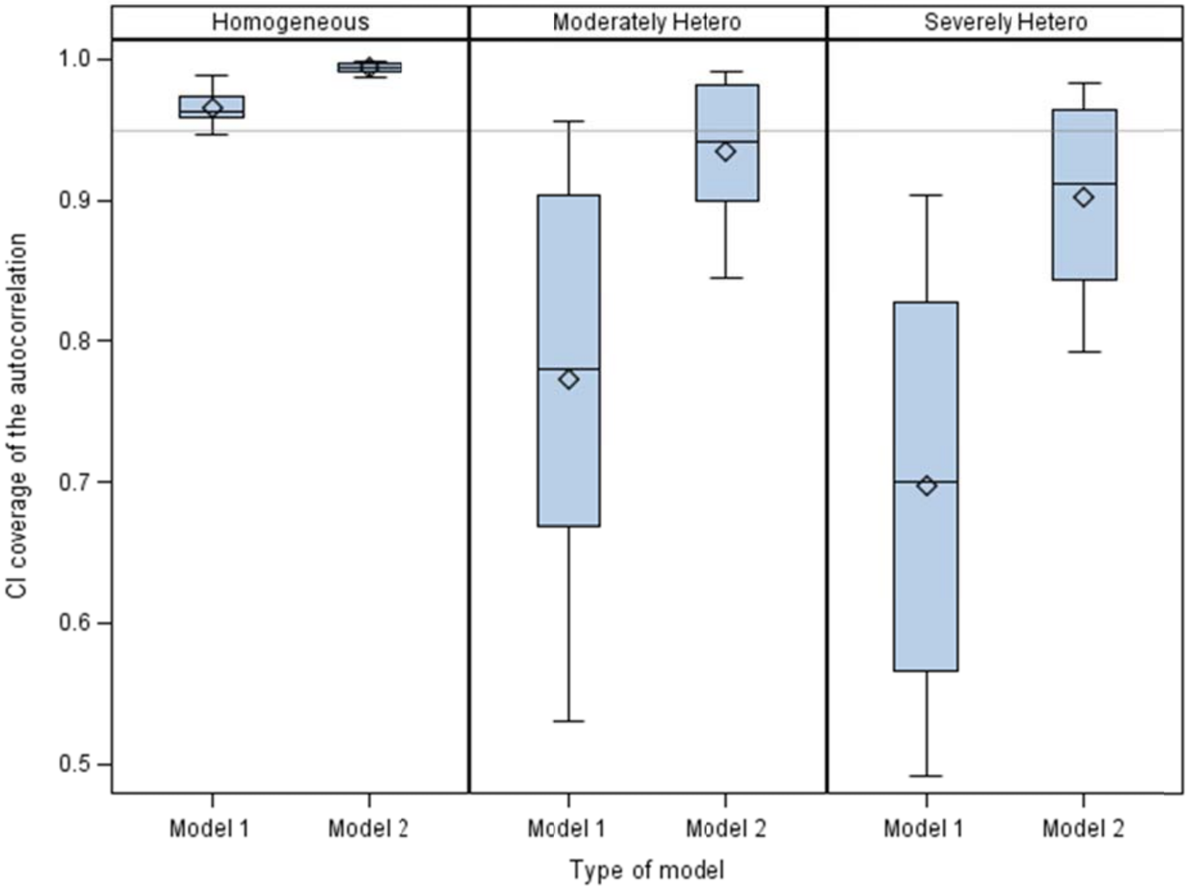


Figure 94. Box plots illustrating the distribution of the CI coverage for the autocorrelation within the three true level-1 error structures.

In order to further explore the variability in the CI coverage of the autocorrelation, GLM models were run. The model explained 98% of variability after including 2-way interactions. The GLM model found two interaction effects that had a medium effect, including the 2-way interaction between the type of model and the true level-1 error structure ($\eta^2 = .07$), and the 2-way interaction between the series length per case and the true level-1 error structure ($\eta^2 = .07$). These interaction effects were illustrated in Figures 95 and 96.

As illustrated in Figure 95, when the true level-1 error structure was homogeneous, the CI coverage was over the nominal level for both Model 1 and Model 2. However, when the true level-1 error structure was one of the heterogeneous error structures, the CI coverage slightly decreased for Model 2, while decreased severely for Model 1. Therefore, the difference of the CI coverage between Model 1 and Model 2 was smaller when the true level-1 error structure was homogeneous than one of the heterogeneous error structures, and as the severity of the heterogeneity in the level-1 error structure increased from moderately heterogeneous to severely heterogeneous, the difference of the CI coverage in the level-1 error standard deviation between Model 1 and Model 2 greatly increased.

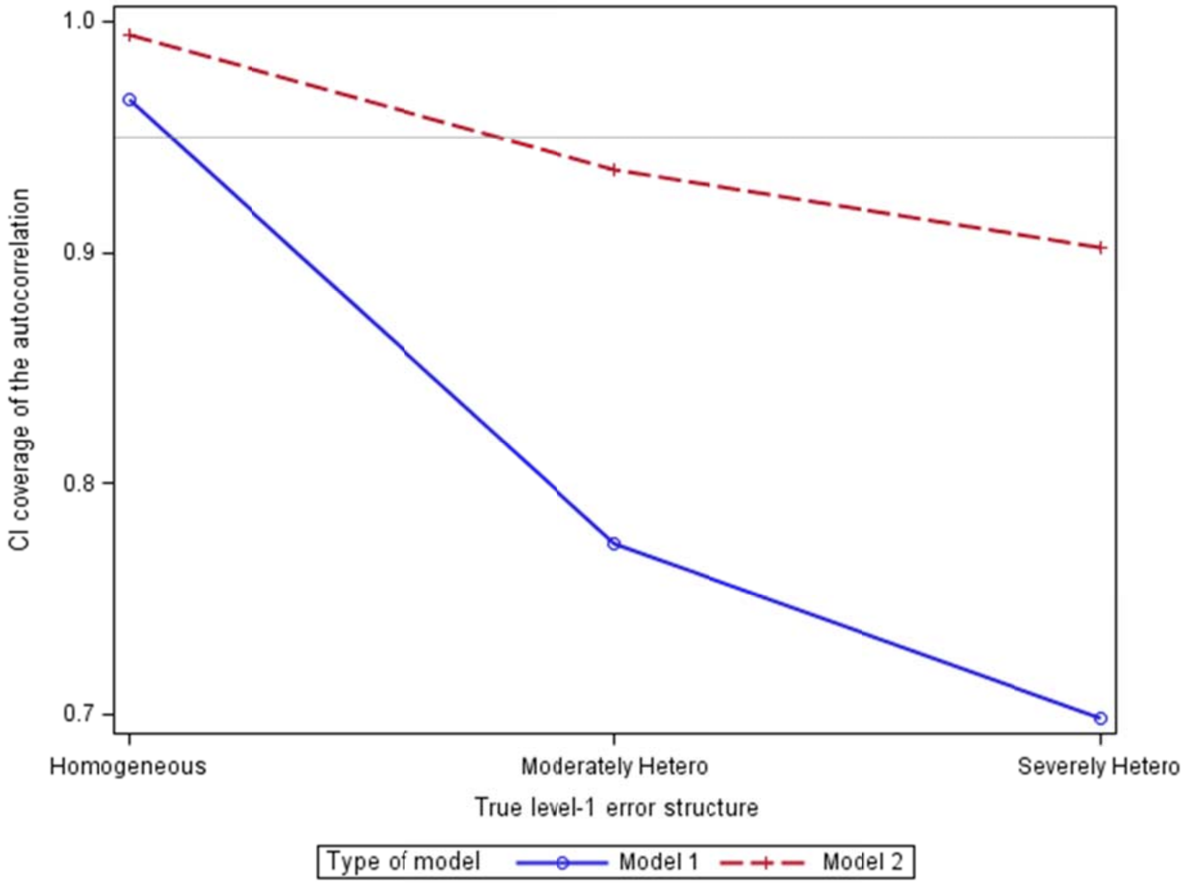


Figure 95. Line graph depicting average CI coverage of the autocorrelation as a function of the two-way interaction effect between the type of model and the true level-1 error structure.

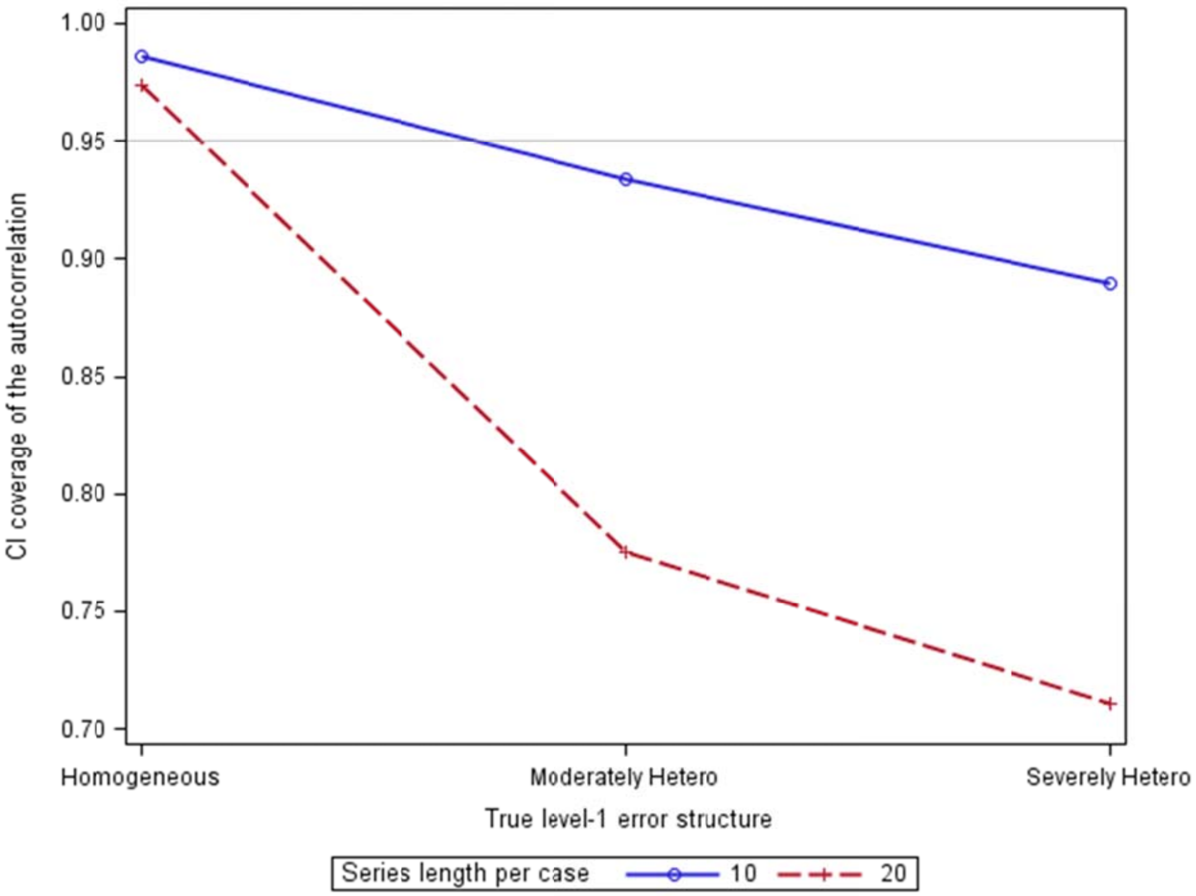


Figure 96. Line graph depicting average CI coverage of the autocorrelation as a function of the two-way interaction effect between the series length per case and the true level-1 error structure.

Similarly, the line graph in Figure 96 portrays that when the true level-1 error structure was homogeneous, the CI coverage was similar and above the nominal level for both the series length per case of 10 or 20 (series length per case 10: *M* = 0.99, *SD* = 0.02; series length per case 20: *M* = 0.97, *SD* = 0.02). However, when the true level-1 error structure was one of the heterogeneous error structures, CI coverage slightly decreased for the series length per case of 10 (moderately hetero: *M* = 0.93, *SD* = 0.07; severely hetero: *M* = 0.89, *SD* = 0.10), while substantially decreased for the series length per case of 20 (moderately hetero: *M* = 0.78, *SD* = 0.15; severely hetero: *M* = 0.71, *SD* = 0.16). Therefore, the difference of the CI coverage between the series length per case of 10 and 20 was smaller when the true level-1 error structure was homogeneous than one of the heterogeneous error structures, and as the severity of the heterogeneity in the level-1 error structure increased from moderately heterogeneous to severely heterogeneous, the difference of the CI coverage between the series length per case of 10 and 20 increased greatly.

Credible Interval Width

The distribution of credible interval width values of the level-2 error standard deviation for intervention effects (shift in level and shift in slope), the level-1 error standard deviation, and the autocorrelation are illustrated in Figures 97 through 115. The full information about the η^2 values for the GLM models is provided in Appendix B.

Level-2 error standard deviation for phase (shift in level). The average credible interval (CI) width values of the level-2 error standard deviation for phase were very similar across the two models (Model 1 and Model 2) (Figure 97). The average CI width for Model 1 was $M = 6.43$, $SD = 3.87$, and the average CI width for Model 2 was $M = 6.44$, $SD = 3.88$. The type of model explained little of the variability ($\eta^2 <.00001$).

The average CI width values of the level-2 error standard deviation for phase across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 98).

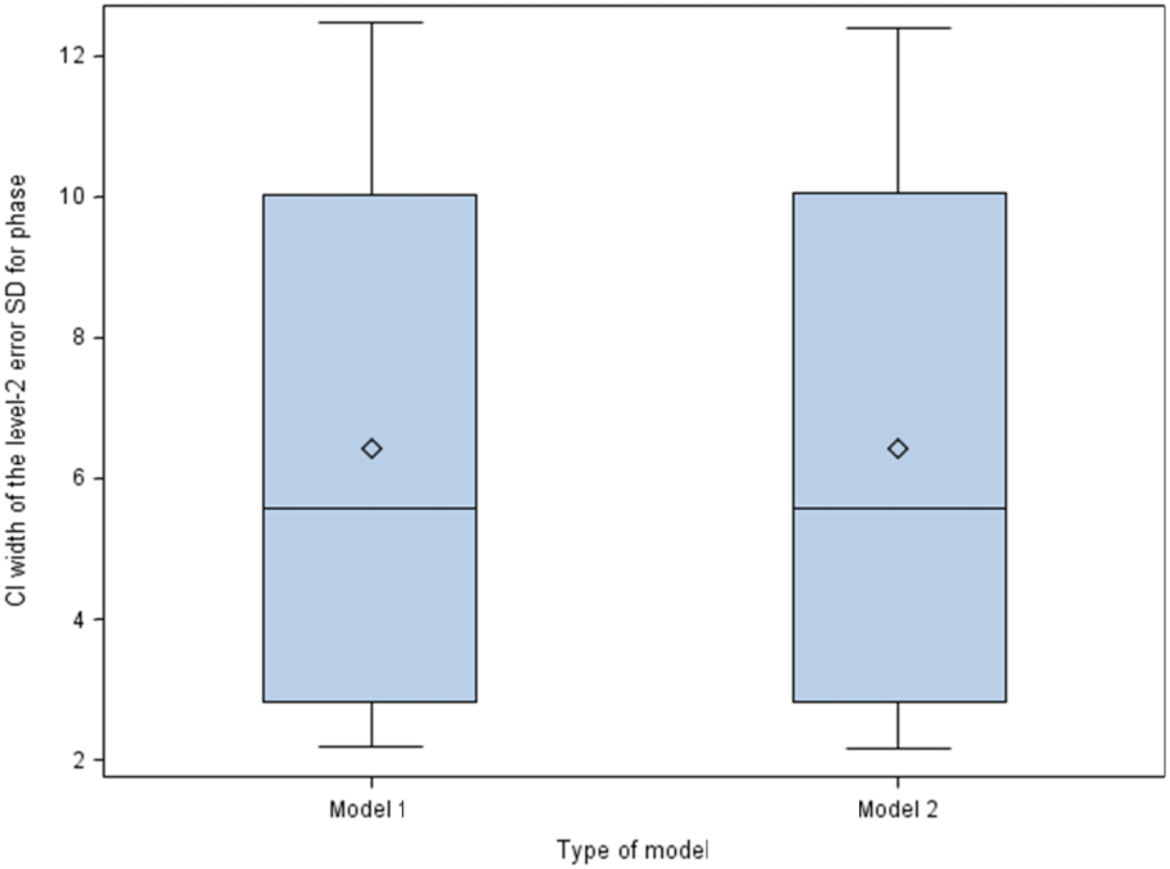


Figure 97. Box plots illustrating the distribution for the CI width of the level-2 error standard deviation for shift in level across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI widths were very similar across the two models within the three true level-1 error structures, and there were no or little differences across the true level-1 error structures, with little of the variability explained by the different types of the true level-1 error structures (η^2 = .0002). The smallest average CI width difference between the two models was found when the true level-1 error structure was moderately heterogeneous ($|M_2\text{-}M_1|$ = .032), and the biggest average CI width difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1|$ = .054).

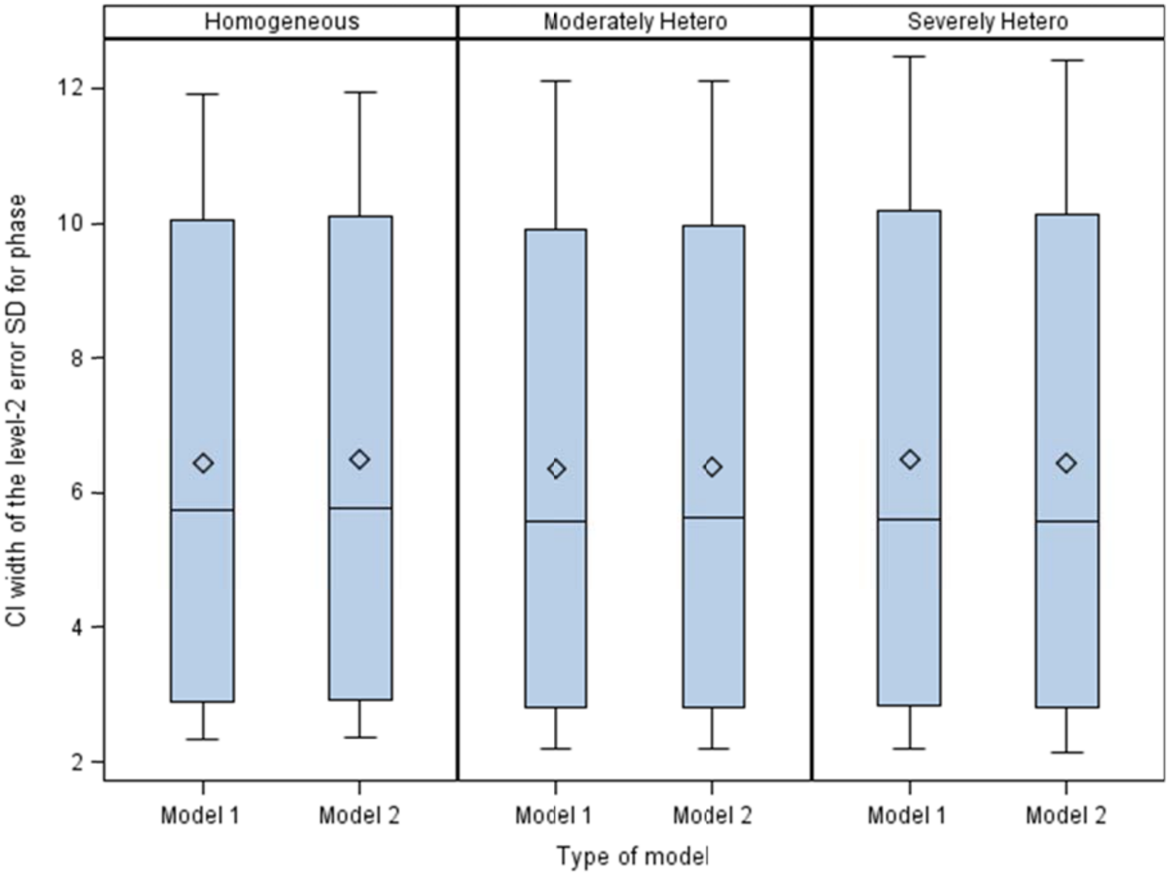


Figure 98. Box plots illustrating the distribution for the CI width of the level-2 error standard deviation for the shift in level across the two models within the three true level-1 error structures.

In order to further explore the variability in the CI width of the level-2 error standard deviation for phase, a GLM model was run. The main effects only model explained 97% of the variability. It was found that two of the design factors had a medium or large effect including the number of cases (η^2 = .88) and the variation in the level-2 errors (η^2 = .08). These main effects are illustrated in Figure 99 and 100.

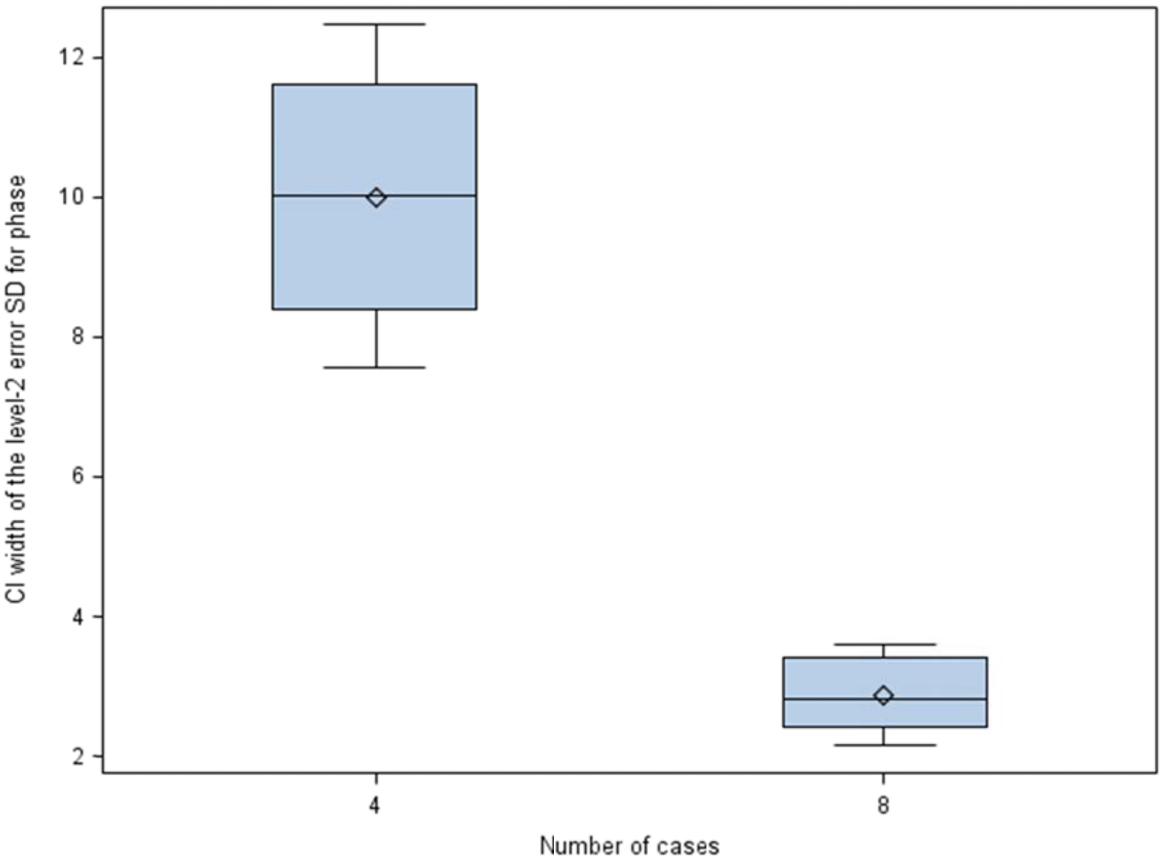


Figure 99. Box plots depicting the estimated CI width of the level-2 error standard deviation for shift in level as a function of the number of cases.

As illustrated in Figure 99, as the number of cases increased from 4 to 8, the average CI width decreased from *M* = 9.995, *SD* = 1.80 to *M* = 2.878, *SD* = 0.54. Similarly, as the variation of error shifted from most of the variance at the level-1 error (.5) to most of the variance at the level-2 error (2), the average CI width increased from *M* = 5.36, *SD* = 3.08 to *M* = 7.52, *SD* = 4.25 (Figure 100).

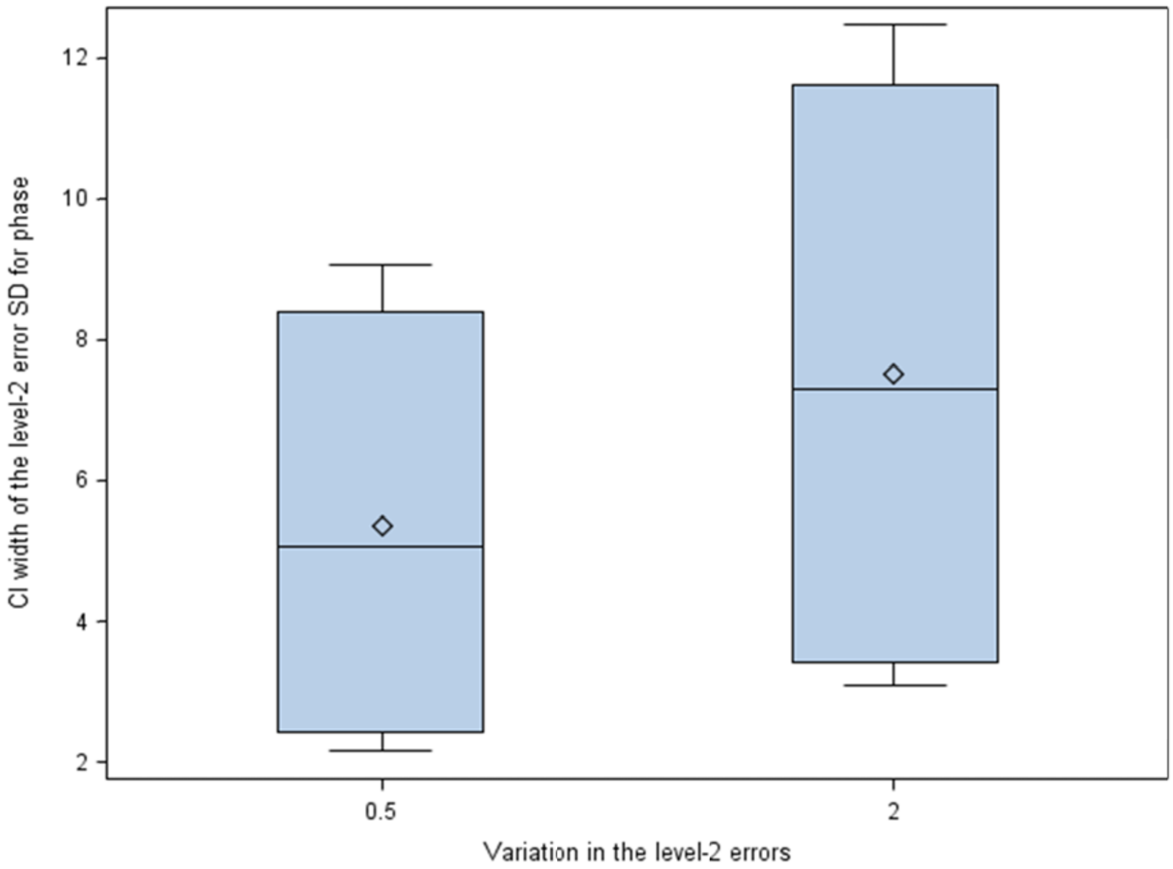


Figure 100. Box plots depicting the estimated CI width of the level-2 error standard deviation for shift in level as a function of the variation in the level-2 errors.

Level-2 error standard deviation for interaction (shift in slope). The average credible interval (CI) width values of the level-2 error standard deviation for the interaction effect were slightly different across the two models (Model 1 and Model 2) (Figure 101). The average CI width for Model 1 was M = 1.62, SD = 1.11, and the average CI width for Model 2 was M = 2.16, SD = 1.46. The type of model explained a small amount of the variability (η^2 =.04). The average CI width values of the level-2 error standard deviation for the interaction effect across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 102).

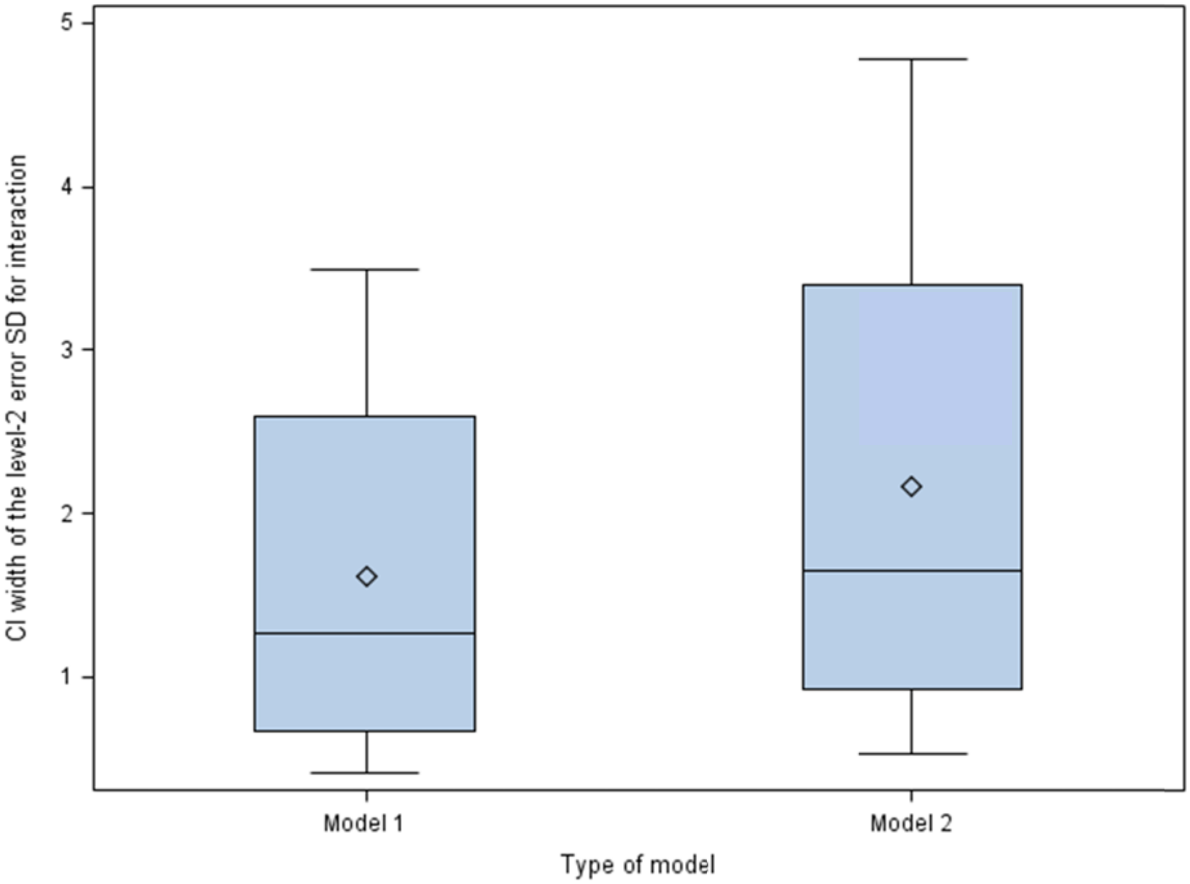


Figure 101. Box plots illustrating the distribution of the CI width for the level-2 error standard deviation for shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI widths were slightly different across the two models within the three true level-1 error structures, and there were no or little differences across the true level-1 error structures, with little of the variability explained by the different types of the true level-1 error structures (η^2 = .0002). The smallest average CI width difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1|$ = .524), and the biggest average CI width difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1|$ = .569).

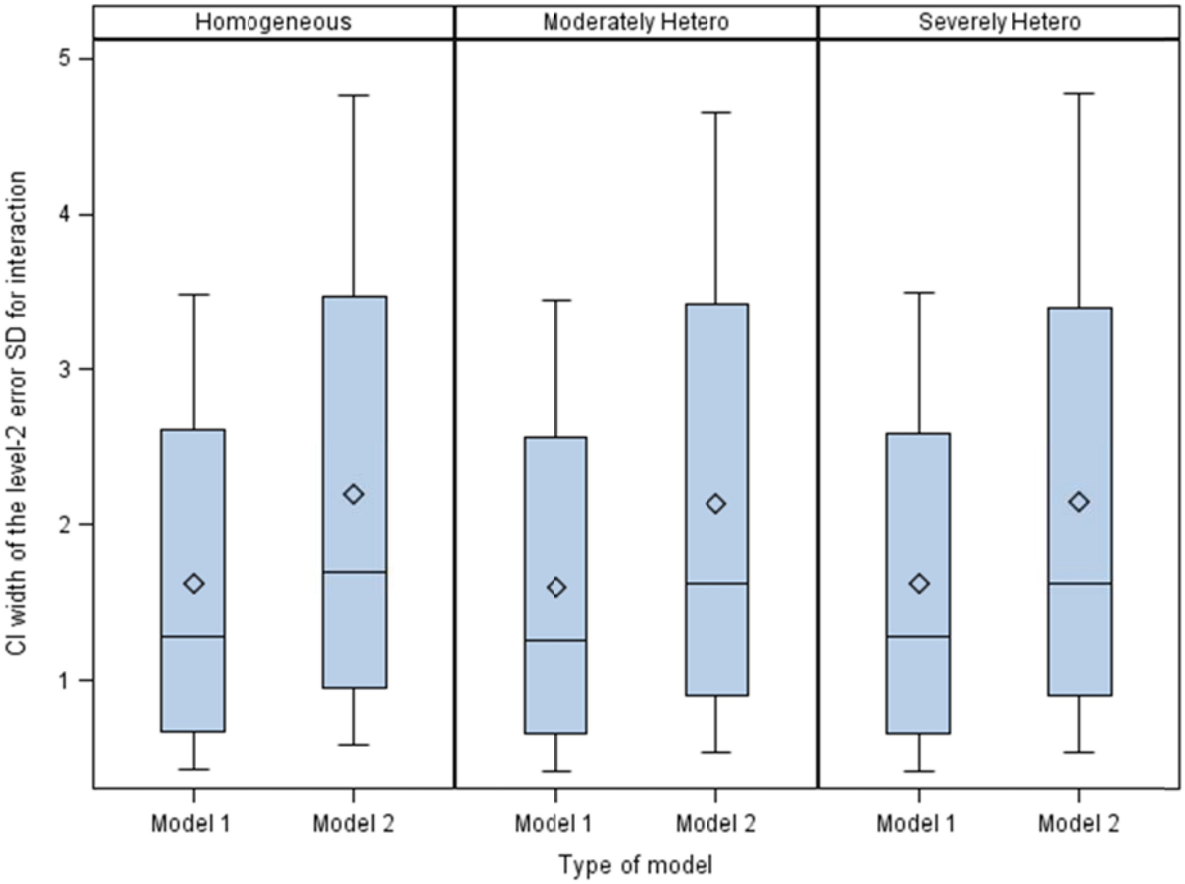


Figure 102. Box plots illustrating the distribution for the CI width of the level-2 error variance for shift in slope across the two models within the three true level-1 error structures.

In order to further explore the variability in the CI width of the level-2 error standard deviation for the interaction effect, GLM models were run. The model explained 99% of the variability after including 2-way interactions. The GLM model found three main effects that had a medium or large effect, including the number of cases (η^2 = .70), variation in the level-2 errors (η^2 = .10), and the series length per case (η^2 = .07). These main effects are illustrated in Figures 103 through 105.

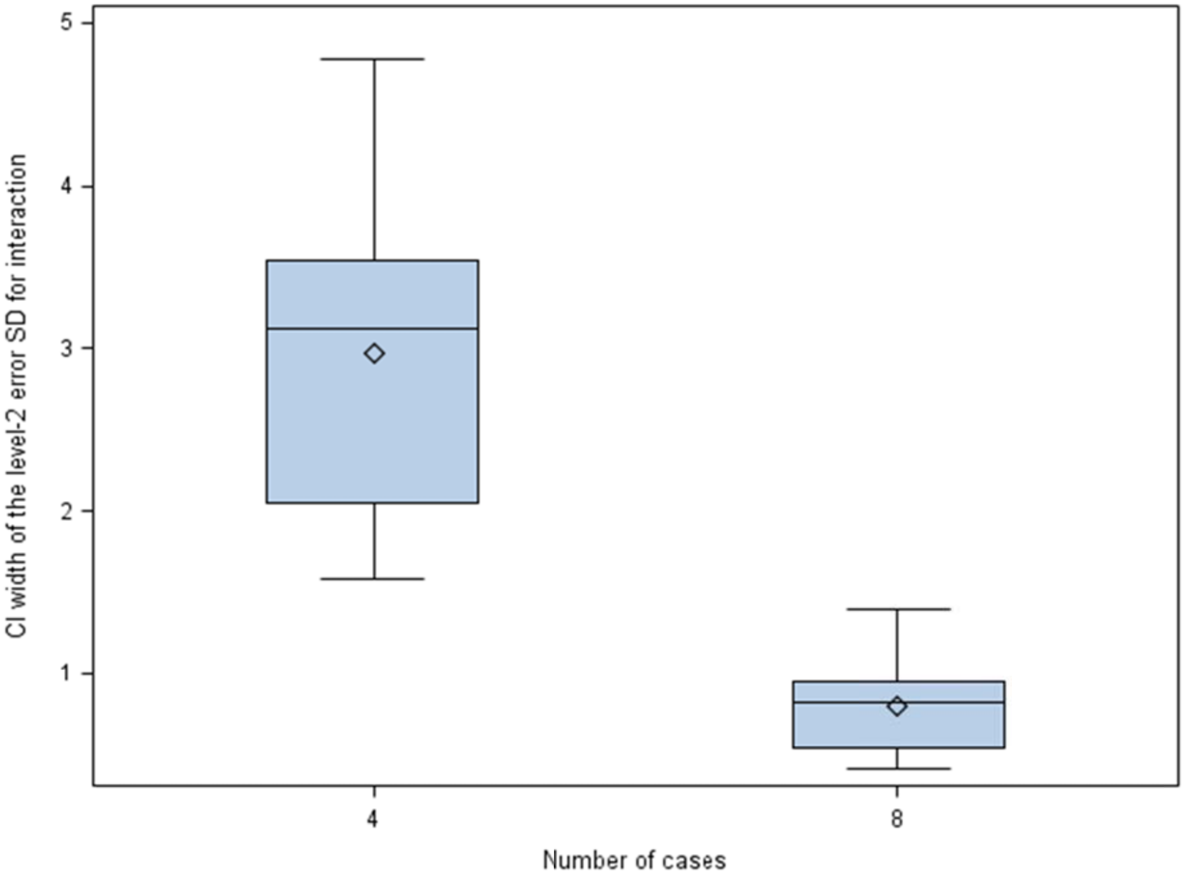


Figure 103. Box plots depicting the estimated CI width of the level-2 error standard deviation for shift in slope as a function of the number of cases.

As illustrated in Figure 103, as the number of cases increased from 4 to 8, the average CI width decreased from *M* = 2.97, *SD* = 0.99 to *M* = 0.80, *SD* = 0.29. Similarly, as the variation in the level-2 errors shifted from most of the variance at the level-1 error (.5) to most of the variance at the level-2 error (2), the average CI width increased from *M* = 1.48, *SD* = 1.06 to *M* = 2.30, *SD* = 1.43 (Figure 104). In addition, as the series length per case increased from 10 to 20, the average CI width decreased from *M* = 2.23, *SD* = 1.48 to *M* = 1.55, *SD* = 1.05 (Figure 105).

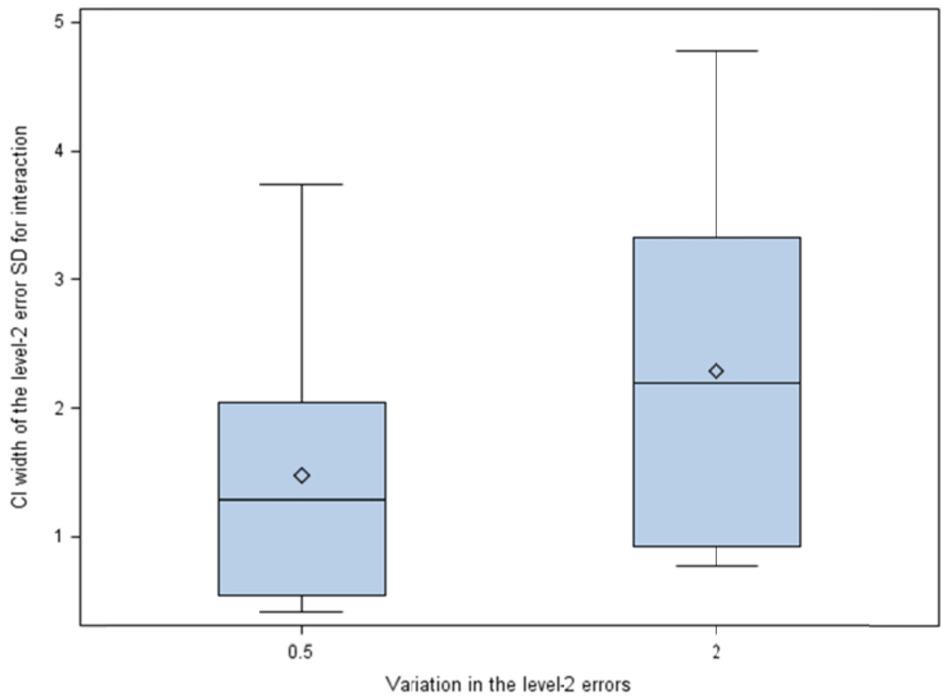


Figure 104. Box plots depicting the estimated CI width of the level-2 error standard deviation for shift in slope as a function of the variation in the level-2 errors.

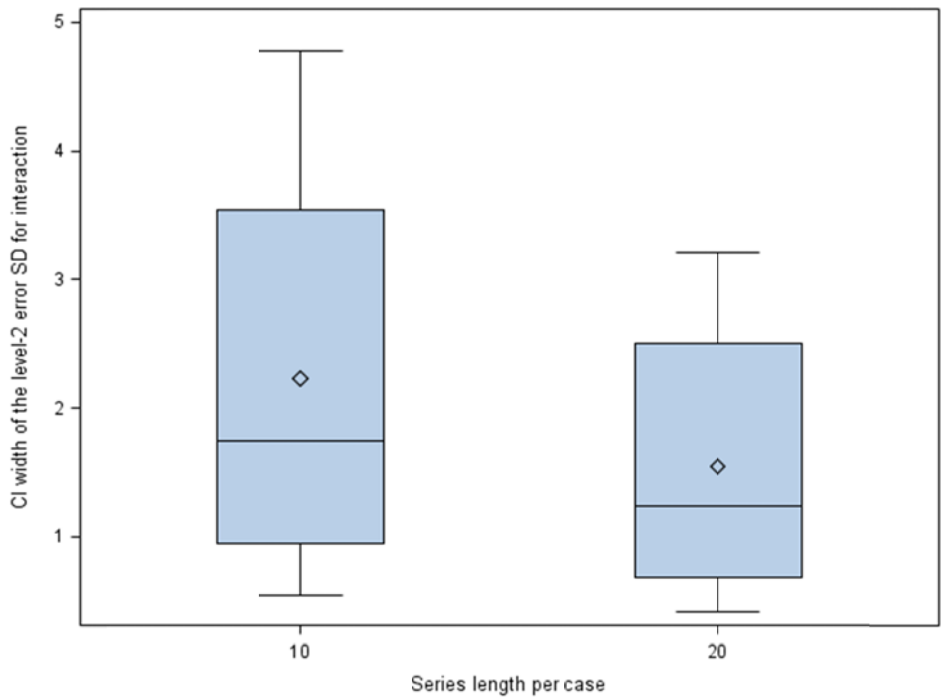


Figure 105. Box plots depicting the estimated CI width of the level-2 error standard deviation for shift in slope as a function of the series length per case.

Level-1 error standard deviation. The average credible interval (CI) width values of the level-1 error standard deviation were different across the two models (Model 1 and Model 2) (Figure 106). The CI width for Model 1 was smaller than the CI width for Model 2. The average CI width for Model 1 was $M = 0.47$, $SD = 0.17$, and the average CI width for Model 2 was $M = 0.81$, $SD = 0.22$. The type of model explained a large amount of the variability ($\eta^2 = .4$).

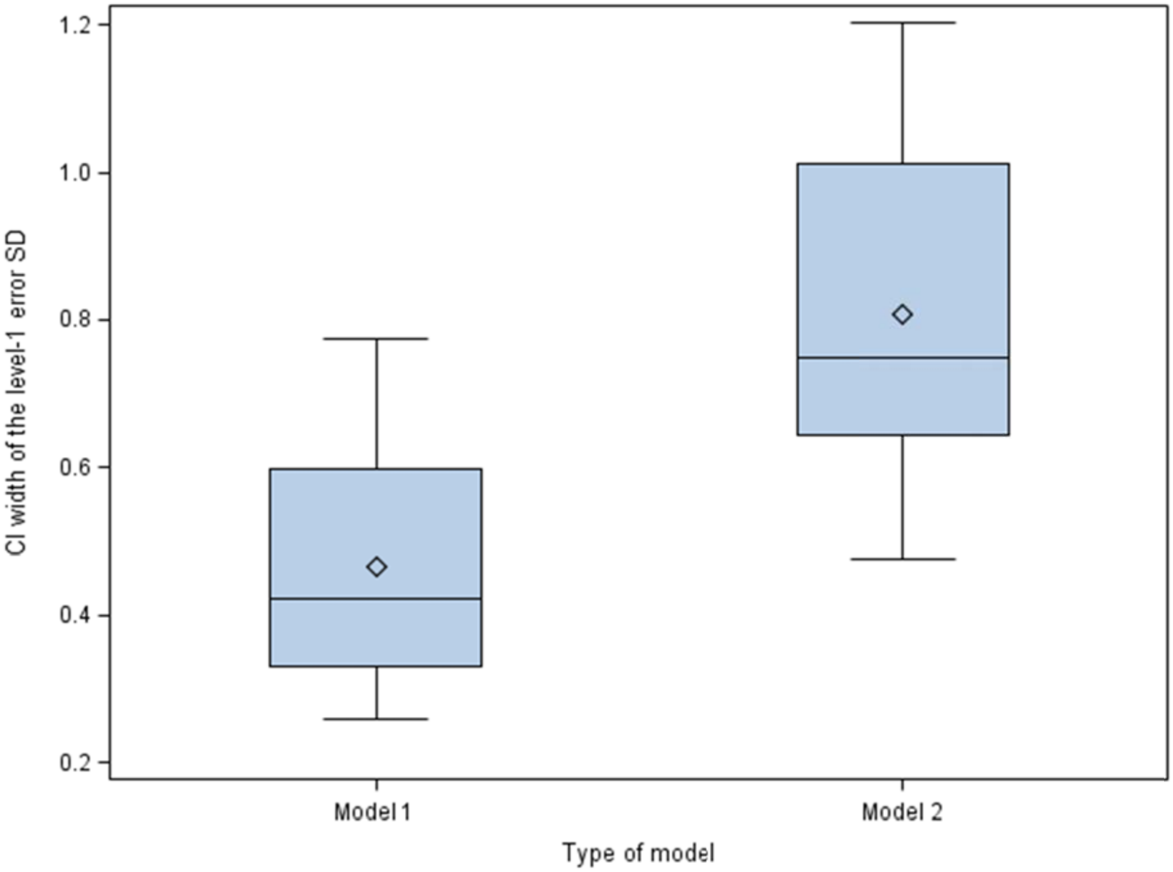


Figure 106. Box plots illustrating the distribution for the CI width of the level-1 error standard deviation across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI width values of the level-1 error standard deviation across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 107). The average CI width were different across the two models within the three true level-1 error structures, and there were little differences across the true level-1 error structures, with a small amount of the variability explained by the different types of the true level-1 error structures (η^2 = .02). The smallest average CI width difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2$-$M_1|$ = .28**)**, and the biggest average CI width difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2$-$M_1|$ = .41). Generally, the CI width for Model 1 was smaller than the CI width for Model 2.

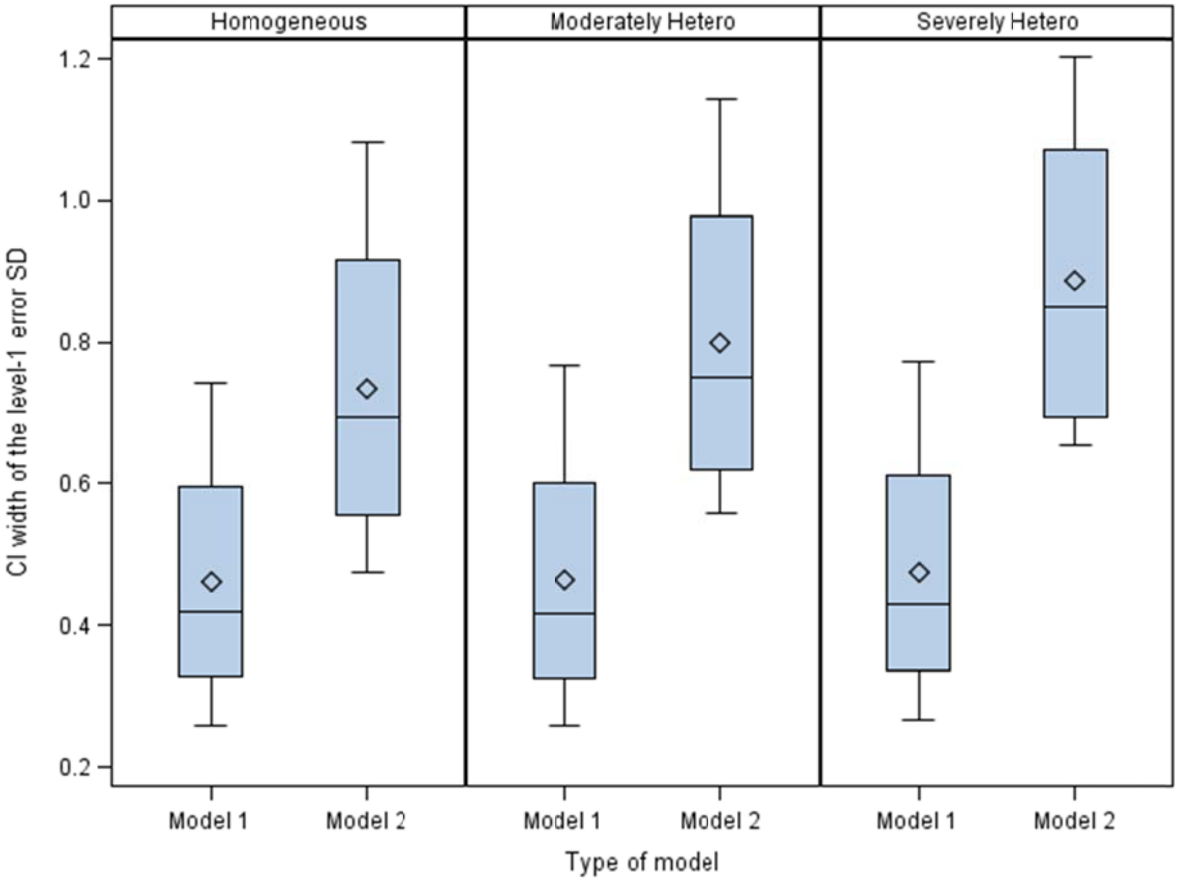


Figure 107. Box plots illustrating the distribution for the CI width of the level-1 error standard deviation across the two models within the three true level-1 error structures.

In order to further explore the variability in the CI width of the level-1 error standard deviation, a GLM model was run. The main effects only model explained 96% of the variability. It was found that three of the design factors had a medium or large effect, including the type of model (η^2 = .44), the number of cases (η^2 = .38), and the series length per case (η^2 = .11). These main effects are illustrated in Figures 108 through 110.

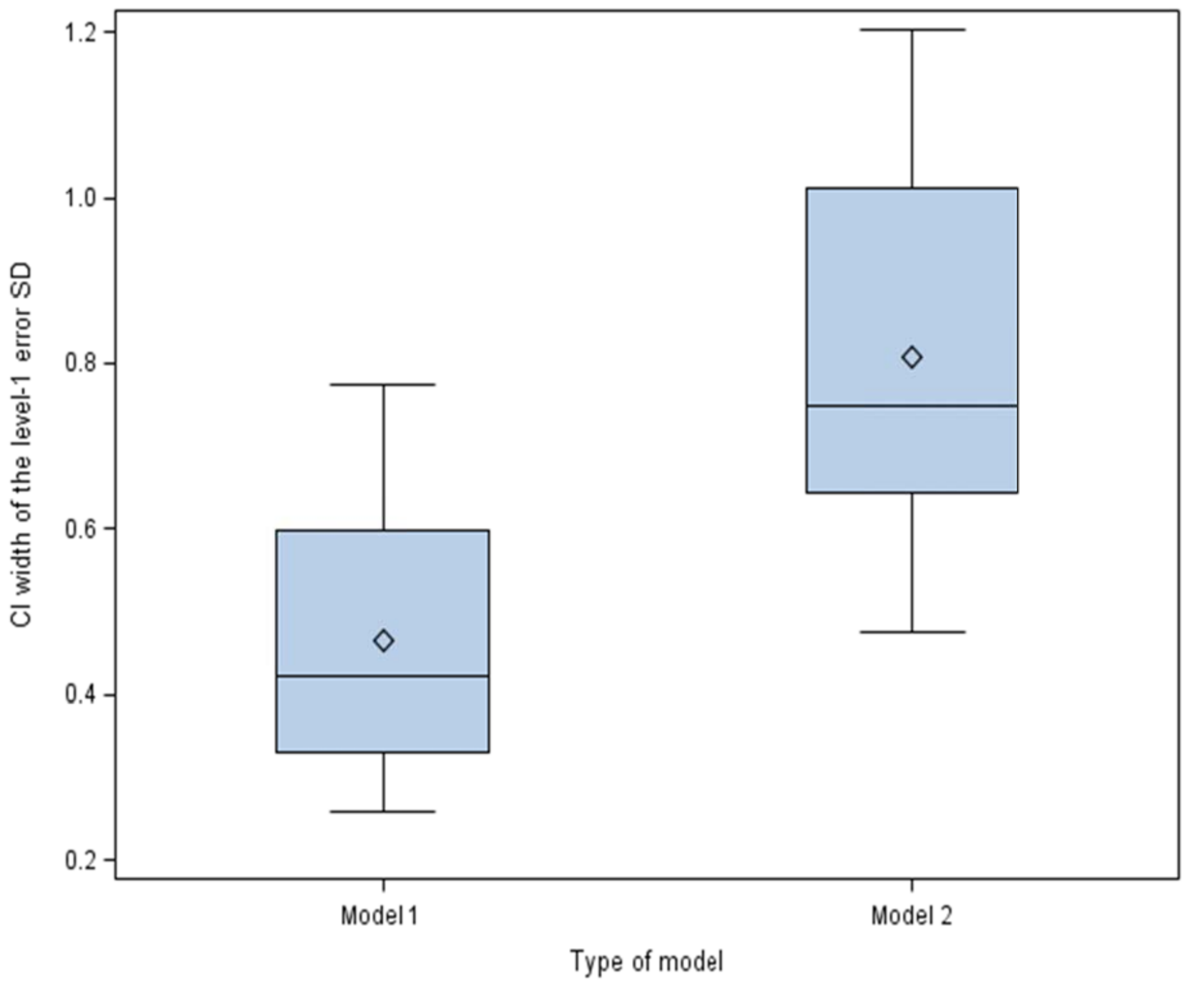


Figure 108. Box plots depicting the estimated CI width of the level-1 error standard deviation as a function of the type of model.

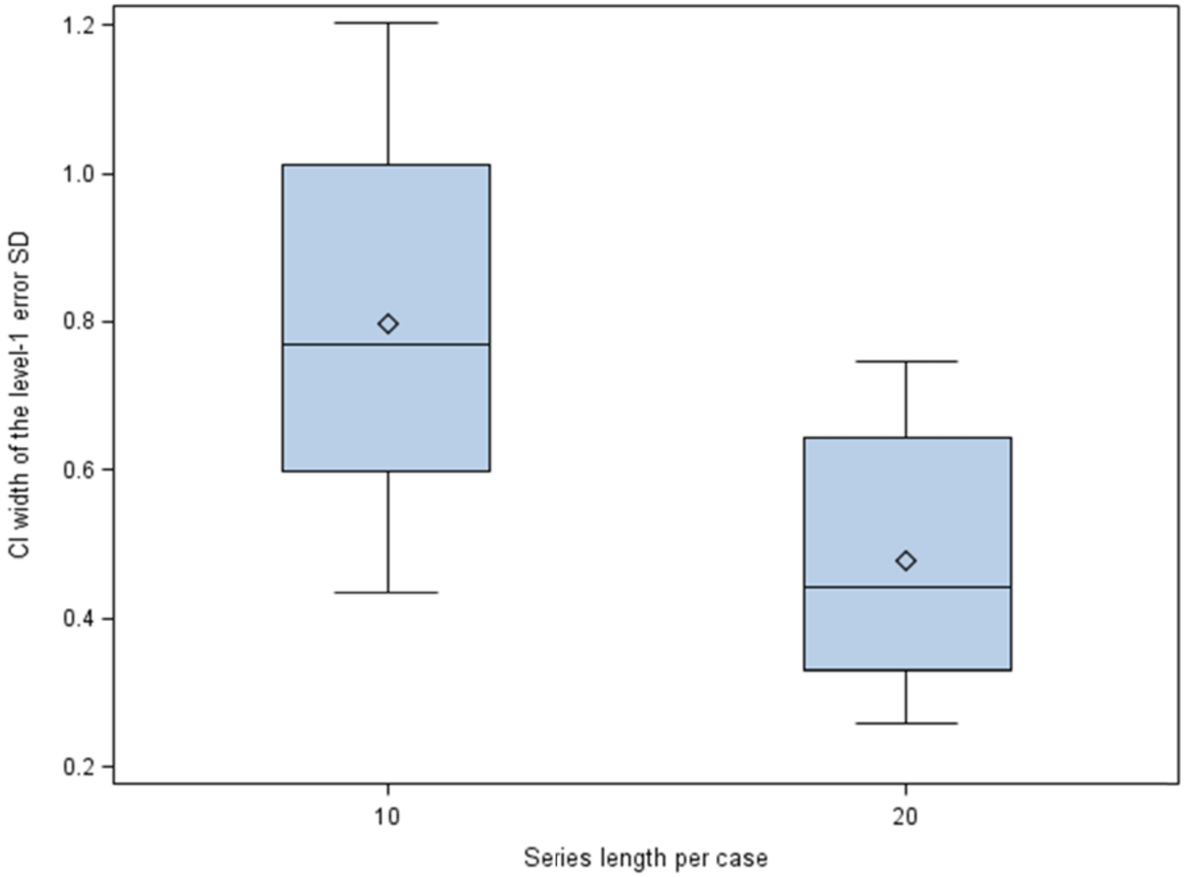


Figure 109. Box plots depicting the estimated CI width of the level-1 error standard deviation as a function of the series length per case.

As illustrated in Figure 108, the average CI coverage for Model 1 was smaller than the average CI coverage for Model 2 (Model 1: *M* = 0.47, *SD* = 0.17; Model 2: *M* = 0.81, *SD* = 0.22. Figure 109 portrays that as the series length per case increased from 10 to 20, the average CI width decreased from *M* = 0.80, *SD* = 0.24 to *M* = 0.48, *SD* = 0.17. Similarly, as the number of cases increased from 4 to 8, the average CI width decreased from *M* = 0.72, *SD* = 0.26 to *M* = 0.55, *SD* = 0.23 (Figure 110).

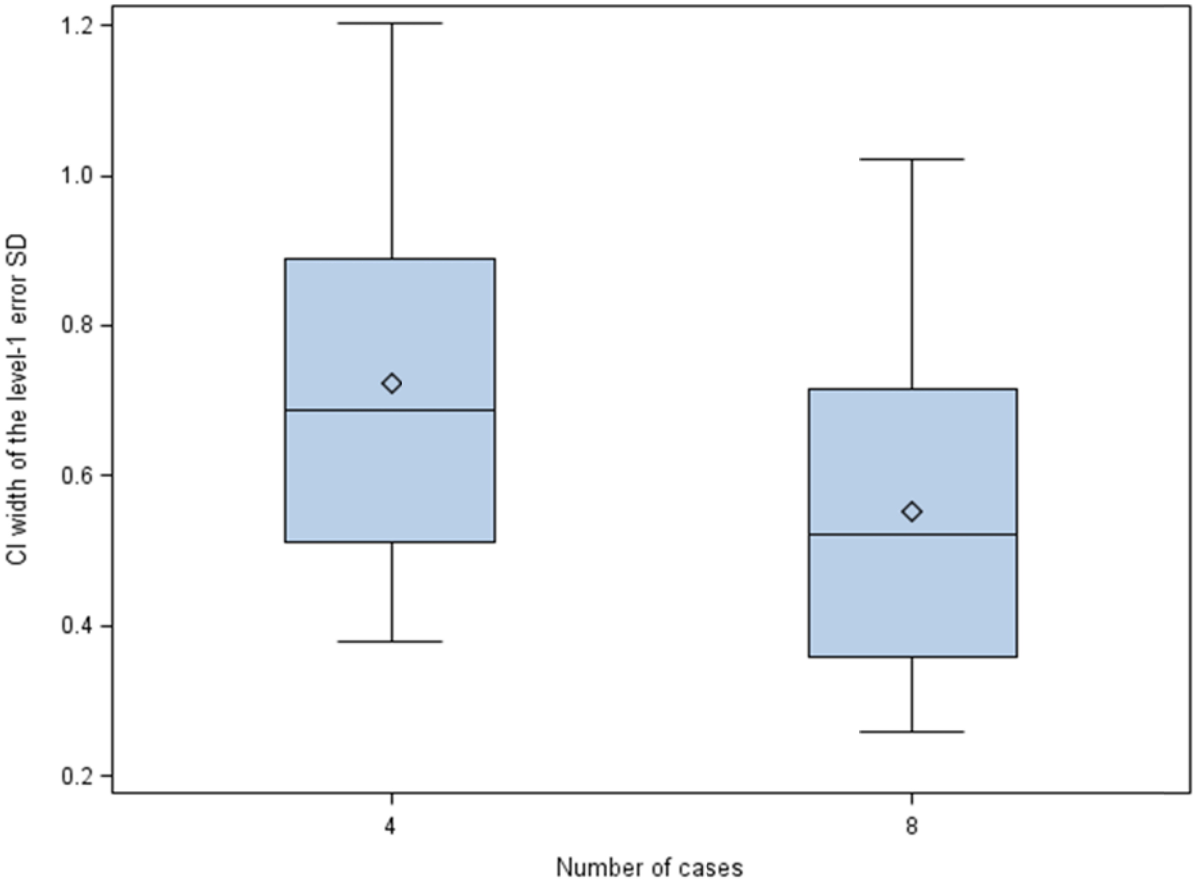


Figure 110. Box plots depicting the estimated CI width of the level-1 error standard deviation as a function of the number of cases.

Autocorrelation. The average credible interval (CI) width values of the autocorrelation were different across the two models (Model 1 and Model 2) (Figure 111). The average CI width for Model 1 was smaller than the average CI width for Model 2. The average CI width for Model 1 was M = 0.74, SD = 0.27, and the average CI width for Model 2 was M = 1.10, SD = 0.24. The type of model explained a large amount of the variability (η^2 =.34).

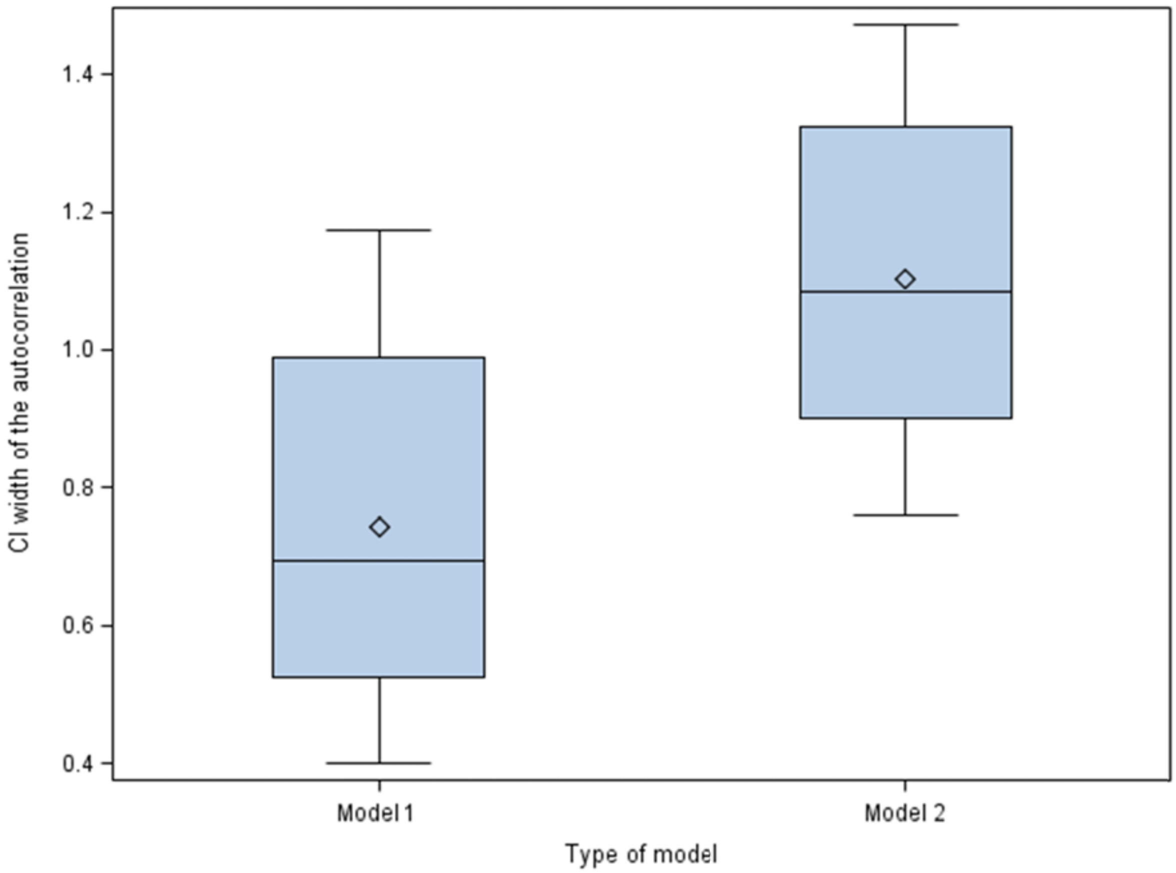


Figure 111. Box plots illustrating the distribution for the CI widths of the autocorrelation across Model 1 which did not model between case variation, and Model 2 which models between case variation.

The average CI width values of the autocorrelation across the two models were also examined within the three different types of true level-1 error structures (homogeneous, moderately heterogeneous, and severely heterogeneous) (Figure 112). The average CI widths were different across the two models within the three true level-1 error structures, and there were no or little differences across the true level-1 error structures, with little of the variability explained by the different types of the true level-1 error structures (η^2 = .0003). The smallest average CI width difference between the two models was found when the true level-1 error structure was homogeneous ($|M_2\text{-}M_1|$ = .32), and the biggest average CI width difference between the two models was found when the true level-1 error structure was severely heterogeneous ($|M_2\text{-}M_1|$ = .39).

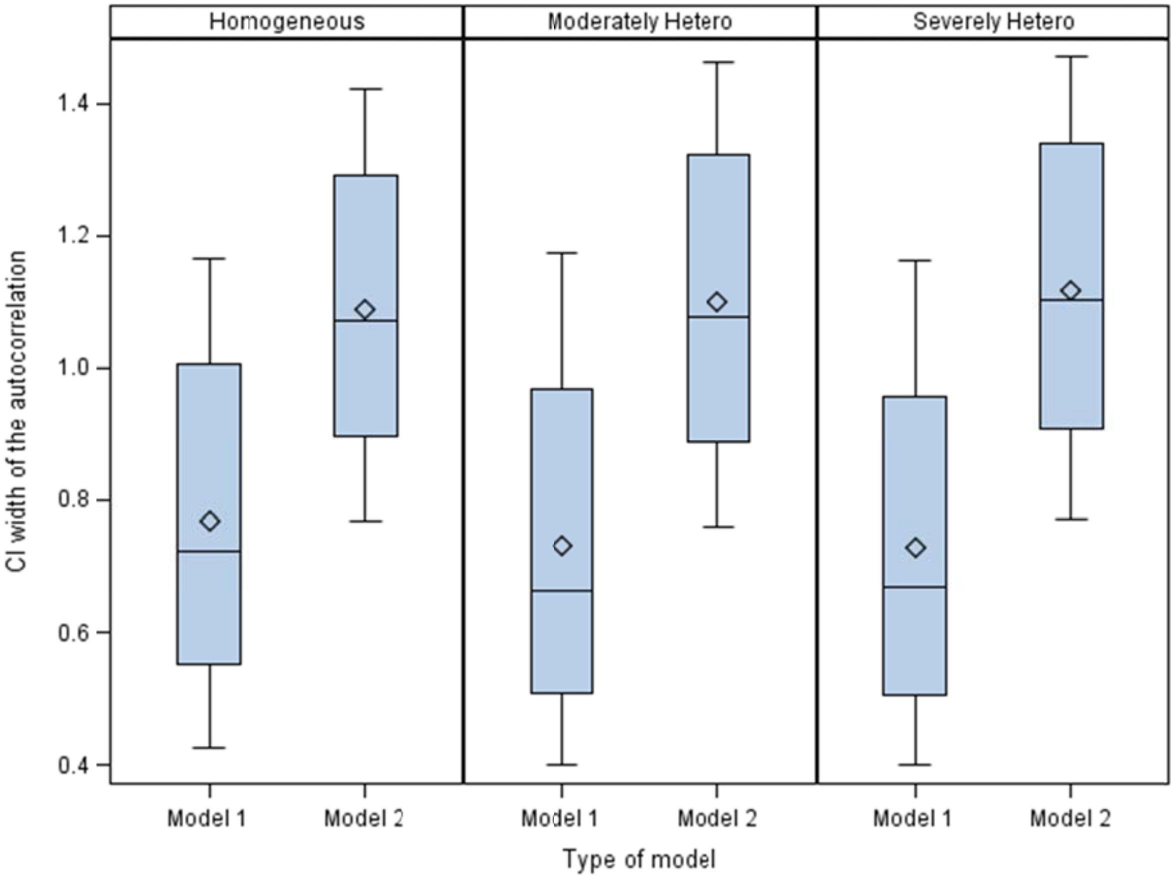


Figure 112. Box plots illustrating the distribution for the CI width of the autocorrelation across the two models within the three true level-1 error structures.

In order to further explore the variability in the CI width for the autocorrelation, a GLM model was run. The main effects only model explained 99% of the variability. It was found that three of the design factors had a large effect, including the series length per case (η^2 = .47), the type of model (η^2 = .34), and the number of cases (η^2 = .17). These main effects are illustrated in Figures 113 through 115.

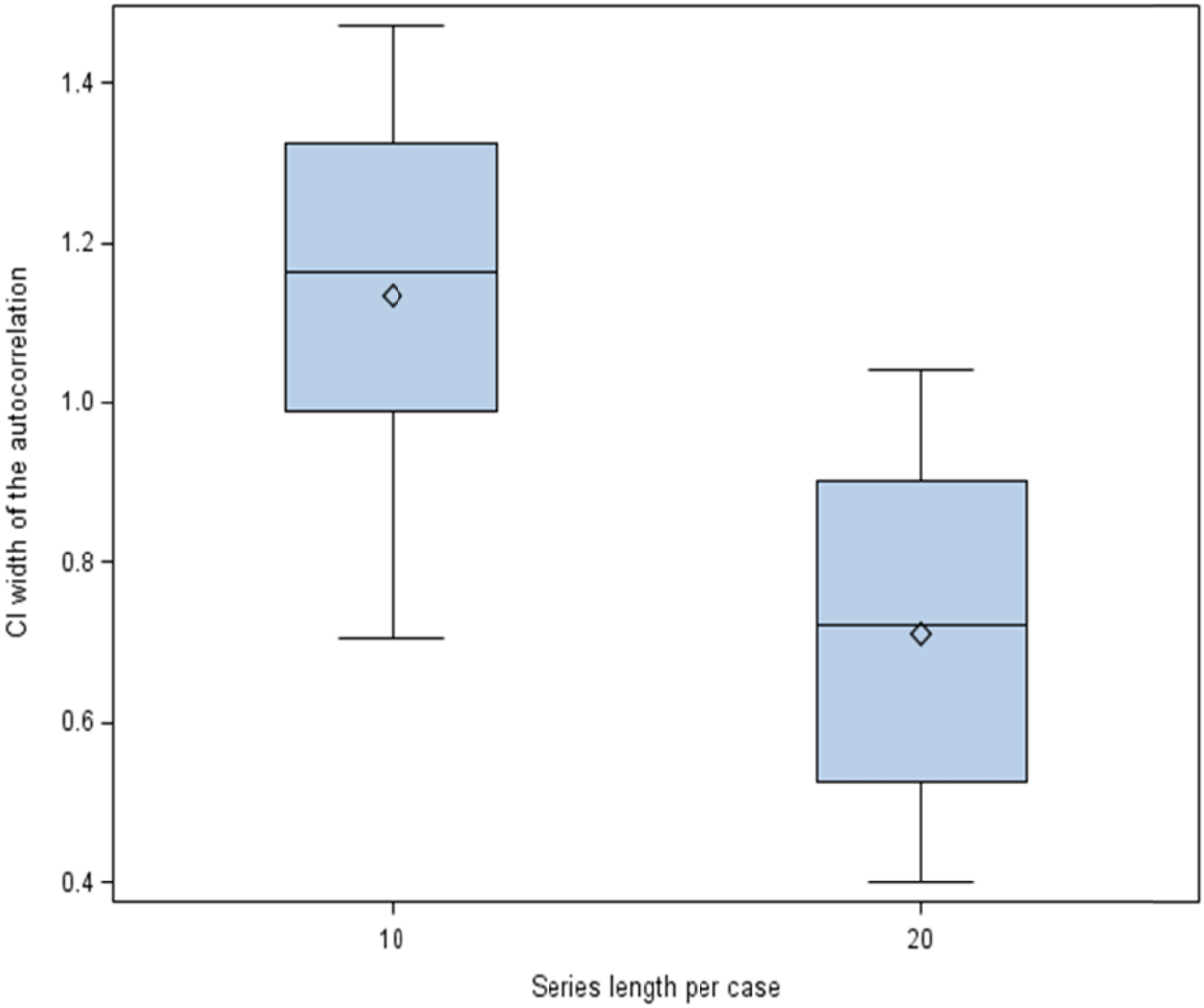


Figure 113. Box plots depicting the estimated CI width of the autocorrelation as a function of the series length per case.

As illustrated in Figure 113, as the series length per case increased from 10 to 20, the average CI width decreased from *M* = 1.14, *SD* = 0.24 to *M* = 0.71, *SD* = 0.22. Figure 114 portrays that the average CI coverage for Model 1 was smaller than the average CI coverage for Model 2 (Model 1: *M* = 0.74, *SD* = 0.27; Model 2: *M* = 1.10, *SD* = 0.24). Similarly, as the number of cases increased from 4 to 8, the average CI width decreased from *M* = 1.05, *SD* = 0.30 to *M* = 0.80, *SD* = 0.28 (Figure 115).

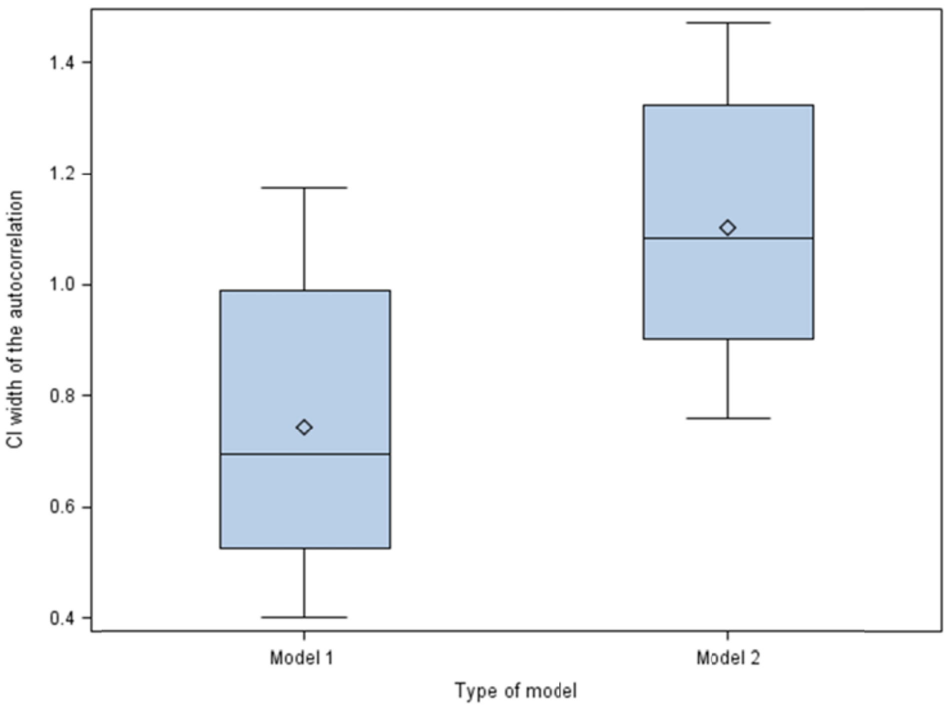


Figure 114. Box plots depicting the estimated CI width of the autocorrelation as a function of the type of model.

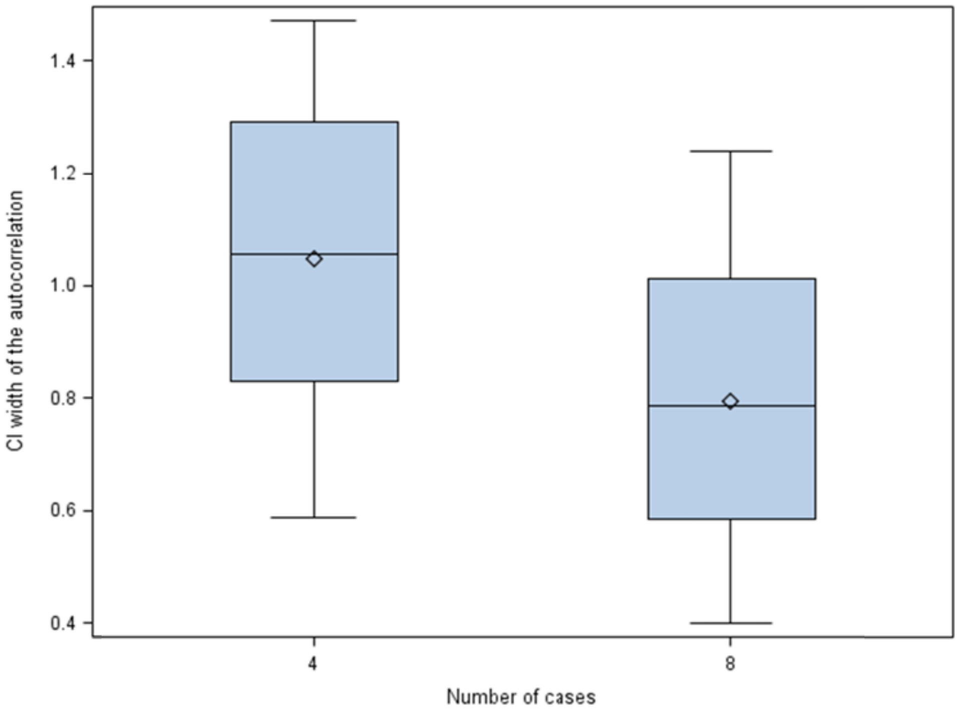


Figure 115. Box plots depicting the estimated CI width of the autocorrelation as a function of the number of cases.

Types of Specifications

The results of the fixed treatment effects and variance components were also looked by three different types of specifications in the level-1 error structure: under-specified (i.e., Model 1 when the data were generated to be heterogeneous), correctly-specified (i.e., Model 1 when the data were generated to be homogeneous, or Model 2 when the data were generated to be heterogeneous), and over-specified (i.e., Model 2 when the data were generated to be homogeneous).

Fixed Treatment Effects

Bias. The average bias values of the treatment effect for phase (shift in level) and the average treatment effect for the interaction effect (shift in slope) by three different types of specifications are illustrated in Figures 116 and 117.

Figure 116 portrays that the average bias values for the shift in level were very similar and close to 0 for all three types of specifications, ranging from 0 to .003. The over-specified models had the smallest variability among the three types of specifications.

Similarly, Figure 117 illustrated that the average bias values for the shift in slope were slightly different, but close to 0 for all three types of specifications, ranging from .001 to .009. The under-specified models had the smallest bias value and variability among the three types of specifications.

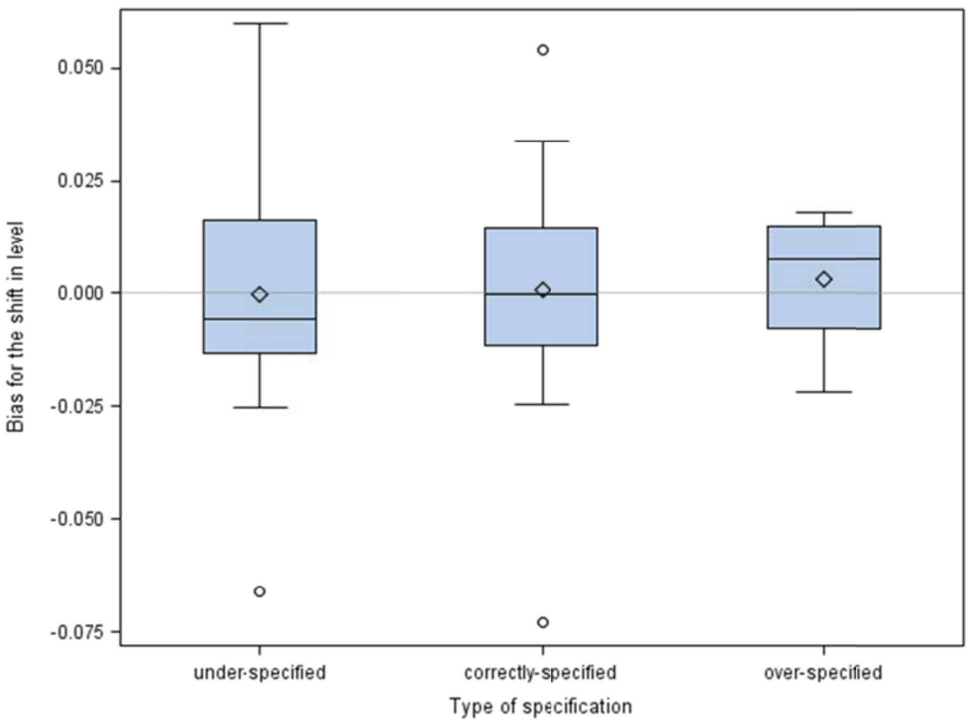


Figure 116. Box plots depicting the estimated bias of the treatment effect for phase across three types of specifications.

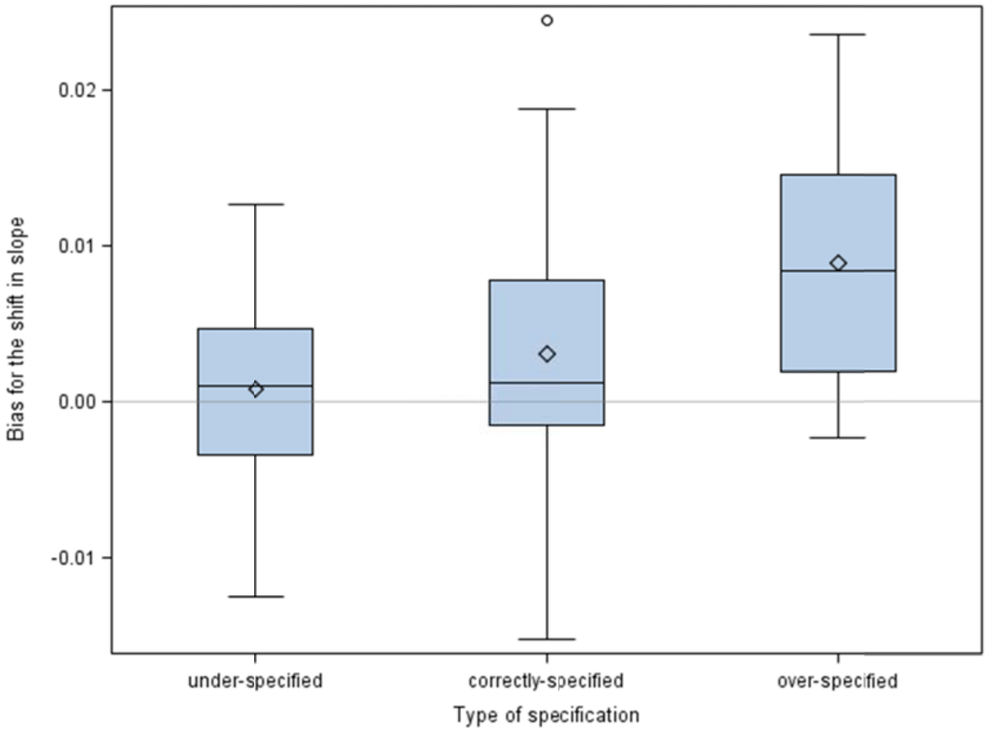


Figure 117. Box plots depicting the estimated bias of the treatment effect for interaction across three types of specifications.

RMSE. The average RMSE values of the treatment effect for phase (shift in level) and the treatment effect for interaction (shift in slope) by three different types of specifications are illustrated in Figures 118 and 119.

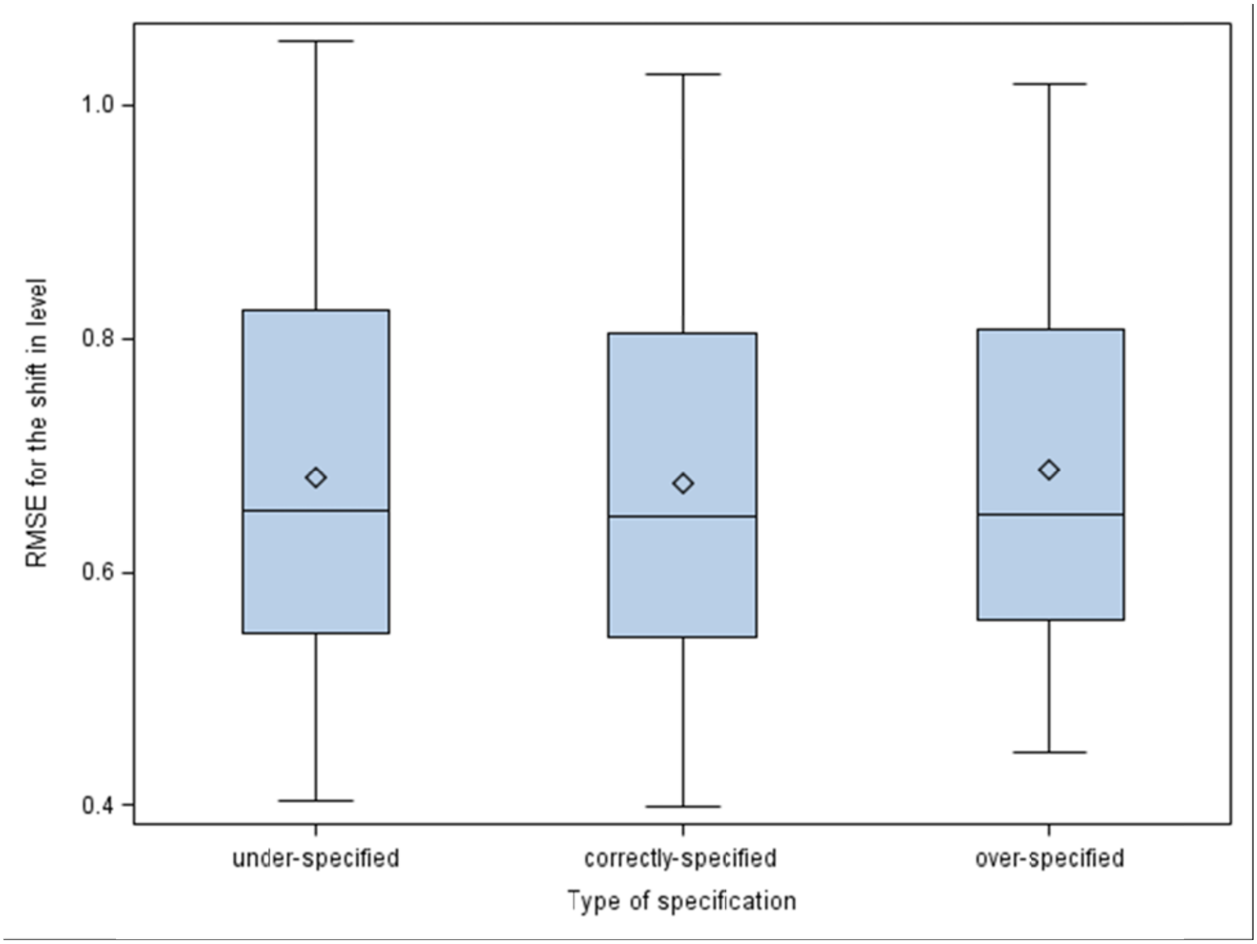


Figure 118. Box plots depicting the estimated RMSE of the treatment effect for phase across three types of specifications.

Figure 118 portrays that the average RMSE values for the shift in level were very similar across all three types of specifications, ranging from .68 to .69.

Similarly, Figure 119 illustrated that the average RMSE values for the shift in slope were very similar across the three types of specifications, ranging from .22 to .23.

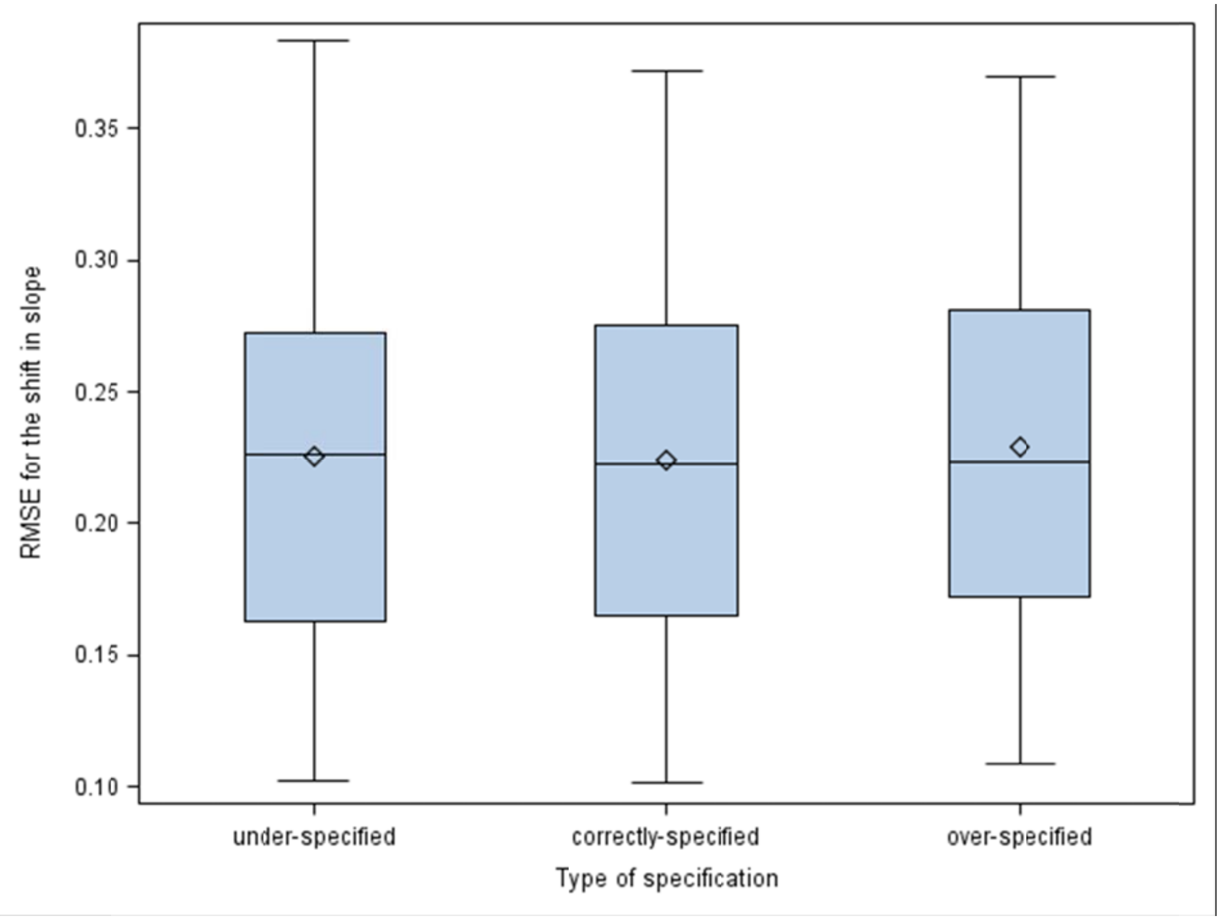


Figure 119. Box plots depicting the estimated RMSE of the treatment effect for the interaction effect across three types of specifications.

CI coverage. The average CI coverage of the treatment effect for phase (shift in level) and the interaction treatment effect (shift in slope) by three different types of specifications are illustrated in Figures 120 and 121.

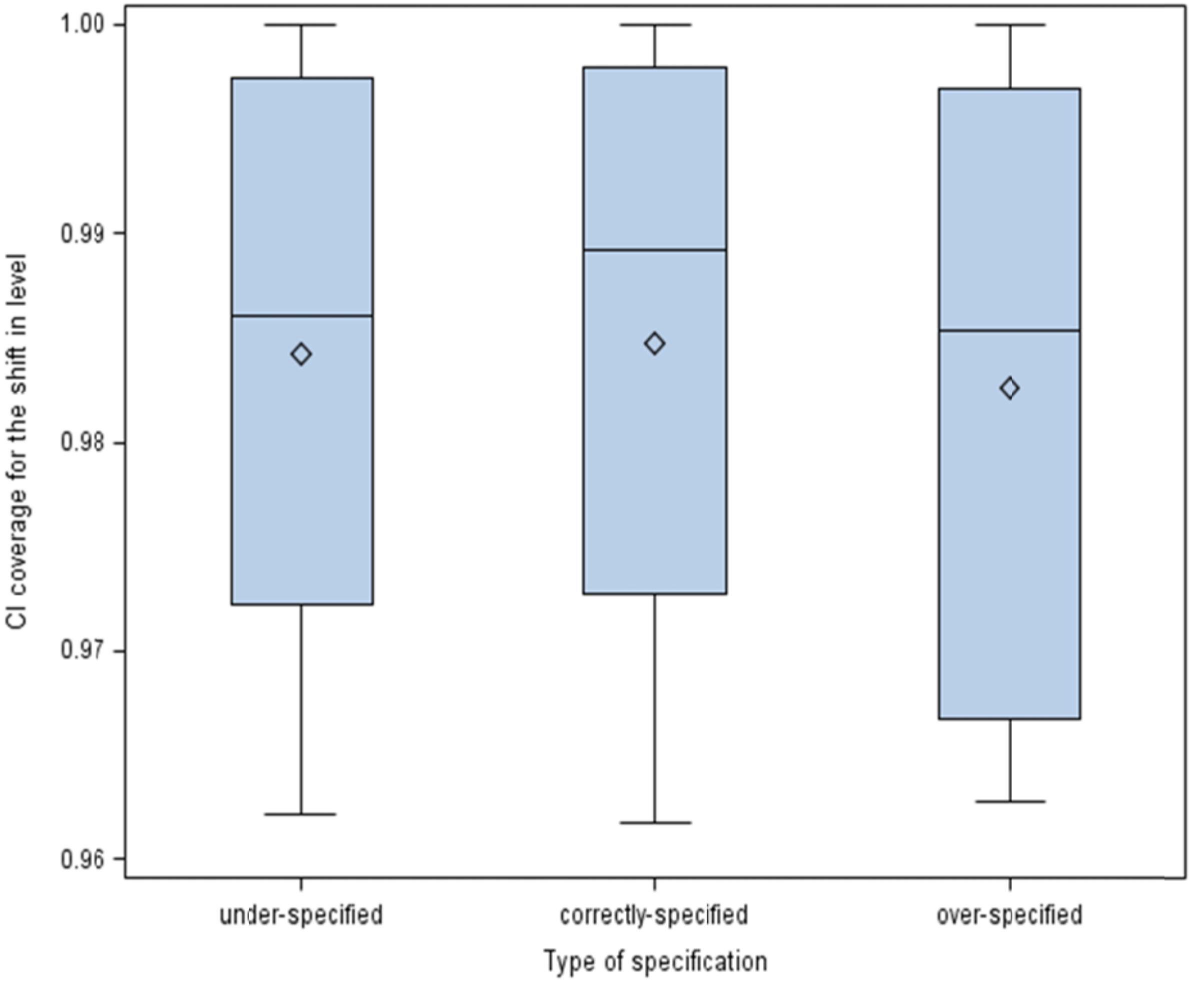


Figure 120. Box plots depicting the estimated CI coverage of the treatment effect for phase across three types of specifications.

Figure 120 portrays that the average CI coverage for the shift in level were similar and tended to over cover across all three types of specifications, ranging from .983 to .985. The over-specified type had more variability than other types of specifications. Similarly, Figure 121 illustrated that the average CI coverage values for the shift in slope were very similar and tended to over cover across the three types of specifications, .985 for all three type of specification.

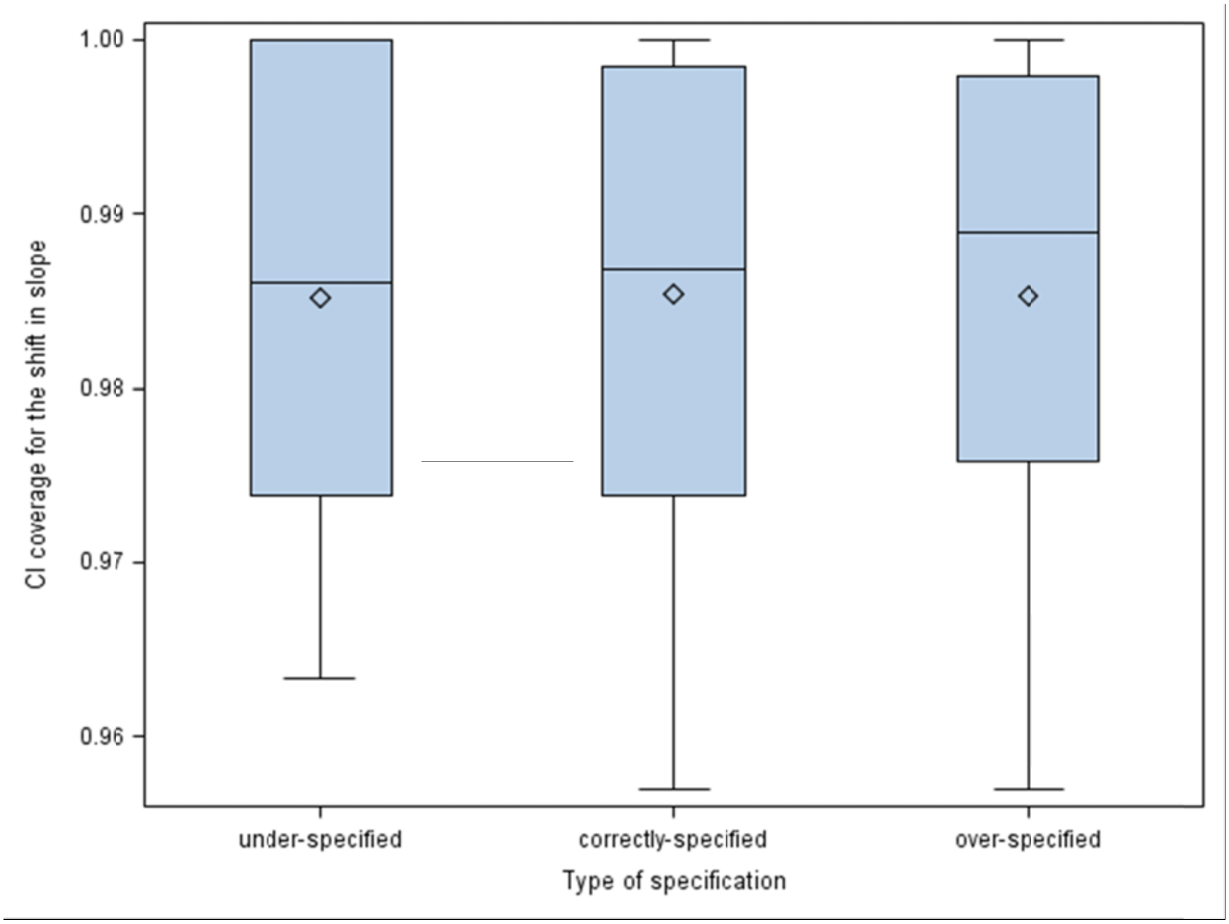


Figure 121. Box plots depicting the estimated CI coverage of the interaction treatment effect across three types of specifications.

CI width. The average CI width of the treatment effect for phase (shift in level) and the interaction treatment effect (shift in slope) by three different types of specifications are illustrated in Figure 122 and 123. Figure 122 portrays that the average CI widths for the shift in level were similar across the three types of specifications, ranging from 4.94 to 4.99. Similarly, Figure 123 illustrated that the average CI widths for the shift in slope were similar across the three types of specifications, ranging from 1.68 to 1.71.

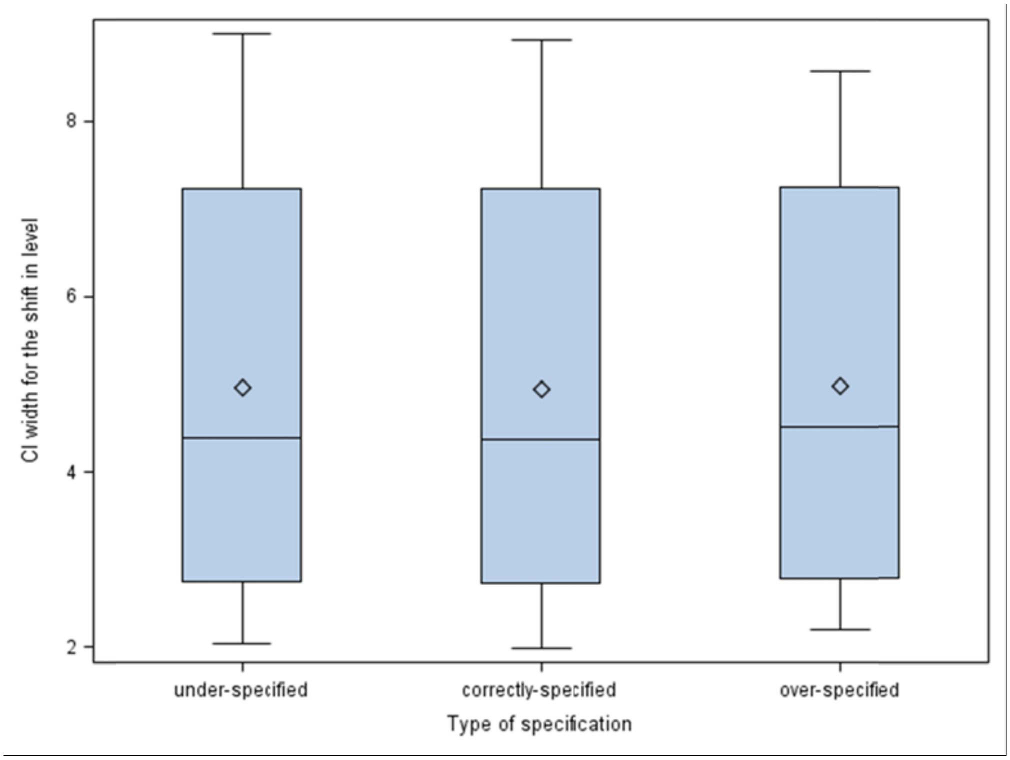


Figure 122. Box plots depicting the estimated CI width of the treatment effect for phase across three types of specifications.

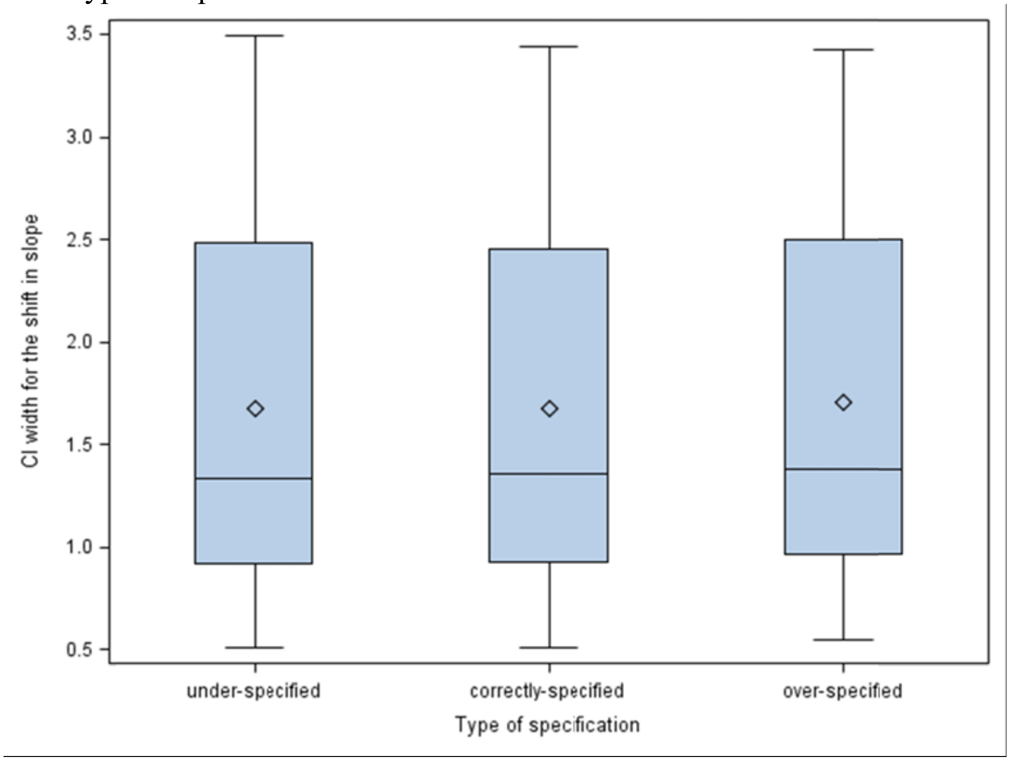


Figure 123. Box plots depicting the estimated CI width of the interaction treatment effect across three types of specifications.

Variance Components

Bias. The average bias values of the level-2 error standard deviation for phase (shift in level) and the level-2 error standard deviation for the interaction (shift in slope) by three different types of specifications are illustrated in Figures 124 and 125.

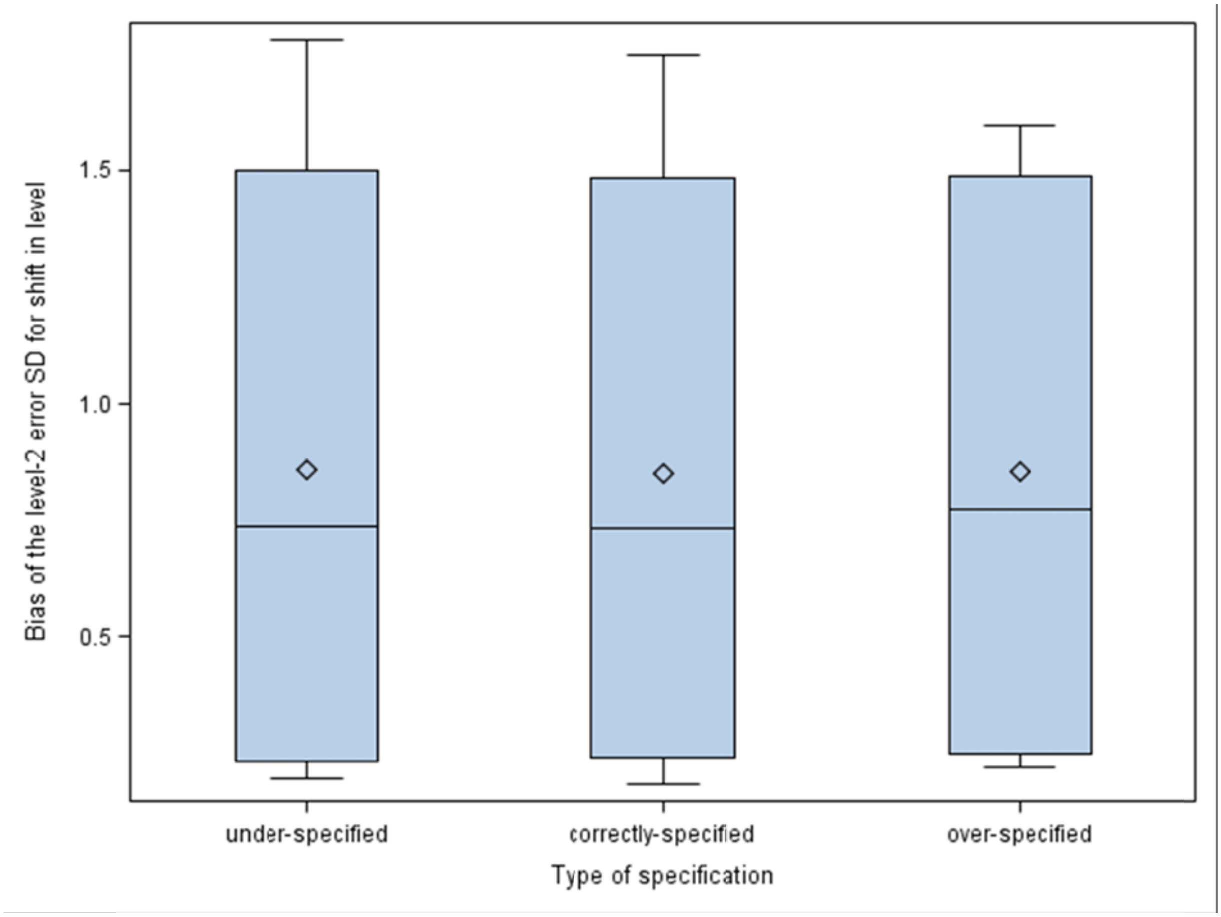


Figure 124. Box plots depicting the estimated bias of the level-2 error standard deviation for phase across three types of specifications.

Figure 124 portrays that the average bias values of the level-2 error standard deviation for the shift in level were very similar and tended to be positively biased for all three types of specifications, ranging from .85 to .86. Similarly, Figure 125 illustrated that the average bias values of the level-2 error standard deviation for the shift in slope were very similar and tended to be positively biased for all three types of specifications, ranging from .30 to .31.

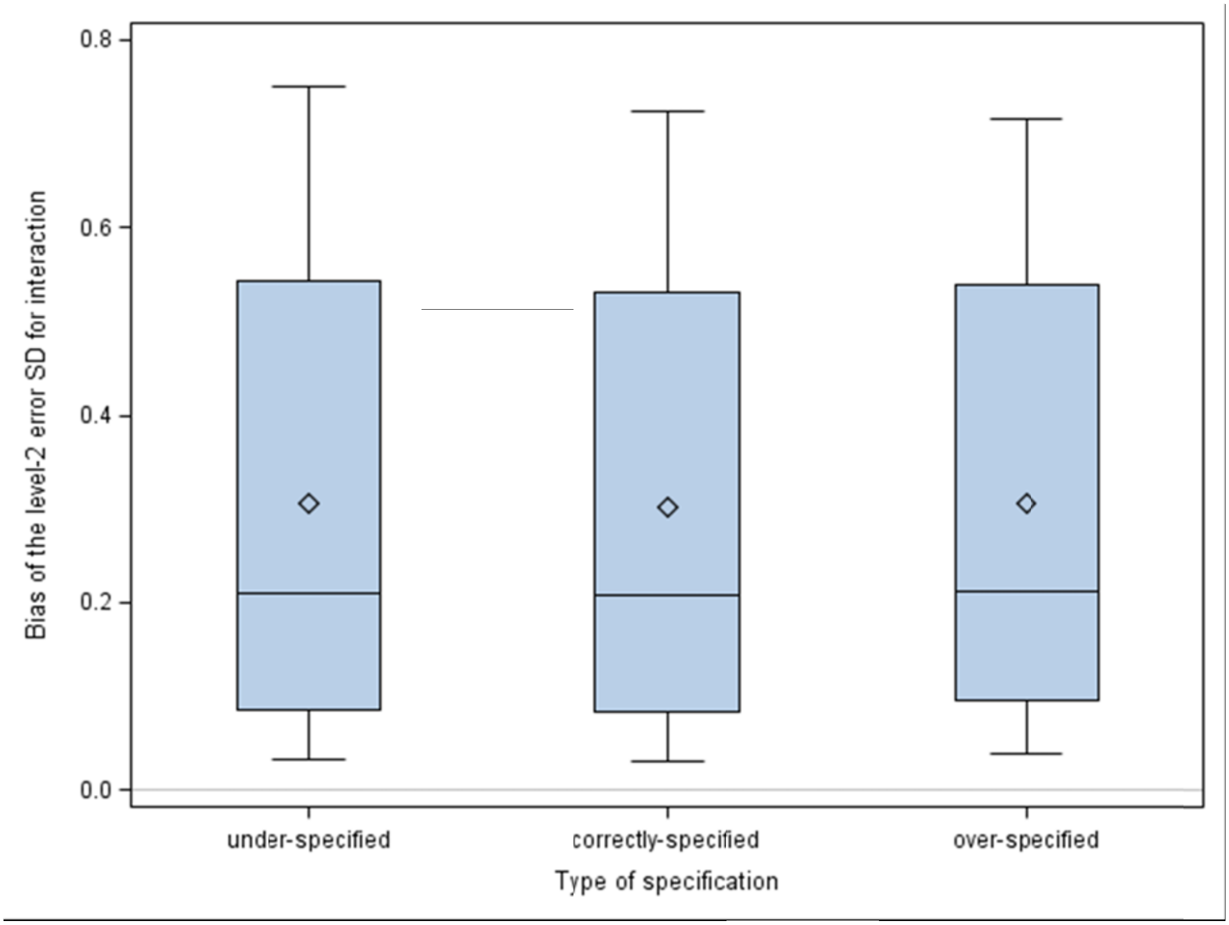


Figure 125. Box plots depicting the estimated bias of the level-2 error standard deviation for the interaction across three types of specifications.

The average bias values of the level-1 error variance by three different types of specifications are illustrated in Figures 126 and 127.

Figure 126 portrays that the average bias values of the level-1 error standard deviation were different and tended to be positively biased for all three types of specifications, ranging from .034 to .056. The over-specified type had the smallest bias and variability, and the under-specified type had the largest bias among the three types of specifications. For further examination of this difference, the average bias values of the level-1 error standard deviation by the three different types of specifications as a function of the true level-1 error structure were also investigated in Figure 127. When under-specified, the average bias values of the level-1 error standard deviation increased as the degree of heterogeneity increased (moderately hetero: *M* =.04; severely hetero: *M* =.07). When correctly-specified, and the true level-1 error structure was homogeneous, it had the smallest average bias value of the level-1 error standard deviation (homogeneous: *M* =.03; moderately hetero: *M* =.05; severely hetero: *M* =.05).

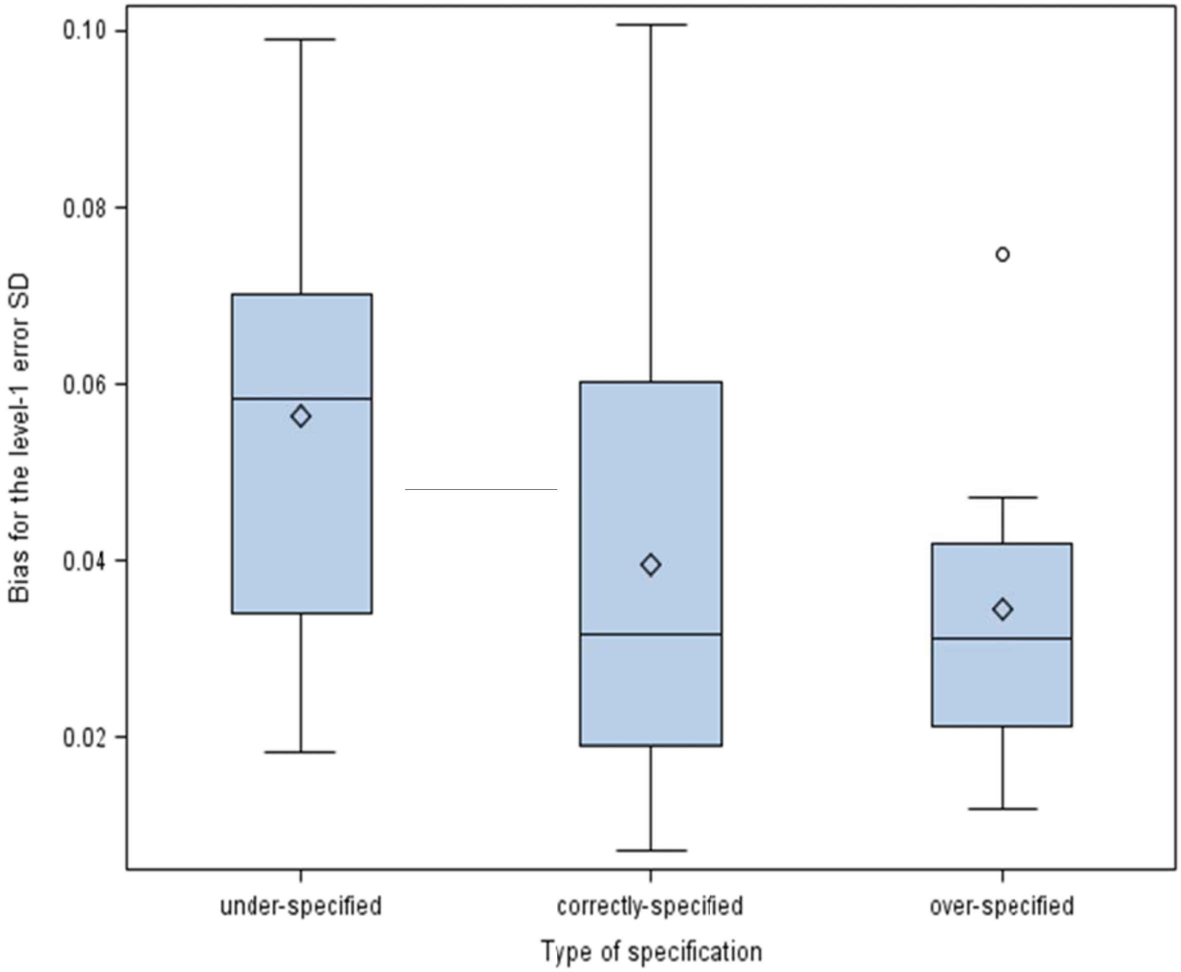


Figure 126. Box plots depicting the estimated bias of the level-1 error standard deviation across three types of specifications.

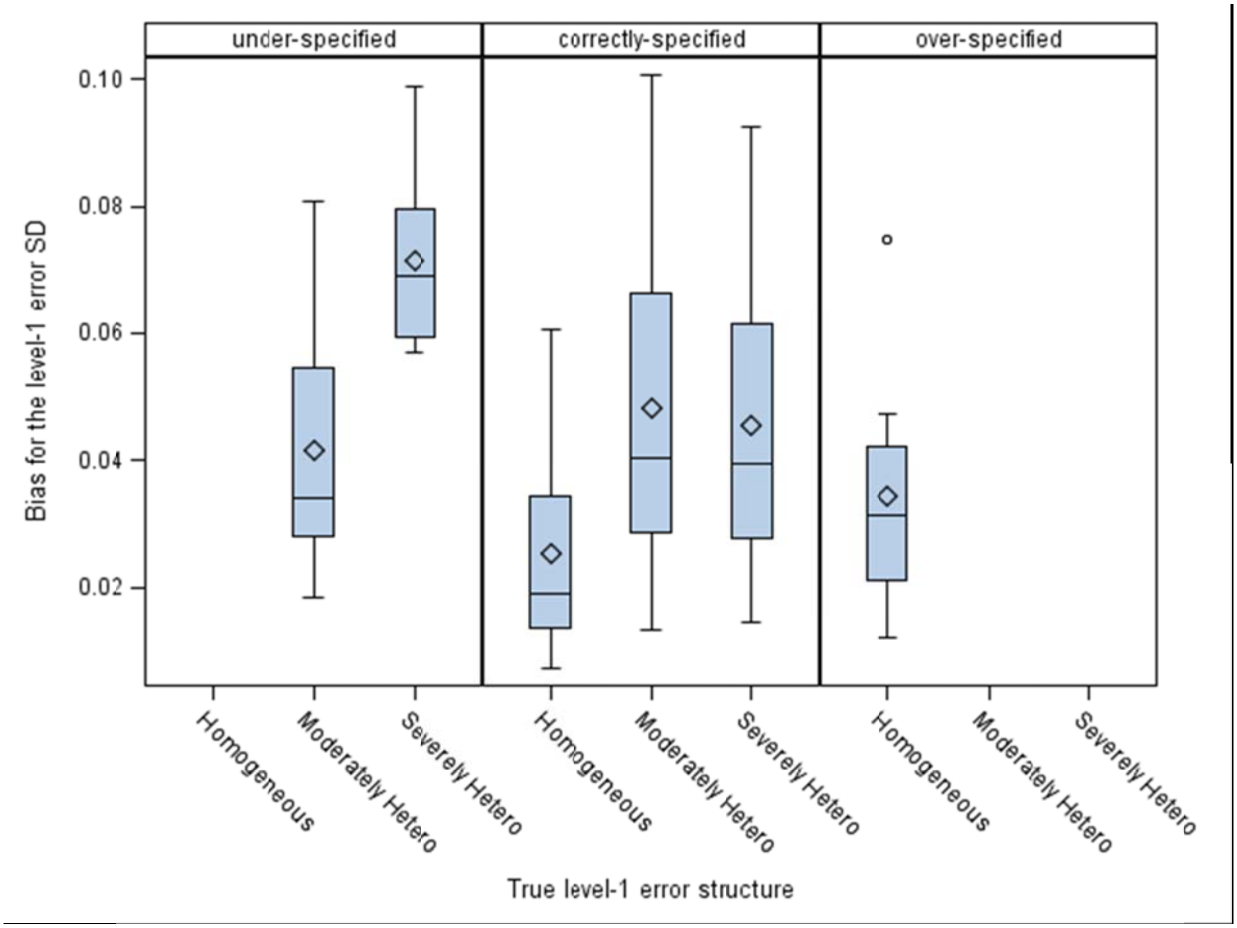


Figure 127. Box plots depicting the estimated bias of the level-1 error standard deviation across three types of specifications as a function of the true level-1 error structure.

The average bias values of the autocorrelation by three different types of specifications are illustrated in Figures 128 and 129.

Figure 128 portrays that the average bias values of the autocorrelation were different and tended to be negatively biased for the both under-specified and correctly-specified (*M* = -0.16, *M* = -0.09, respectively) models, but close to 0 for the over-specified models (*M* = 0.01). The over-specified models had the smallest variability, and the correctly-specified models had the largest variability among the three types of specifications. For further examination of this difference, the average bias values of the autocorrelation by the three different types of specifications as a function of the true level-1 error structure were plotted in Figure 129. When under-specified, the average bias values of the autocorrelation were similar across the moderately and the severely heterogeneous error structures (moderately hetero: *M* =-.16; severely hetero: *M* =-.17). When correctly-specified, the average bias values of the autocorrelation were substantially smaller when the true level-1 error structure was homogeneous than one of the heterogeneous error structures (homogeneous: *M* =.02; moderately hetero: *M* =-.14; severely hetero: *M* =-.14).

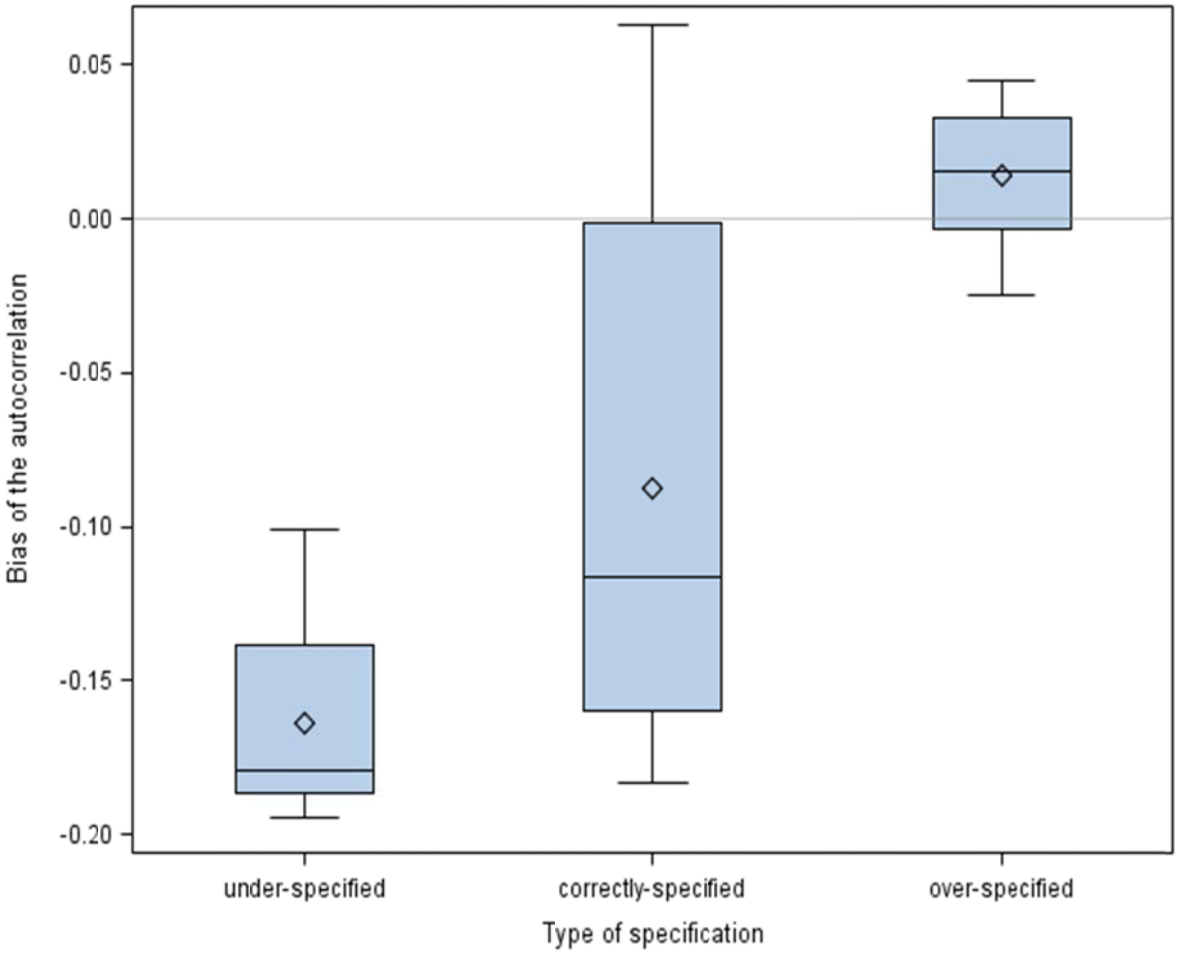


Figure 128. Box plots depicting the estimated bias of the autocorrelation across three types of specifications.

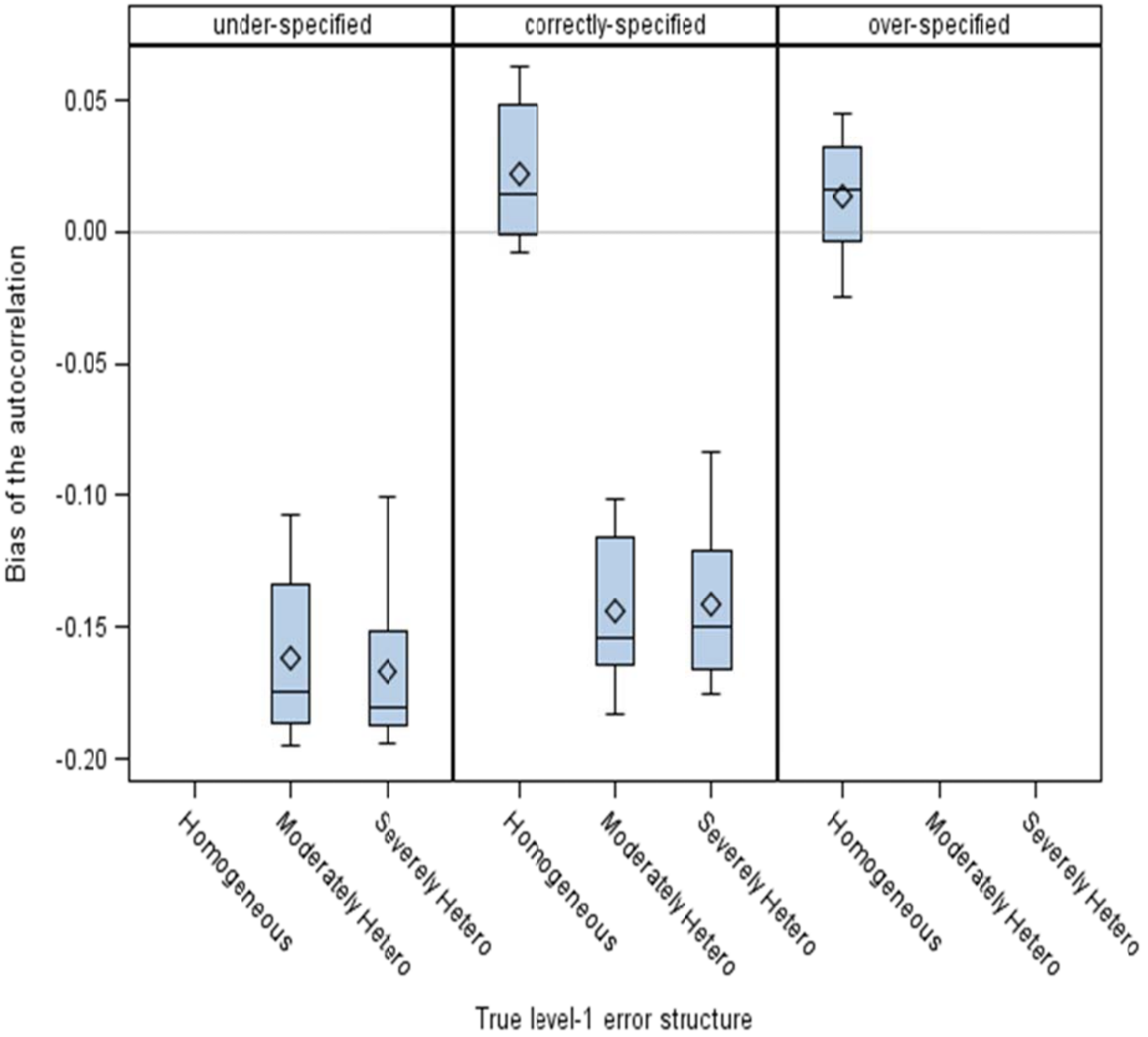


Figure 129. Box plots depicting the estimated bias of the autocorrelation across three types of specifications as a function of the true level-1 error structure.

RMSE. The average RMSE values of the level-2 error standard deviation for phase (shift in level) and the level-2 error standard deviation for the interaction (shift in slope) by three different types of specifications are illustrated in Figures 130 and 131.

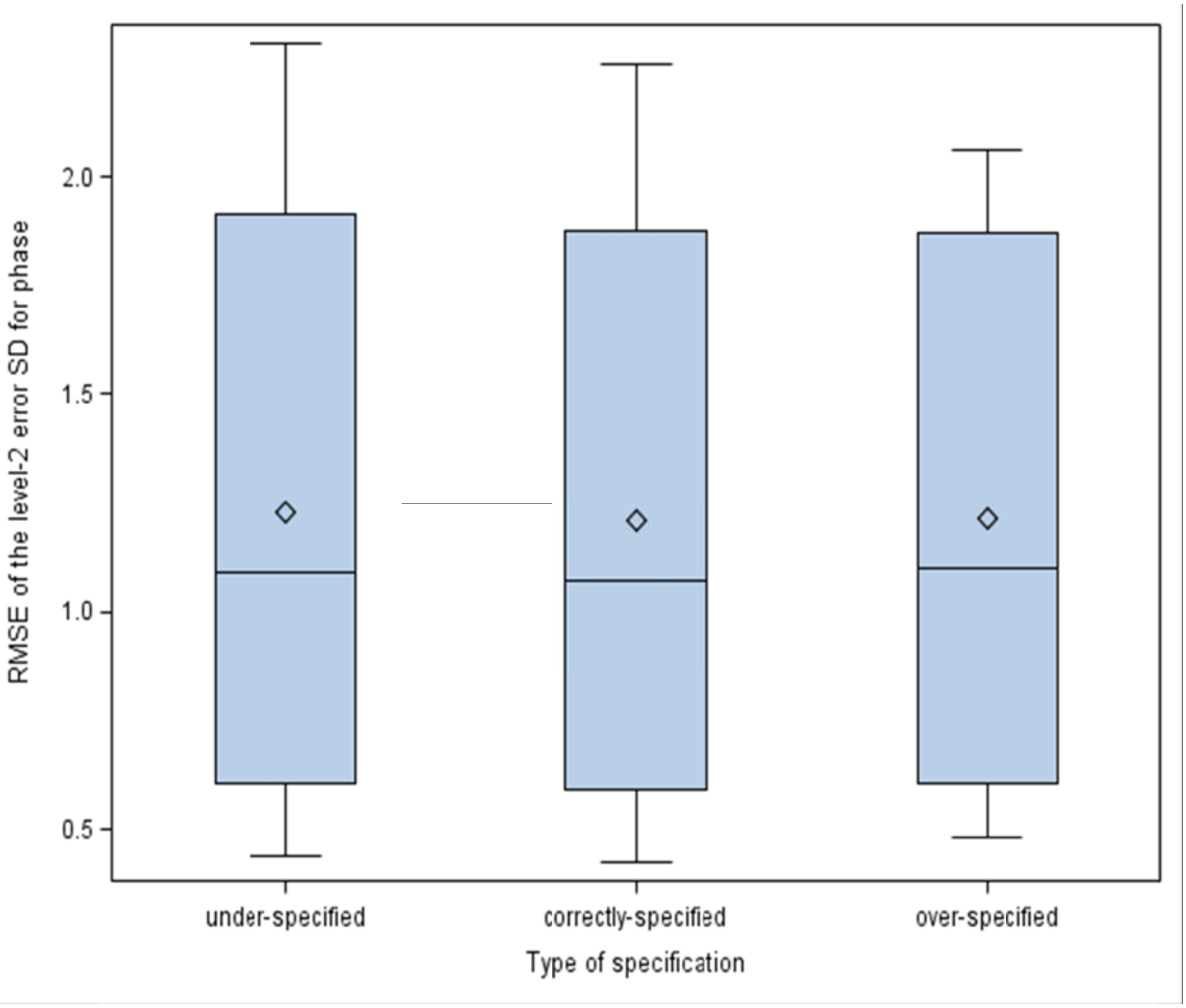


Figure 130. Box plots depicting the estimated RMSE of the level-2 error standard deviation for phase across three types of specifications.

Figure 130 portrays that the average RMSE values of the level-2 error standard deviation for shift in level were very similar across all three types of specifications, ranging from 1.21 to 1.23. Similarly, Figure 131 illustrated that the average RMSE values of the level-2 error standard deviation for shift in slope were very similar across the three types of specifications, ranging from .41 to .42.

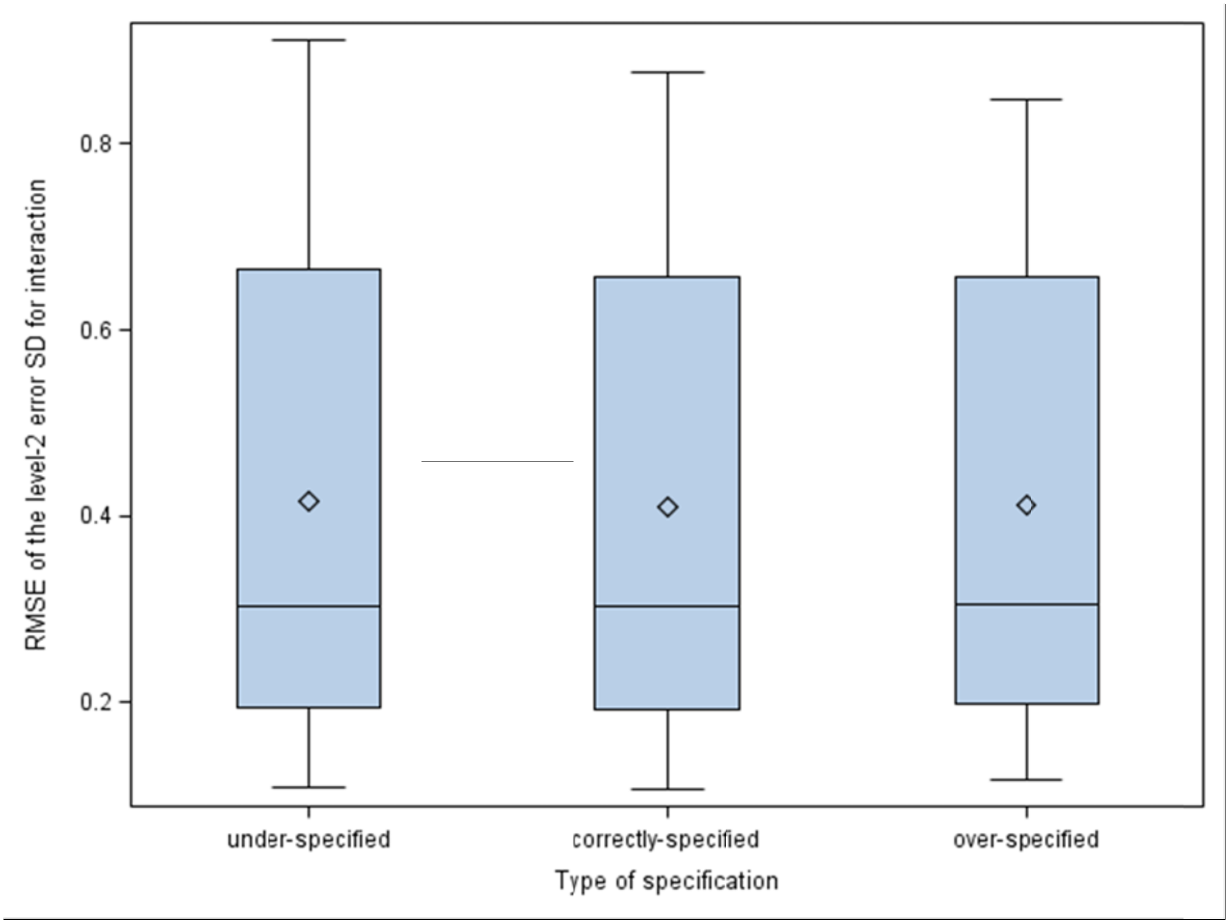


Figure 131. Box plots depicting the estimated RMSE of the level-2 error standard deviation for interaction across three types of specifications.

The average RMSE values of the level-1 error standard deviation by three different types of specifications are illustrated in Figures 132 and 133.

Figure 132 portrays that the average RMSE values of the level-1 error standard deviation were different, ranging from .14 to .27. The over-specified type had the smallest average RMSE value and variability, and the under-specified type had the largest average RMSE value and variability among the three types of specifications. For further examination of this difference, the average RMSE values of the level-1 error standard deviation by the three different types of specifications as a function of the true level-1 error structure were provided in Figure 133. In the under-specified conditions, the average RMSE values of the level-1 error standard deviation increased as the degree of heterogeneity of the level-1 error structure increased (moderately hetero: *M* =.20; severely hetero: *M* =.35). In the correctly-specified conditions, when the true level-1 error structure was homogeneous, it had the smallest average RMSE values of the level-1 error standard deviation (homogeneous: *M* =.11; moderately hetero: *M* =.18; severely hetero: *M* =.22).

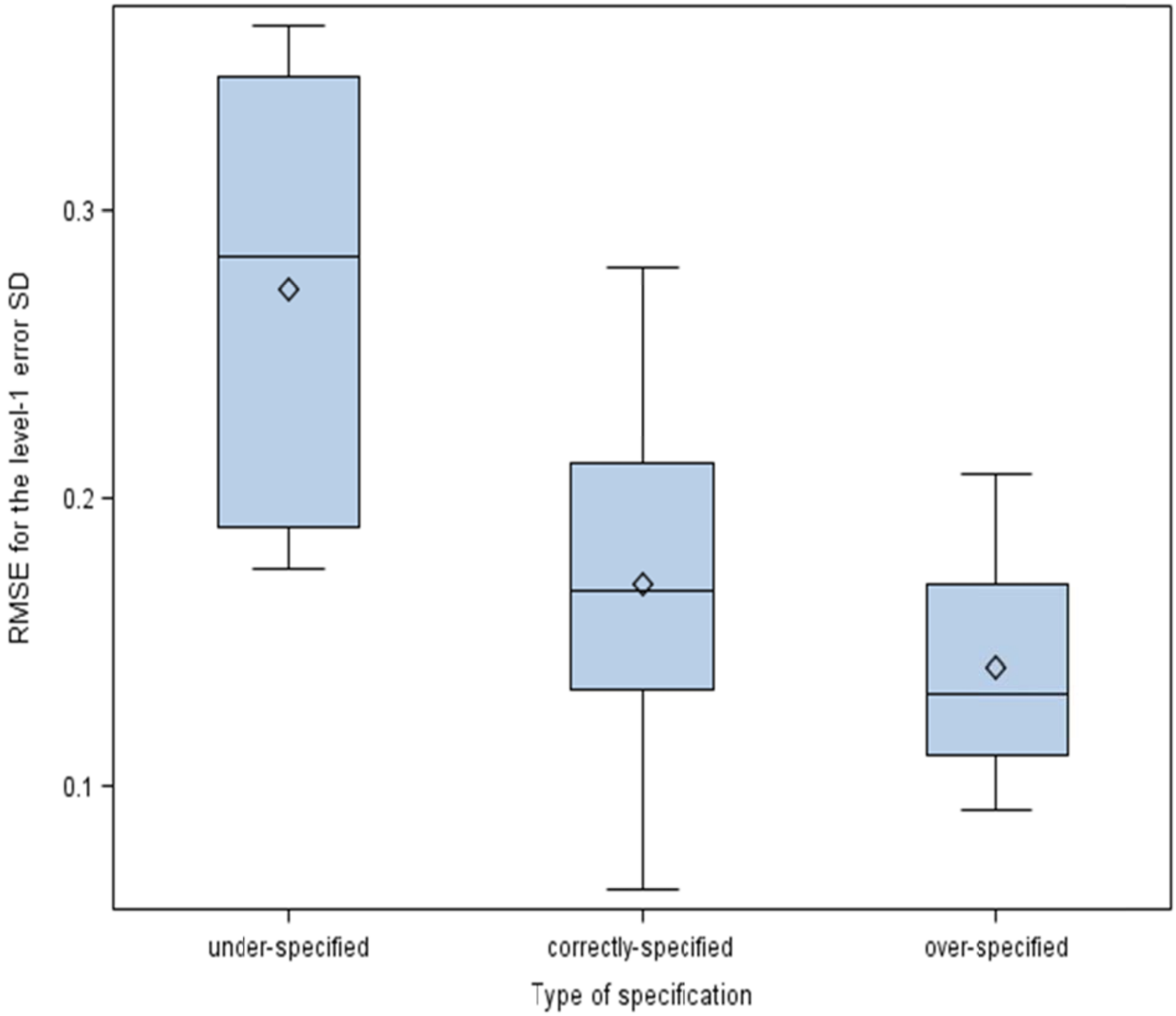


Figure 132. Box plots depicting the estimated RMSE of the level-1 error standard deviation across three types of specifications.

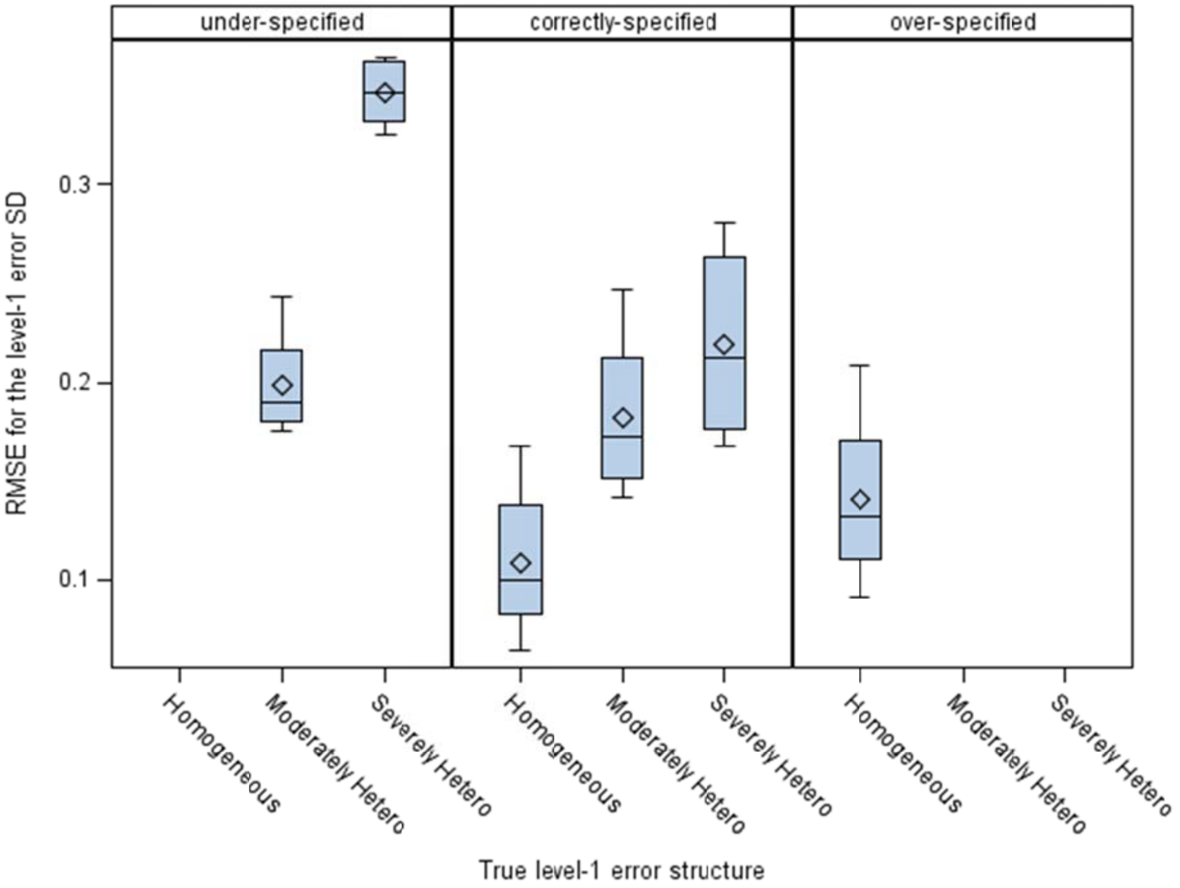


Figure 133. Box plots depicting the estimated RMSE of the level-1 error standard deviation across three types of specifications as a function of the true level-1 error structure.

The average RMSE values of the autocorrelation by three different types of specifications are illustrated in Figures 134 and 135.

Figure 134 portrays that the average RMSE values of the autocorrelation were different, ranging from .18 to .30. The over-specified type had the smallest average RMSE value and variability, and the under-specified type had the largest variability among the three types of specifications.

For further examination of this difference, the average RMSE values of the autocorrelation by the three different types of specifications as a function of the true level-1 error structure were graphed in Figure 135. When the models were under-specified, the average

RMSE values of the autocorrelation increased as the degree of heterogeneity in the level-1 error structure increased (moderately hetero: *M* =.27; severely hetero: *M* =.33). When the models were correctly-specified, and the true level-1 error structure was homogeneous, it had the smallest average RMSE value of the autocorrelation (homogeneous: *M* =.17; moderately hetero: *M* =.26; severely hetero: *M* =.31).

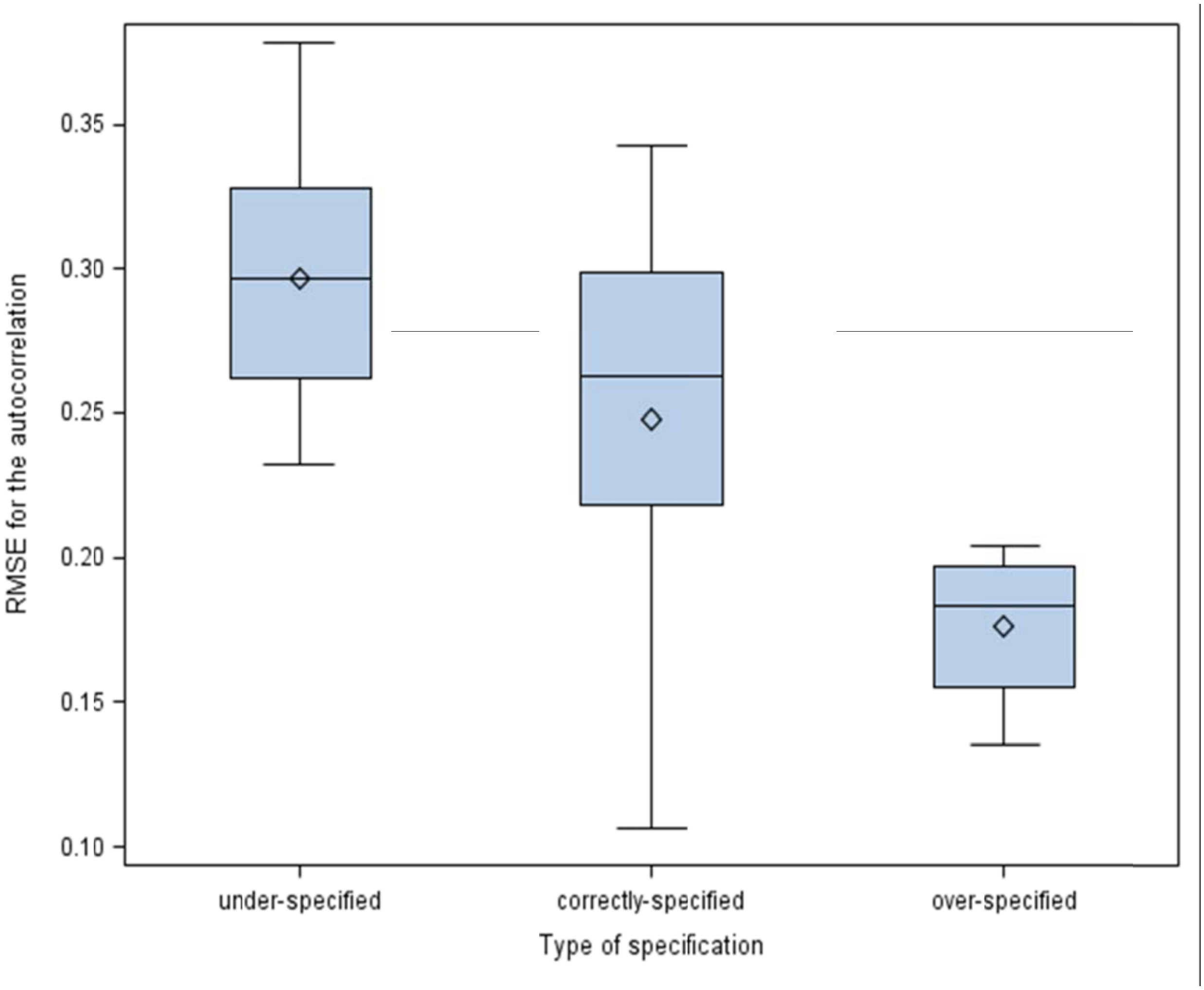


Figure 134. Box plots depicting the estimated RMSE of the autocorrelation across three types of specifications.

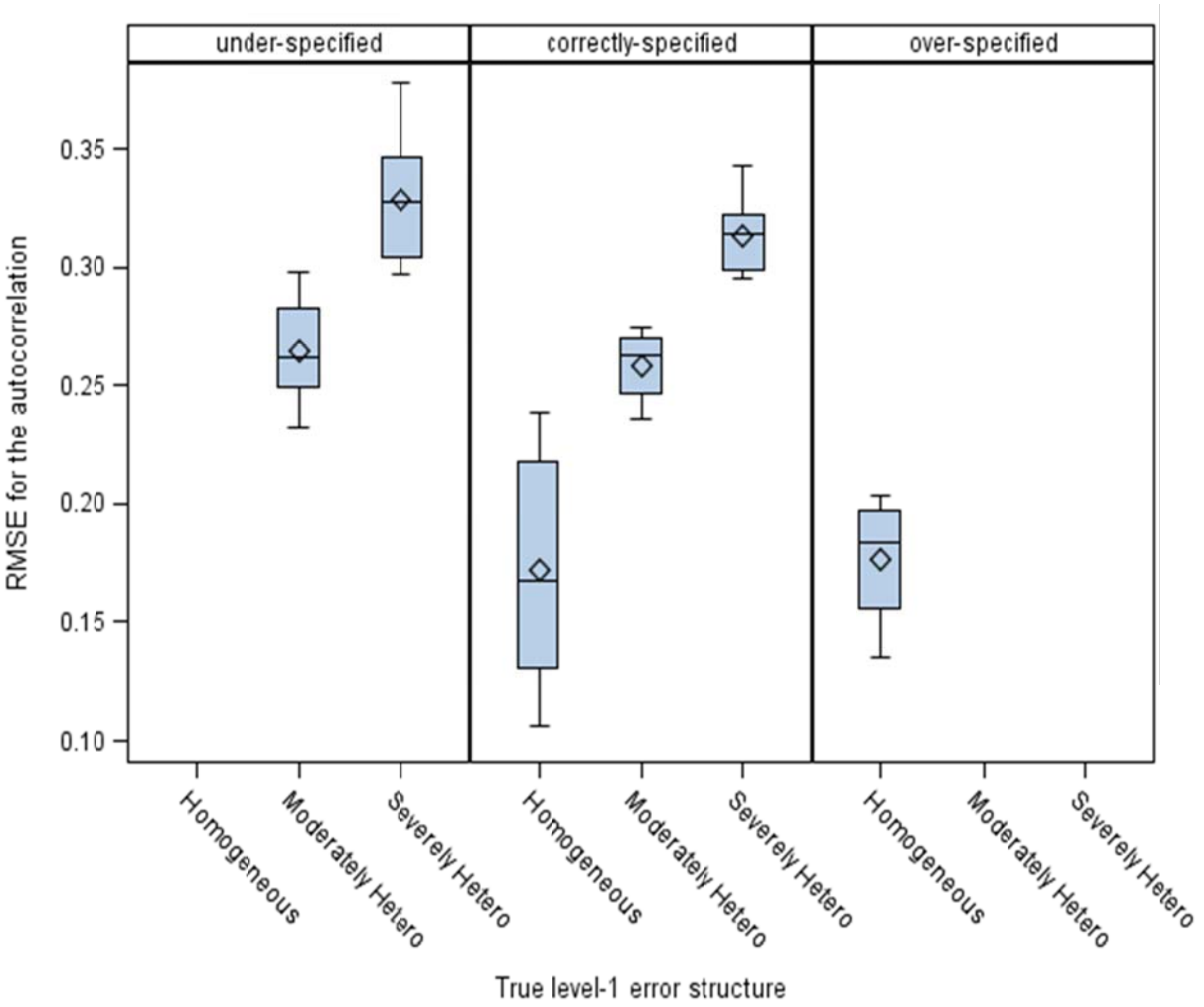


Figure 135. Box plots depicting the estimated RMSE of the autocorrelation across three types of specifications as a function of the true level-1 error structure.

CI coverage. The average CI coverage of the level-2 error standard deviation for phase (shift in level) and the level-2 error standard deviation for the interaction (shift in slope) by three different types of specifications are illustrated in Figures 136 through 138.

Figure 136 portrays that the average CI coverages of the level-2 error standard deviation for the shift in level were similar and tended to be over the nominal level across all three types of specifications, ranging from .967 to .974. The over-specified type had more conservative CI coverage than the other types of specifications.

Similarly, Figure 137 illustrated that the average CI coverage values of the level-2 error standard deviation for the shift in slope were slightly different and tended to exceed the nominal level across all three types of specifications, ranging from .970 to .982. The over-specified type had more conservative CI coverage than other types of specifications.

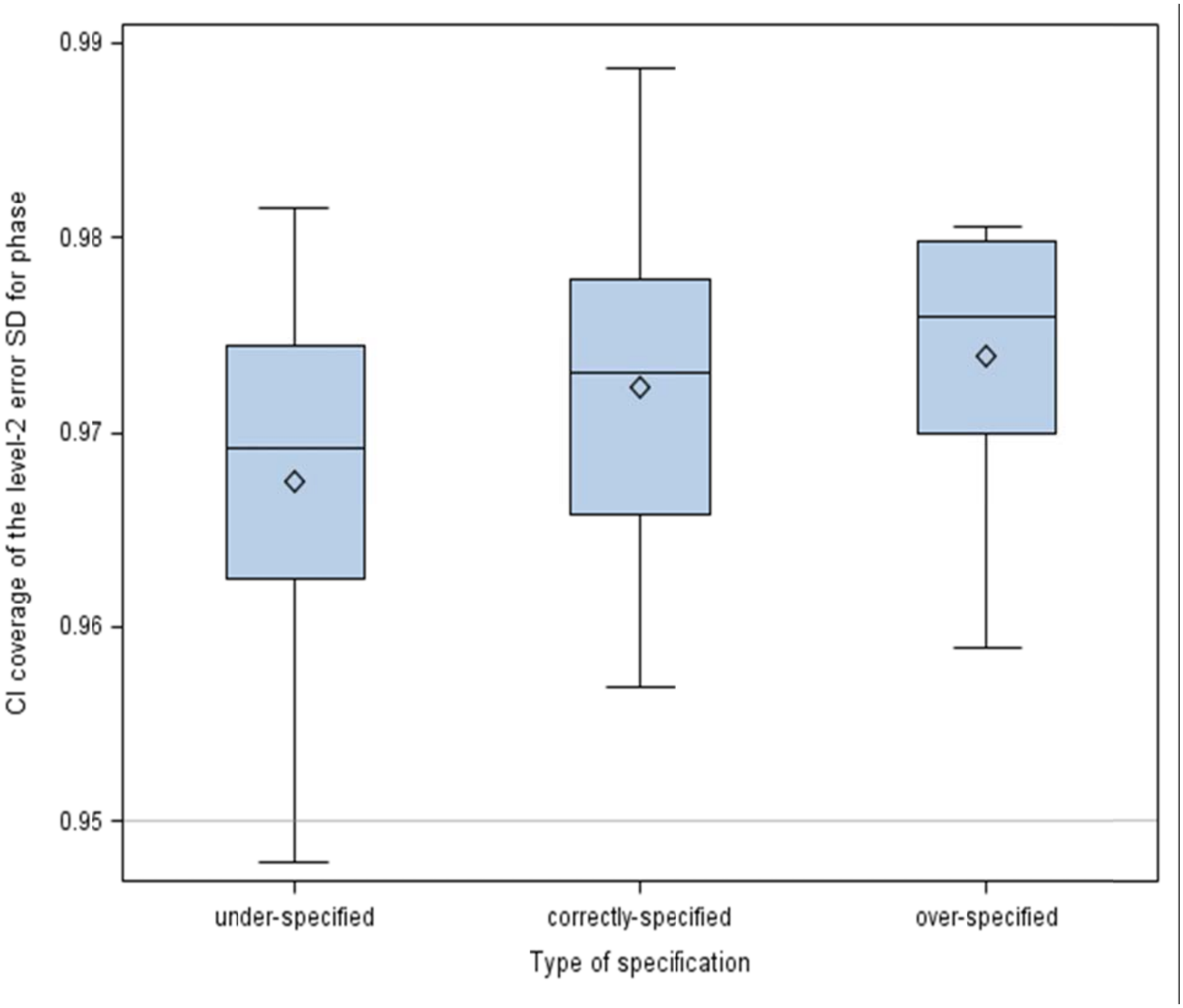


Figure 136. Box plots depicting the estimated CI coverage of the level-2 error standard deviation for phase across three types of specifications.

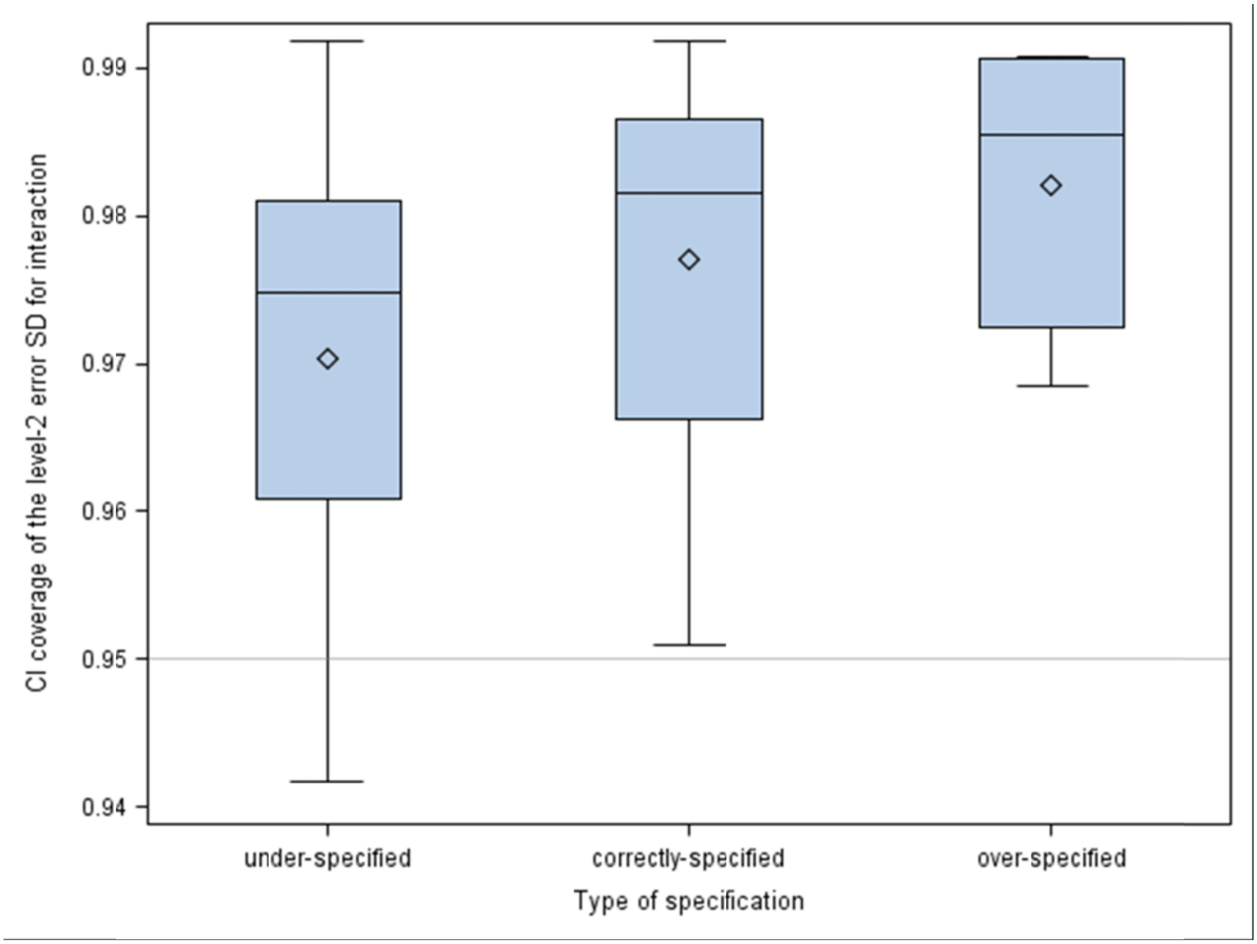


Figure 137. Box plots depicting the estimated CI coverage of the level-2 error standard deviation for the interaction across three types of specifications.

For further examination of this difference, the average CI coverage values of the level-2 error standard deviation for the shift in slope by the three different types of specifications as a function of the true level-1 error structure were presented in Figure 138. As illustrated in Figure 138, there were no or little differences among the true level-1 error structures within the different types of the specifications.

The average CI coverage values of the level-1 error standard deviation by three different types of specifications are illustrated in Figures 139 and 140.

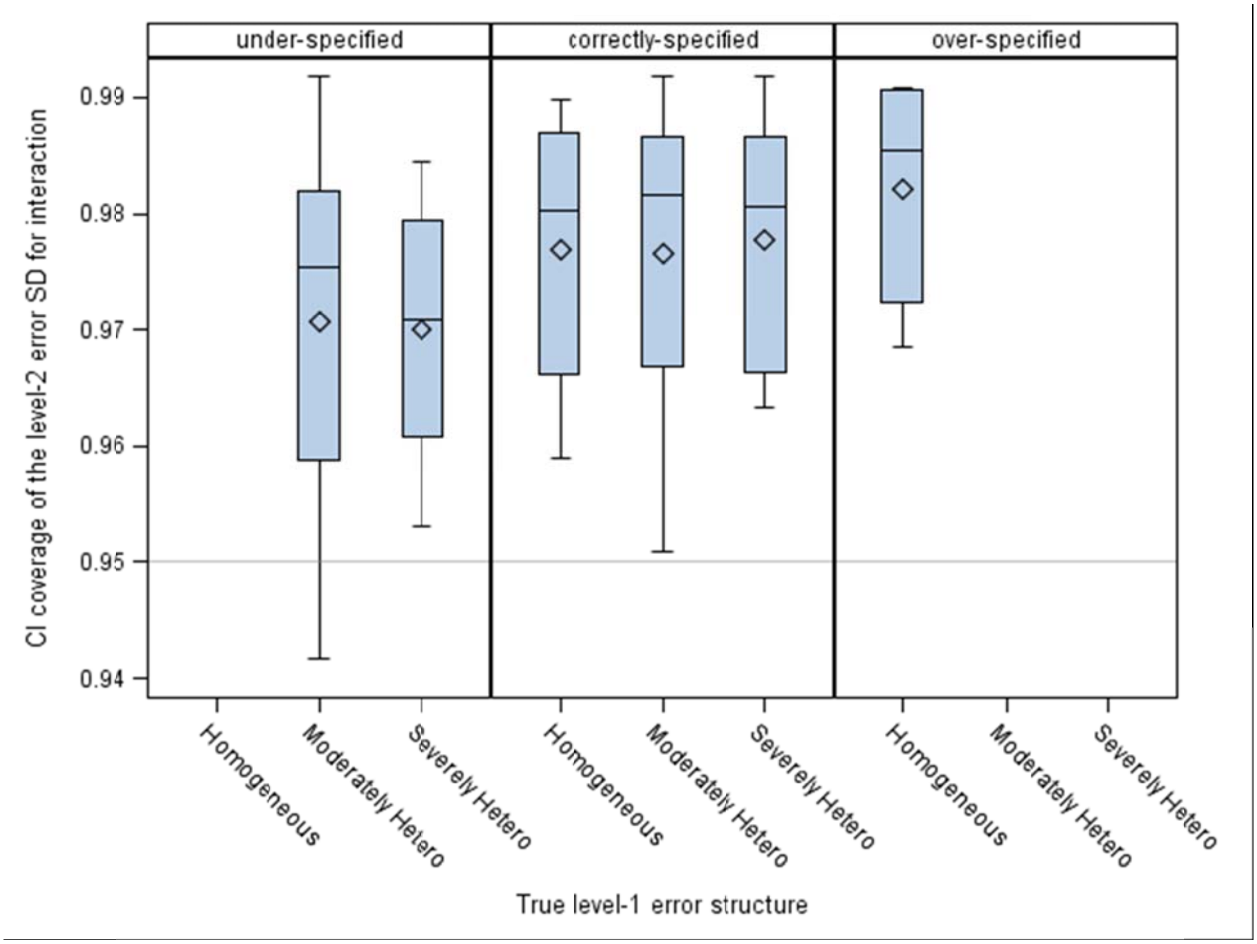


Figure 138. Box plots depicting the estimated CI coverage of the level-2 error standard deviation for the interaction across three types of specifications as a function of the true level-1 error structure.

Figure 139 portrays that the average CI coverage values of the level-1 error standard deviation were different, and tended to be under covered for the under-specified (*M* = .79) type, close to the nominal level for the correctly-specified (*M* = .96) type, and over covered for the over-specified (*M* = .99) type. The under-specified type had substantially larger variability than the other types of specifications. For further examination of this difference, the average CI coverage values of the level-1 error standard deviation by the three different types of specifications as a function of the true level-1 error structure were also shown in Figure 140. When the models were under-specified, the average CI coverage values of the level-1 error standard deviation tended to be under the nominal level for both the heterogeneous error structures, and as the degree of heterogeneity increased, the average CI coverage values substantially decreased (moderately hetero: *M* =.88; severely hetero: *M* =.70). When the models were correctly-specified, the average CI coverage values of the level-1 error standard deviation tended to be close to the nominal level, .95, regardless of the true level-1 error structure (homogeneous: *M* =.96; moderately hetero: *M* =.96; severely hetero: *M* =.95).

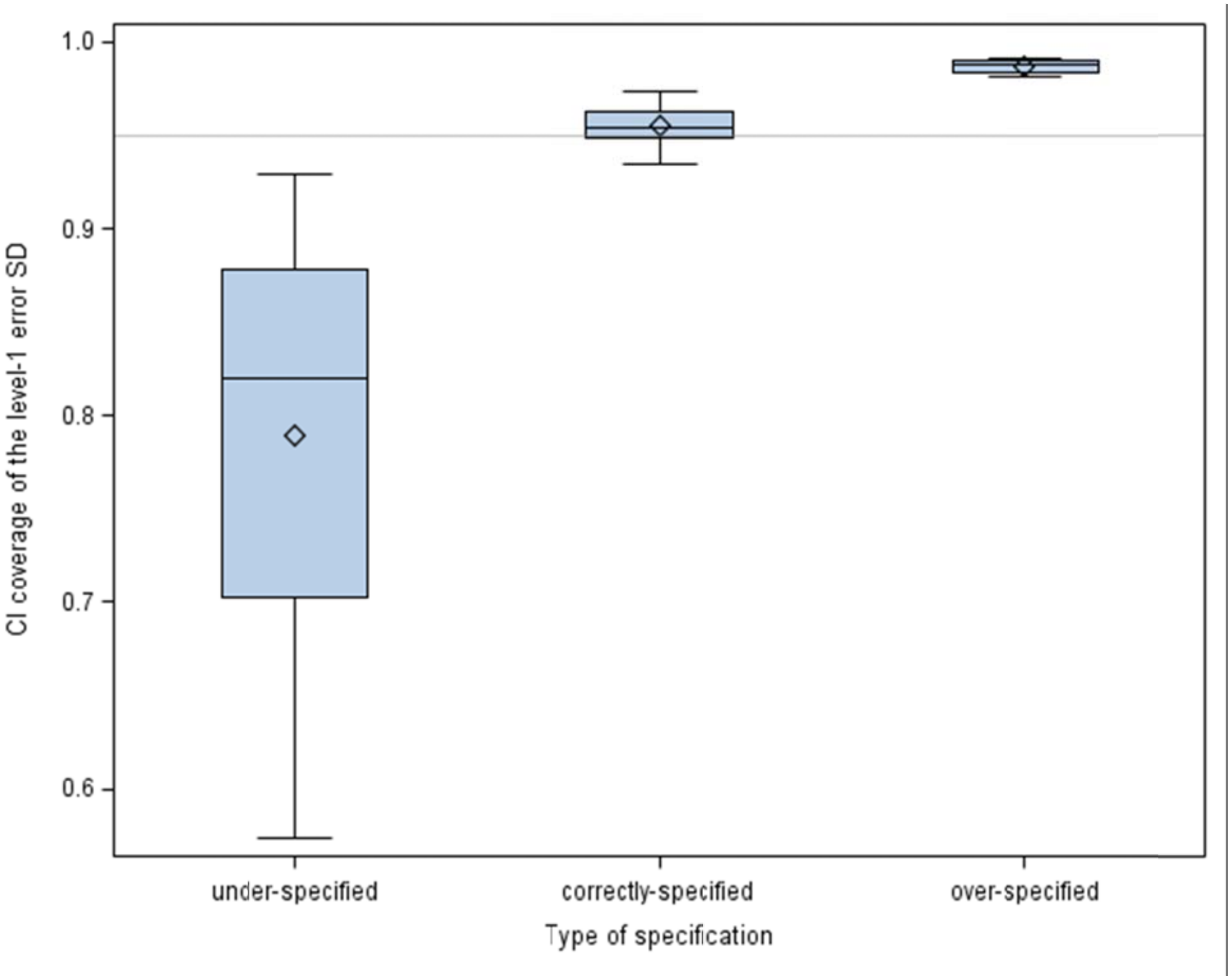


Figure 139. Box plots depicting the estimated CI coverage of the level-1 error standard deviation across three types of specifications.

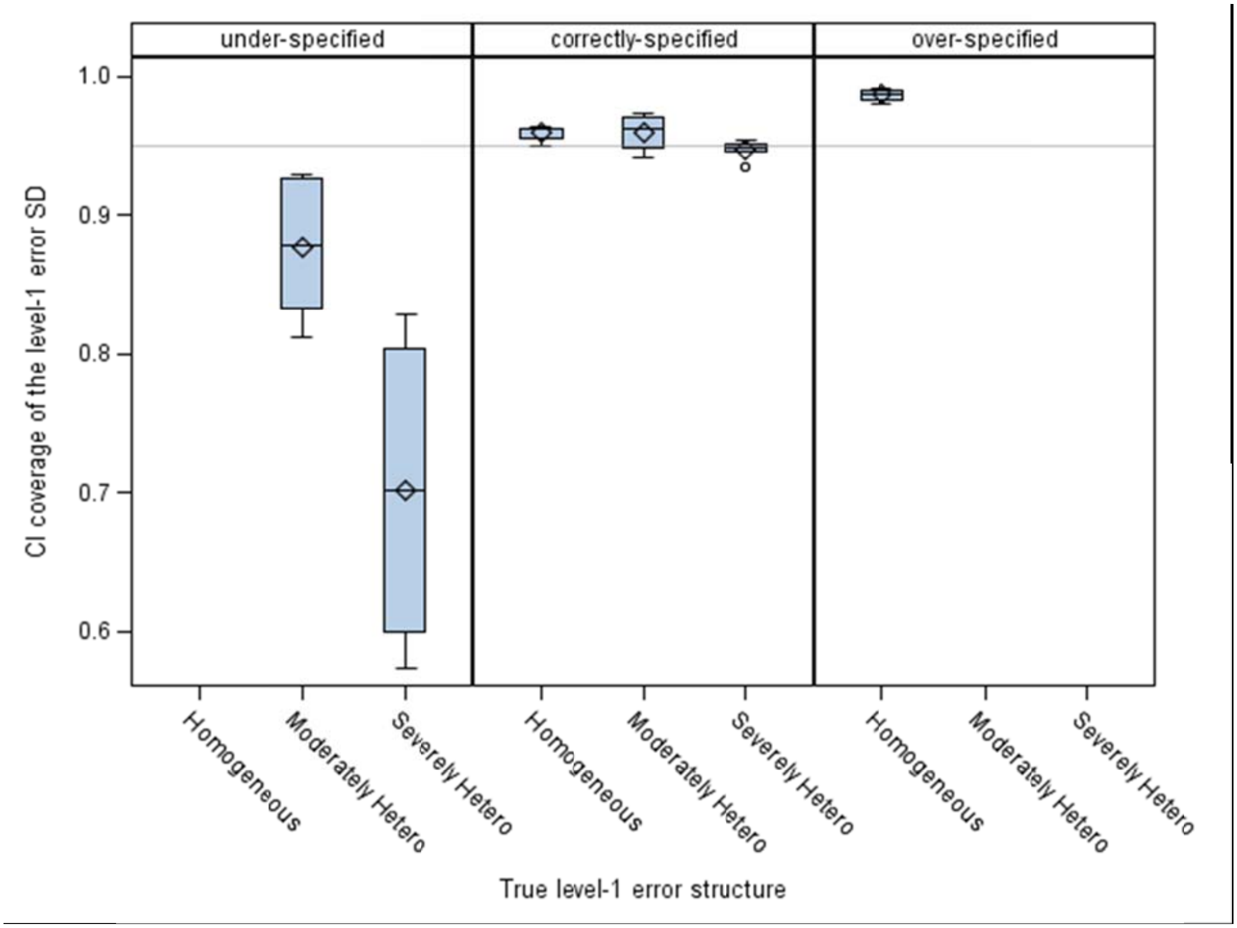


Figure 140. Box plots depicting the estimated CI coverage of the level-1 error standard deviation across three types of specifications as a function of the true level-1 error structure.

The average CI coverage values of the autocorrelation by three different types of specifications are illustrated in Figures 141 and 142.

Figure 141 portrays that the average CI coverage values of the autocorrelation were different, and tended to be under the nominal level for the under-specified (*M* = .74) type, close to the nominal level for the correctly-specified (*M* = .94**)** type, and over the nominal level for the over-specified (*M* = .99) type. The under-specified type had substantially larger variability than the other types of specifications.

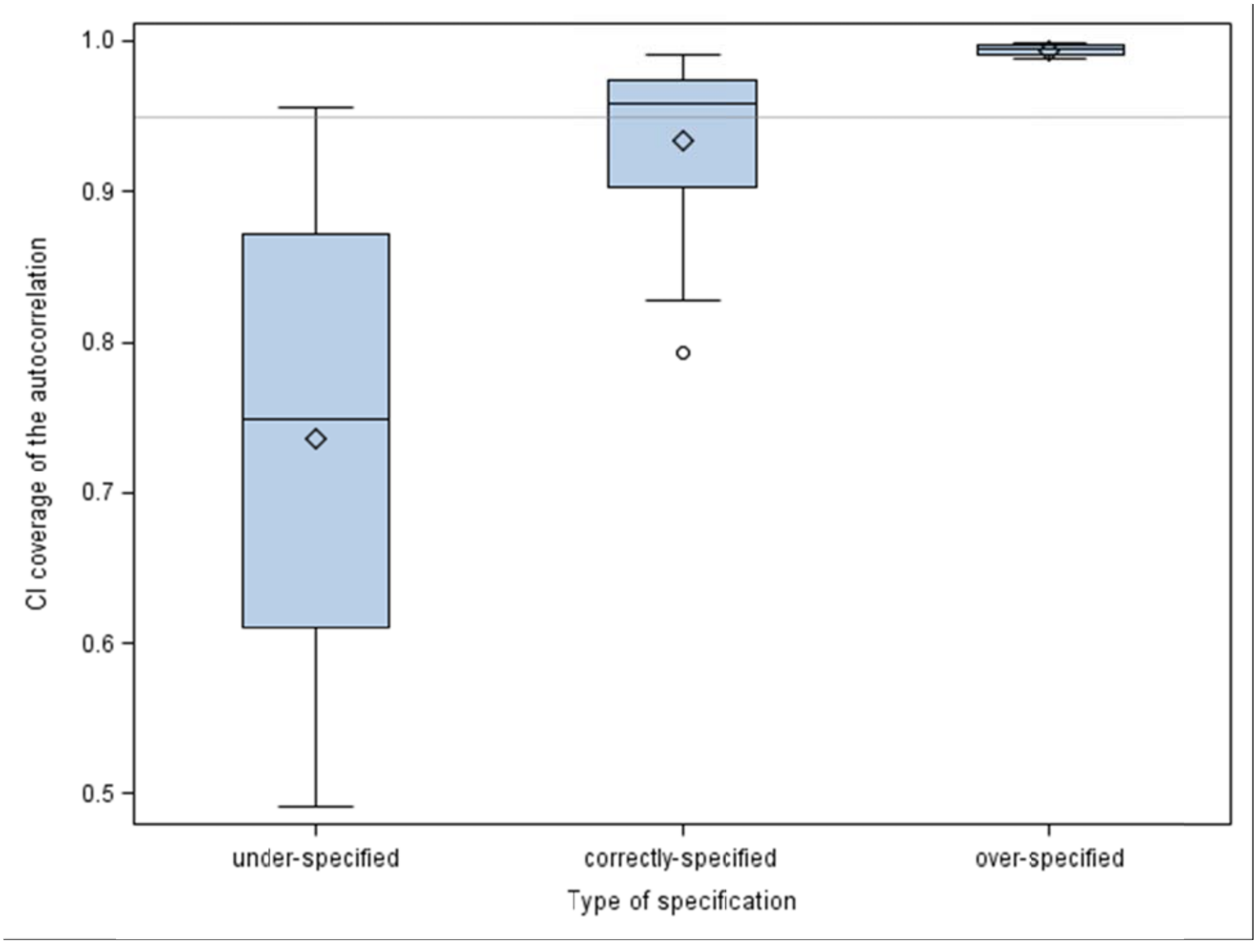


Figure 141. Box plots depicting the estimated CI coverage of the autocorrelation across three types of specifications.

For further examination of this difference, the average CI coverage values of the autocorrelation by the three different types of specifications as a function of the true level-1 error structure were also displayed in Figure 142. When models were under-specified, the average CI coverage values of the autocorrelation tended to be under the nominal level for both heterogeneous error structures, and as the degree of heterogeneity in the level-1 error structure increased, the average CI coverage value decreased (moderately hetero: *M* =.77; severely hetero: *M* =.70). When the models were correctly-specified, the average CI coverage values of the autocorrelation decreased as the degree of heterogeneity in the level-1 error structure increased (homogeneous: *M* =.97; moderately hetero: *M* =.94; severely hetero: *M* =.90). When the true level-1 error structure was moderately heterogeneous, the average CI coverage value of the autocorrelation was close to the nominal level, .95.

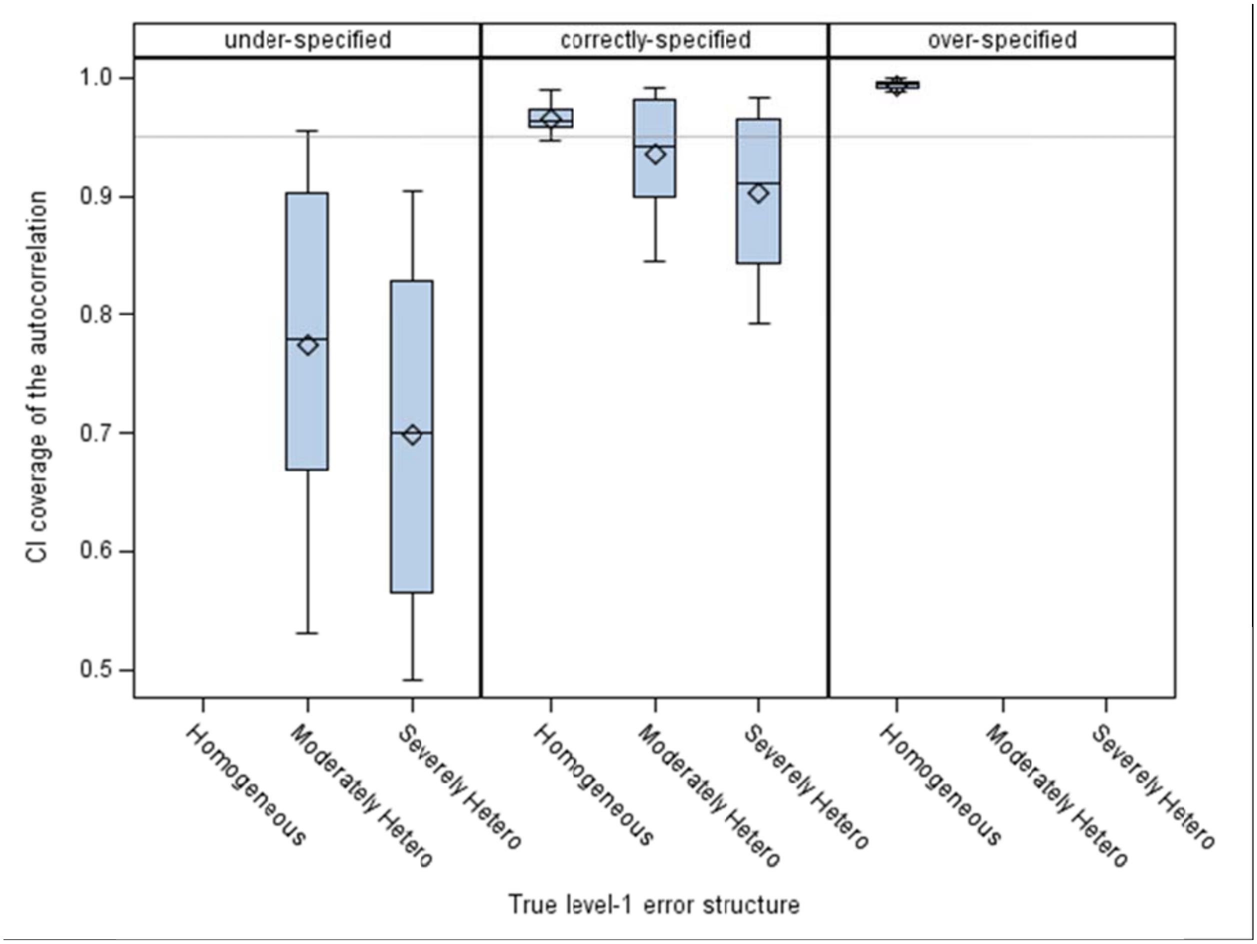


Figure 142. Box plots depicting the estimated CI coverage of the autocorrelation across three types of specifications as a function of the true level-1 error structure.

CI width. The average CI width values of the level-2 error standard deviation for phase (shift in level) and the level-2 error standard deviation for the interaction (shift in slope) by three different types of specifications are illustrated in Figures 143 through 145.

Figure 143 portrays that the average CI width values of the level-2 error standard deviation for the shift in level were similar across the three types of specifications, ranging from 6.43 to 6.48. In contrast, Figure 144 illustrates that the average CI width values of the level-2 error standard deviation for the shift in slope were slightly different across the three types of specifications, ranging from 1.61 to 2.20. The Over-specified type had the widest average CI width among the three types of specifications.

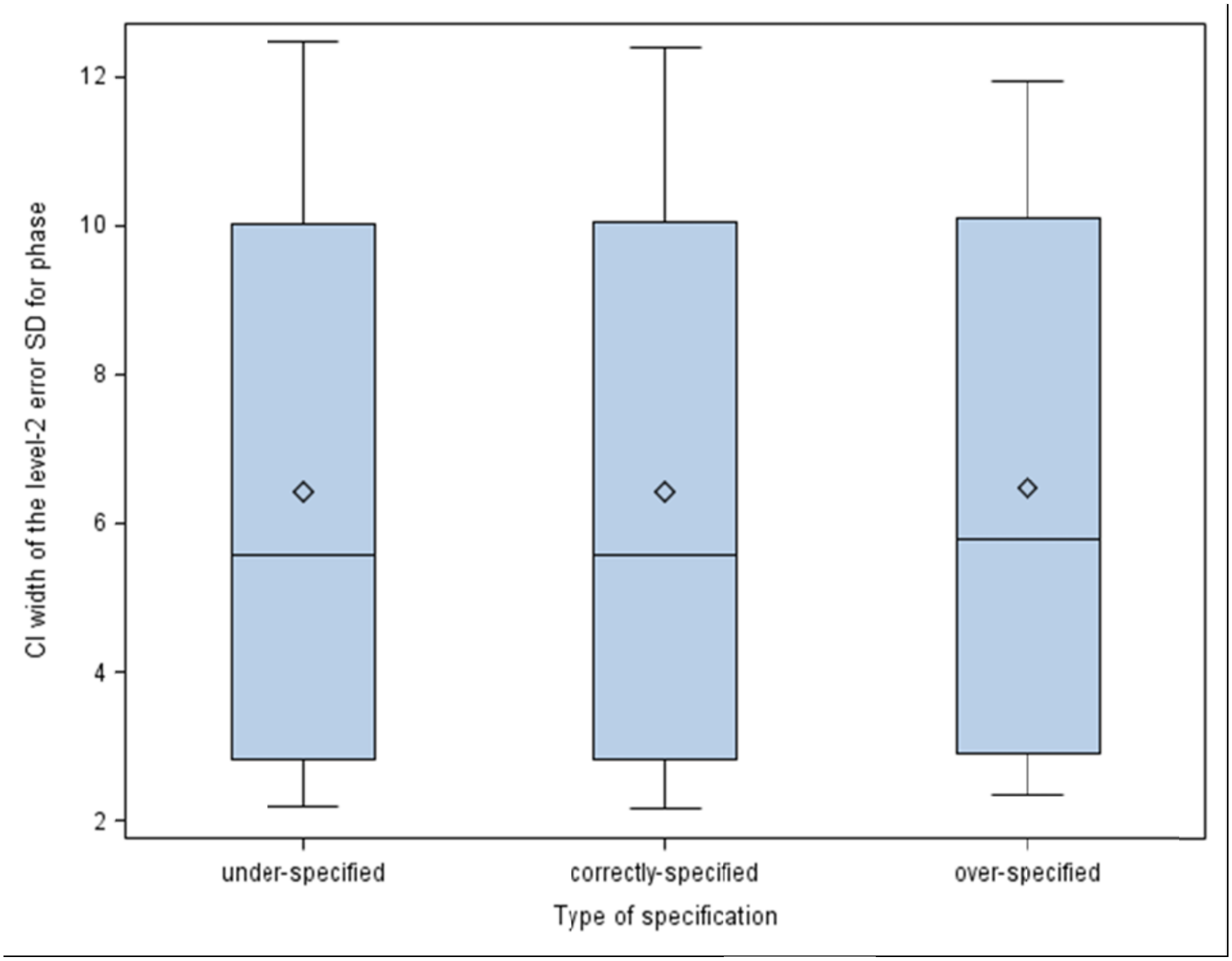


Figure 143. Box plots depicting the estimated CI width of the level-2 error standard deviation for phase across three types of specifications.

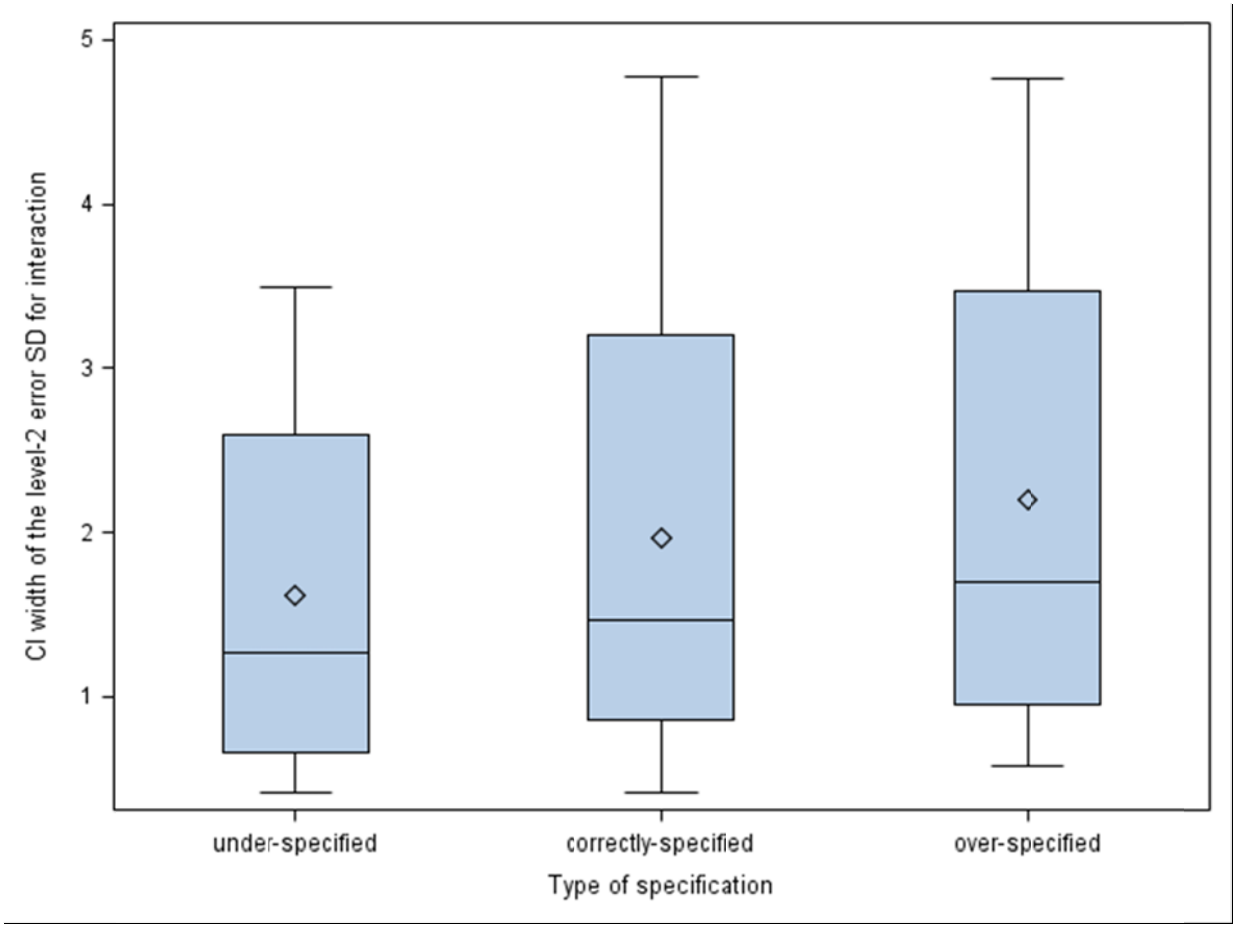


Figure 144. Box plots depicting the estimated CI width of the level-2 error standard deviation for interaction across the three types of specifications.

For further examination of this difference, the average CI width values of the level-2 error standard deviation for the shift in slope by the three different types of specifications as a function of the true level-1 error structure were also graphed in Figure 145. When models were under-specified, there were little differences across the true level-1 error structures (moderately hetero: *M* =1.61; severely hetero: *M* =1.62). When the models were correctly-specified, the average CI width value was smaller when the true level-1 error structure was homogeneous than one of the heterogeneous error structures (homogeneous: *M* =1.63; moderately hetero: *M* =2.14; severely hetero: *M* =2.15).

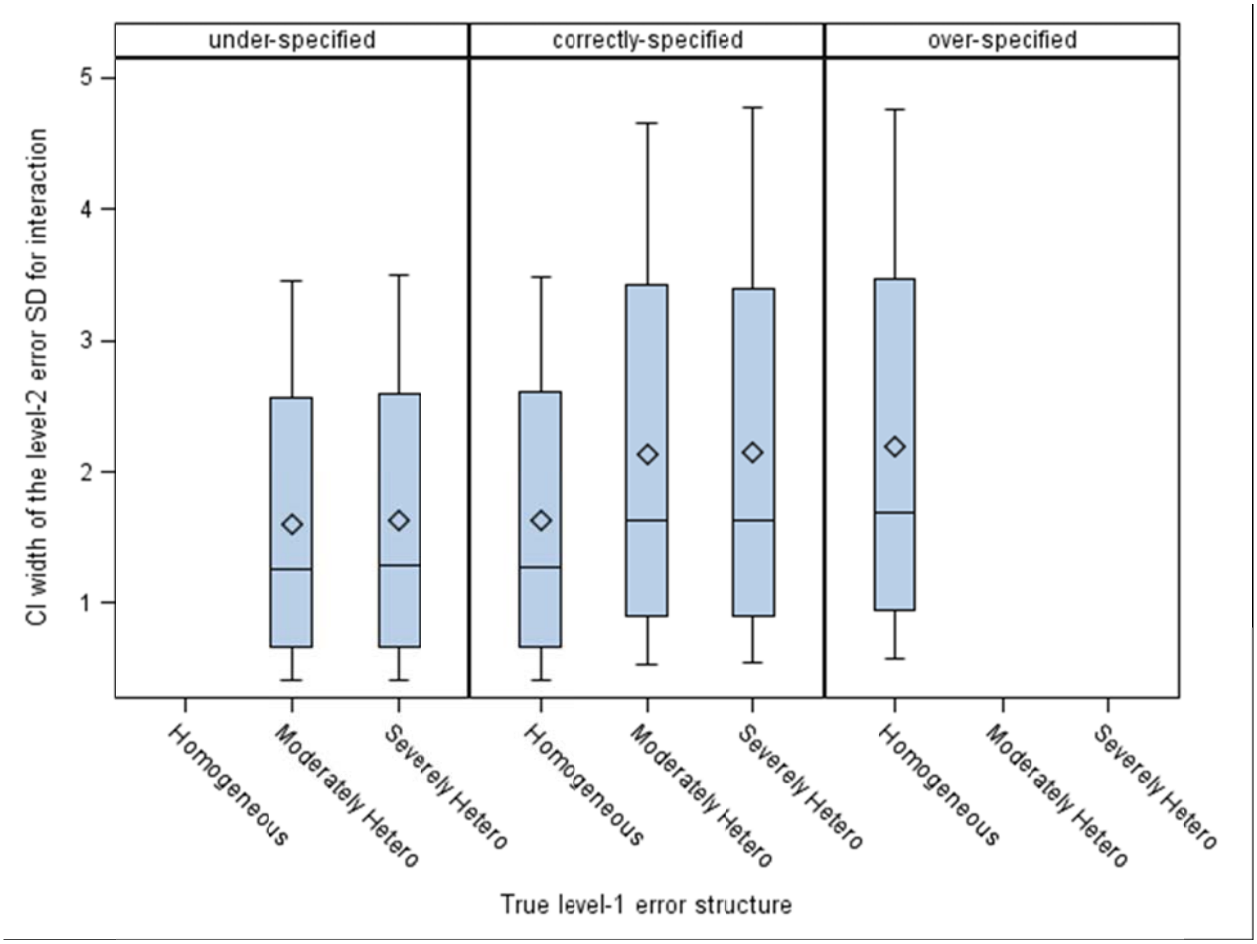


Figure 145. Box plots depicting the estimated CI coverage of the autocorrelation across three types of specifications as a function of the true level-1 error structure.

The average CI width values of the level-1 error standard deviation by three different types of specifications are illustrated in Figures 146 and 147.

Figure 146 portrays that the average CI width values of the level-1 error standard deviation were different, ranging from .47 to .74. The under-specified type had a smaller average CI width value than the other types of specifications.

For further examination of this difference, the average CI width values of the level-1 error standard deviation by the three different types of specifications as a function of the true level-1 error structure were displayed in Figure 147. When models were under-specified, the average CI width values of the level-1 error standard deviation were similar across the true level-1 error structures (moderately hetero: *M* =.46; severely hetero: *M* =.48). When the models were correctly-specified, and the true level-1 error structure was homogeneous, it had the smallest average CI width value of the level-1 error standard deviation (homogeneous: *M* =.46; moderately hetero: *M* =.80; severely hetero: *M* =.89).

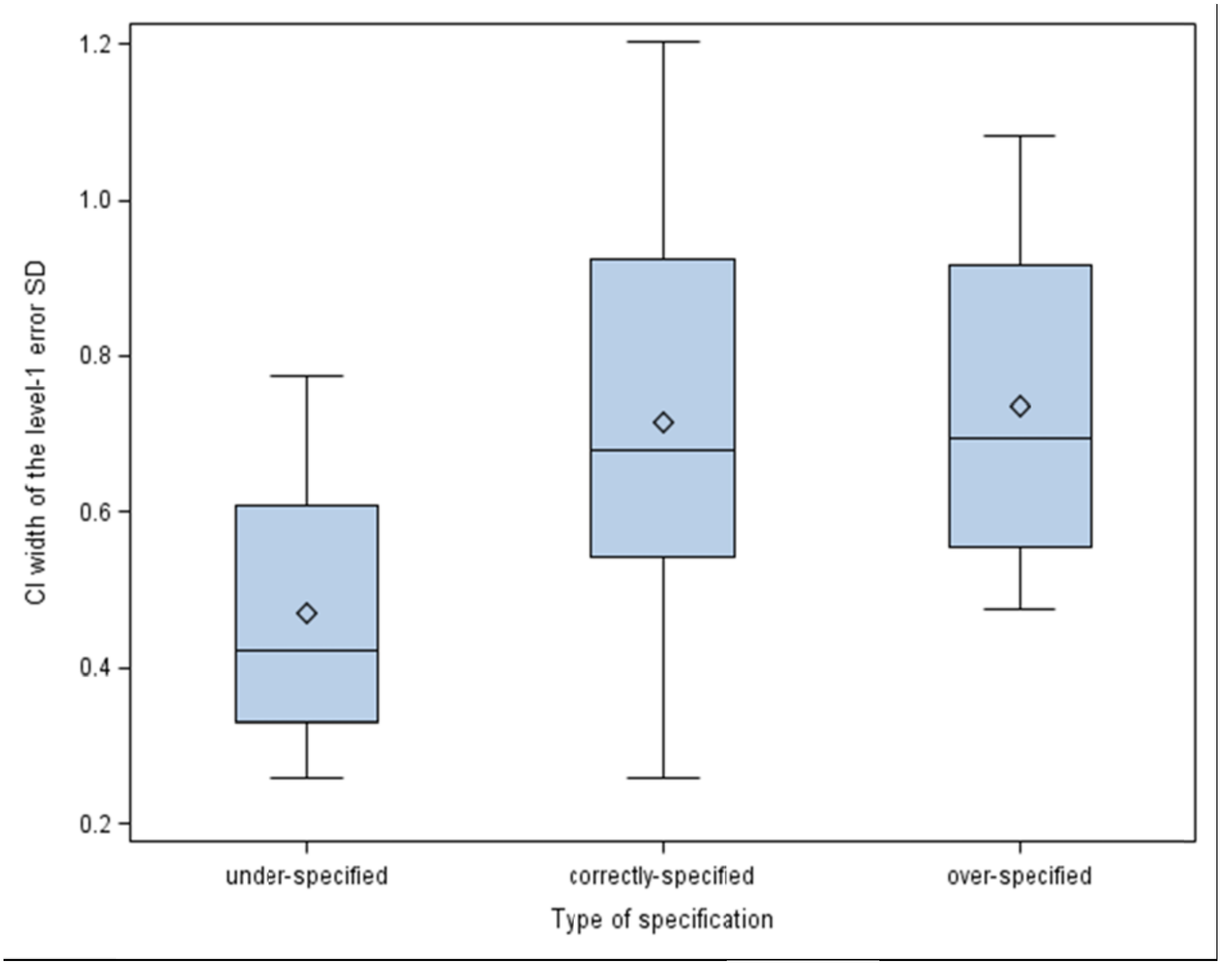


Figure 146. Box plots depicting the estimated CI width of the level-1 error standard deviation across three types of specifications.

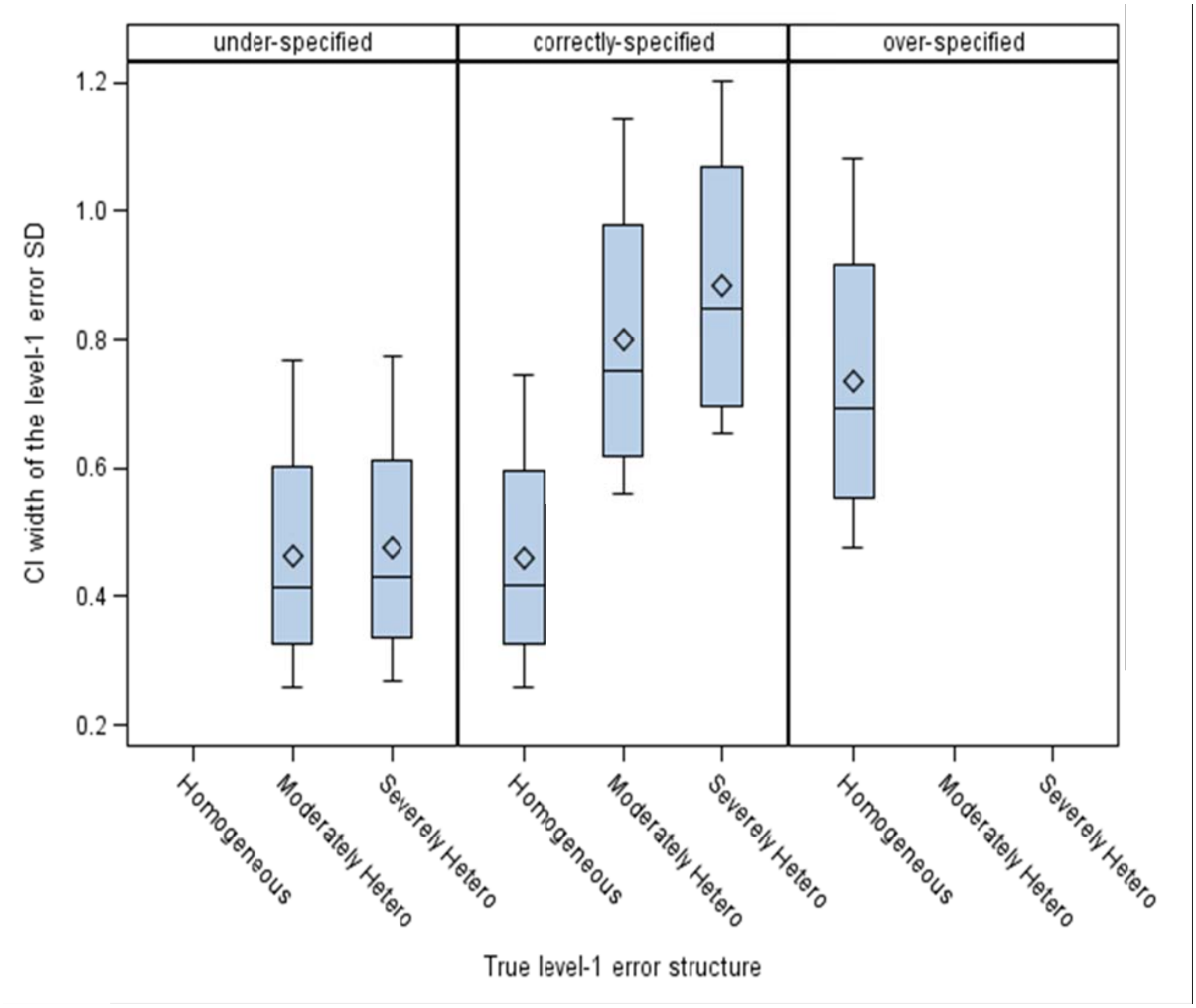


Figure 147. Box plots depicting the estimated CI width of the level-1 error standard deviation across three types of specifications as a function of the true level-1 error structure.

The average CI width values of the autocorrelation by three different types of specifications are illustrated in Figures 148 and 149.

Figure 148 portrays that the average CI width values of the autocorrelation were different, ranging from .73 to 1.09. The under-specified type had a smaller average CI width value than the other types of specifications. For further examination of this difference, the average CI width values of the autocorrelation by the three different types of specifications as a function of the true level-1 error structure were displayed in Figure 149. When the models were under-specified, the average CI width values of the autocorrelation were similar across the true level-1 error structures (moderately hetero: *M* =.73; severely hetero: *M* =.73). When the models were correctly-specified, and the true level-1 error structure was homogeneous, it had the smallest average CI width value of the autocorrelation (homogeneous: *M* =.77; moderately hetero: *M* =1.10; severely hetero: *M* =1.12).

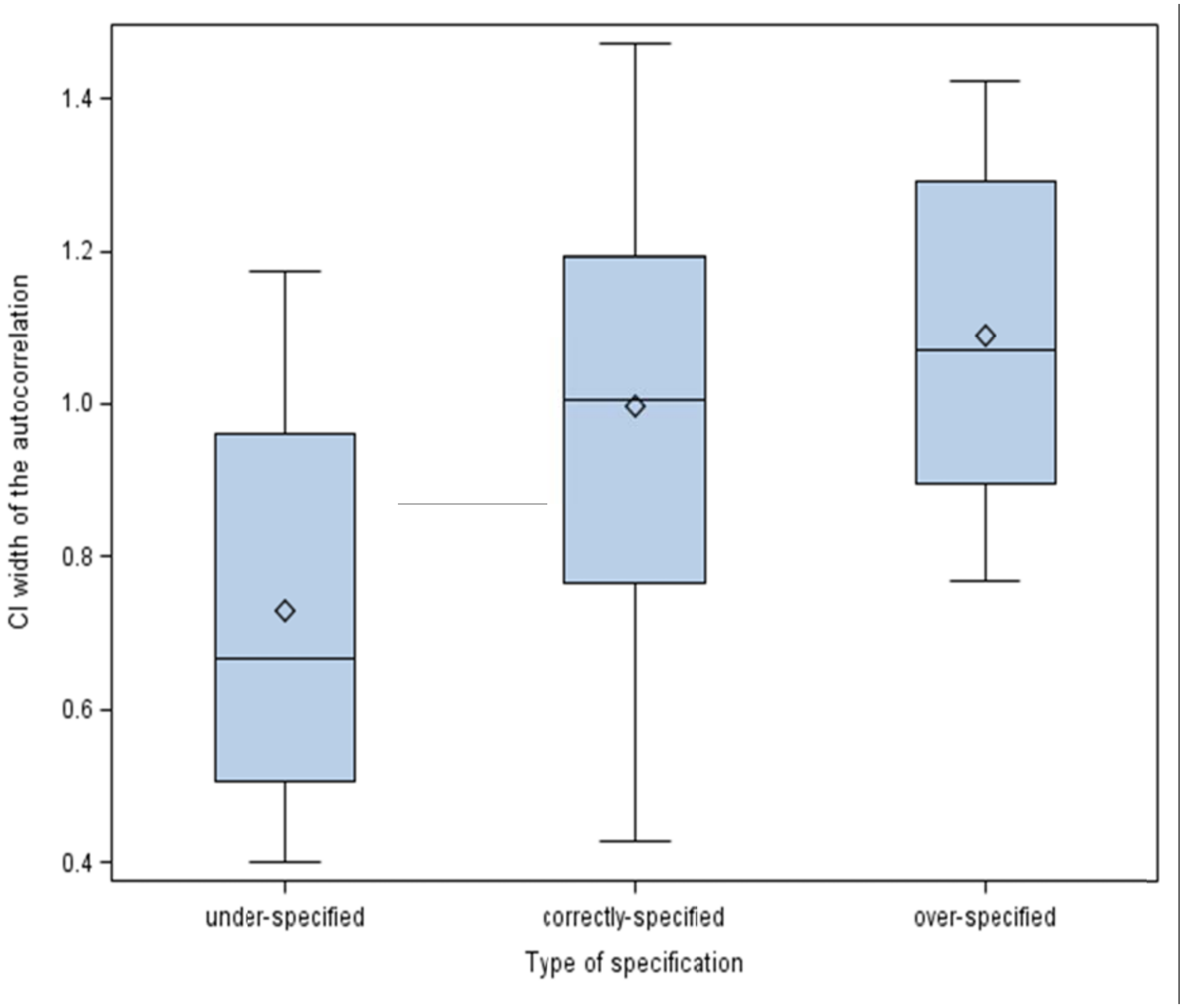


Figure 148. Box plots depicting the estimated CI width of the autocorrelation across three types of specifications.

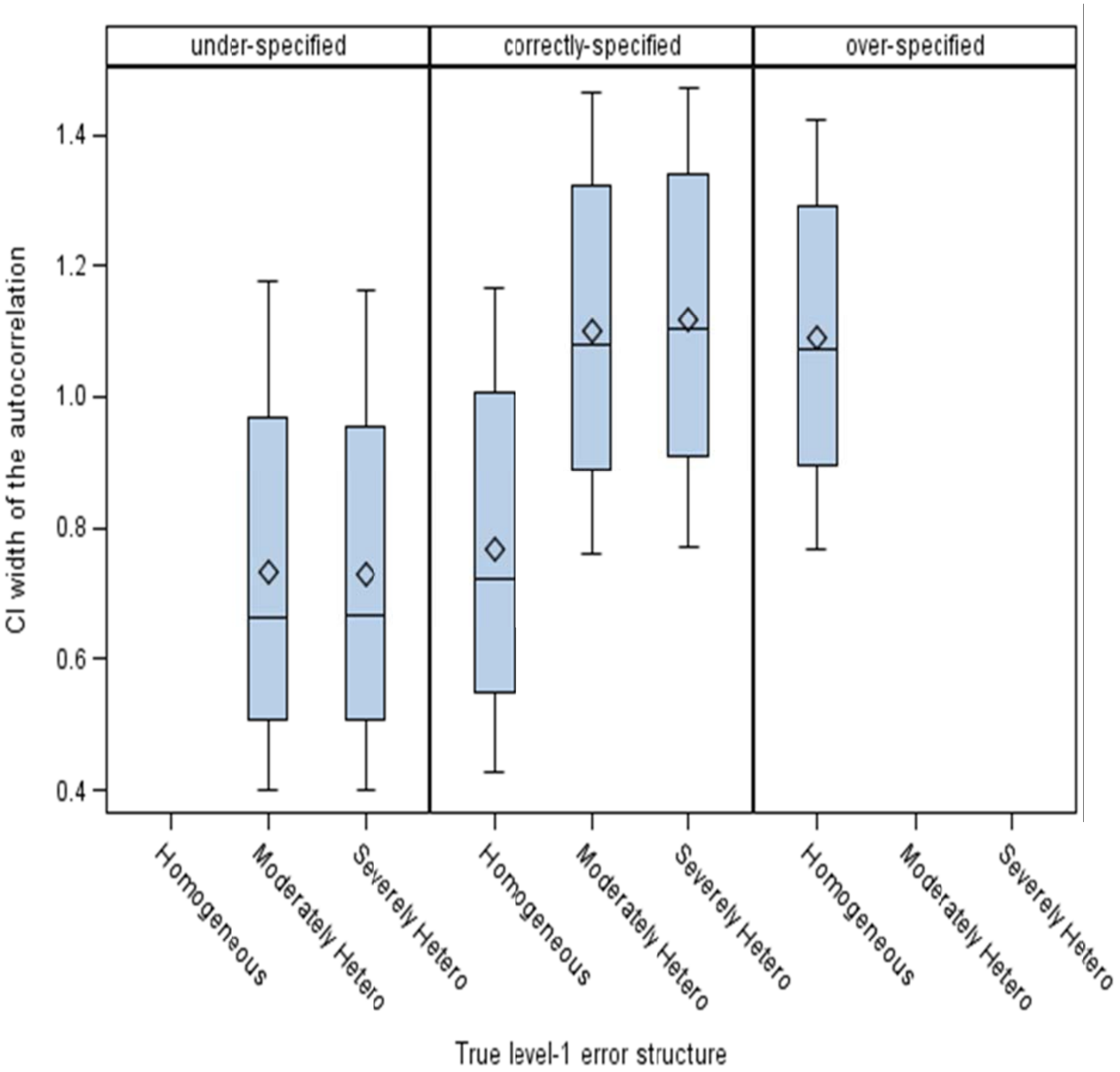


Figure 149. Box plots depicting the estimated CI width of the autocorrelation across three types of specifications as a function of the true level-1 error structure.

Summary of the Study

The main study with 48 conditions found that different methods of modeling level-1 error structure had little to no impact on the estimates of the fixed treatment effects, but substantial impact on the estimates of the variance components, especially the level-1 error standard deviation and the autocorrelation parameters. Similarly, whether the level-1 error structure was under-specified, over-specified, or correctly-specified had little to no impact on the estimates of the fixed treatment effects, but a substantial impact on the estimates of the variance components, especially the level-1 error standard deviation and the autocorrelation. In addition, it was found that the different type of true level-1 error structure had substantial impact on the estimates of the level-1 error standard deviation and the autocorrelation. The summary tables of these findings are provided in Tables 2 through 5.

The fixed treatment effects were not biased for both Model 1 and 2. The average RMSE values for the fixed treatment effects were similar across the models. The interval coverage for the fixed treatment effects tended to be over the nominal level for both models. The interval width values were similar across the two models. In addition, under- or over-specification of the level-1 error structure had little to no impact on the estimates of the fixed treatment effects.

For the variance components, all level-2 error standard deviation estimates were positively biased for both Model 1 and 2. The average RMSE values for the level-2 error standard deviation estimates were similar across the two models. The interval coverage for the level-2 error standard deviations tended to be over the nominal level for both models. The interval width values were similar across the two models. In addition, different types of specifications in the level-1 error structure had little to no impact on the estimates of the level-2 error standard deviations.

Unlike the level-2 error standard deviations that had similar results across Model 1 and 2, the level-1 error standard deviation and autocorrelation show some differences in terms of the results across Model 1 and 2. The level-1 error standard deviation was similar and positively biased for both models, but the average RMSE values were different across the two models. The average RMSE value was smaller and had less variability when estimated by Model 2 than

Model 1. In addition, the interval coverage had a substantial difference across the two models. It was under the nominal level when estimated by Model 1, but close to the nominal level when estimated by Model 2. The interval width was smaller when estimated by Model 1 than Model 2. Similarly, the autocorrelation was similar but negatively biased for both models, and the average RMSE value was similar across the two models. The interval coverage was substantially different across the two models. It was under the nominal level when estimated by Model 1 but close to the nominal level when estimated by Model 2. The interval width was smaller when estimated by Model 1 than Model 2.

In addition, different types of specifications in the level-1 error structure had a substantial impact on the estimates of the variance components, especially the level-1 error standard deviation and the autocorrelation. For the average bias and RMSE values of the level-1 error standard deviation and the autocorrelation, over-specified models had the smallest bias and RMSE values, and for the CI coverage of the level-1 error standard deviation and the autocorrelation, the correctly-specified models led to coverage that was the closest to the nominal level.

Moreover, different types of the true level-1 error structures had a substantial impact on the estimates of the level-1 error standard deviation and the autocorrelation. As the degree of heterogeneity in the level-1 error structures increased, estimates of the level-1 error standard deviation and the autocorrelation tended to be more accurate when estimated by Model 2 than Model 1.

Table 2

Summary of the results for the fixed treatment effects

Parameter estimate	Bias	RMSE	CI coverage	CI width
Shift in level	• Close to 0 for both Model 1 (M=.002) and Model 2 (M<.001)	• Similar across the two models (M=.68 for both models)	• Over the nominal level for both models (M=.98 for both models)	• Similar across the two models (Model 1:M=4.96; Model 2: M=4.95)
	• The moderately heterogeneous error structure had the smallest bias but largest variability	• No or little difference across the true level-1 error structures (η2 = .001)	• No or little difference across the true level-1 error structures (η2 = .008)	• No or little difference across the true level-1 error structures (η2 = .0002)
	• One medium effect (η2 = .10) for the 4-way interaction among number of cases, series length per case, variation in the level-2 errors, and true level-1 error structure	• Three of the design factors had a medium or large effect, including the number of cases (η2 = .48), variation in the level-2 errors (η2 = .38), and the series length per case (η2 = .11)	• One large effect (η2 = .88) for the number of case	• Two of the design factors had a medium or large effect including the number of cases (η2 = .84) and the variation in the level-2 errors (η2 = .12)
Shift in slope	• Close to 0 for both Model 1 (M=.004) and Model 2 (M=.003)	• Similar across the two models (Model 1: M=.23; Model 2: M=.22)	• Over the nominal level for the both Model 1 and Model 2. M=.99 for both models	• Similar across the two models (M=1.68 for both models)
	• The severely heterogeneous error structure had the smallest bias, and the homogeneous error structure had the largest bias	• No or little difference across the true level-1 error structures (η2 = .00143)	• No or little difference across the true level-1 error structures (η2 = .0005)	• No or little difference across the true level-1 error structures (η2 = .0005)

Table 2 (continued)

Summary of the results for the fixed treatment effects

	• Three interaction effects had a medium effect, including the 3-way interaction among the number of cases, the series length per case, and the true level-1 error structure (η2 = .10), the 3-way interaction among the number of cases, the series length per case, and the variation in the level-2 errors (η2 = .09), and the 3-way interaction among the series length per case, the true level-1 error structure, and the variation in the level-2 errors (η2 = .08)	• Three of the design factors had a large effect, including the series length per case (η2 = .47), the number of cases (η2 = .28), and the variation in the level-2 errors (η2 = .22)	• Two of the design factors had a medium or large effect, including the number of cases (η2 = .83) and the series length per case (η2 = .08)	• Three of the design factors had a medium or large effect including the number of cases (η2 = .65), the series length per case (η2 = .19), and the variation in the level-2 errors (η2 = .10)

Table 3

Summary of the results for the variance components

Parameter estimate	Bias	RMSE	CI coverage	CI width
Level-2 error standard deviation for shift in level	• Similar across the two models and both positively biased (Model 1: M=.86; Model 2:M=.85)	• Similar across the two models (Model 1: M=1.23; Model 2:M=1.21)	• Over the nominal level across the two models (Model 1: M=0.97; Model 2:M=0.97)	• Similar across the two models (Model 1: M=6.43; Model 2:M=6.44)
	• No or little difference across the true level-1 error structures (η2 = .0004)	• No or little difference across the true level-1 error structures (η2 = .0003)	• Little difference across the true level-1 error structures (η2 = .02)	• No or little difference across the true level-1 error structures (η2 = .0002)
	• One of the design factors, the number of cases, had a large effect (η2 = .96)	• Two of the design factors had a medium or large effect, including the number of cases (η2 = .89), and the variation in the level-2 errors (η2 = .08)	• Two main effects and one interaction effect had a medium or large effect, including the variation in the level-2 errors (η2 = .29), the type of model (η2 = .07), the 3-way interaction among the series length per case, the number of cases and the true level-1 error structure (η2 = .11)	• Two of the design factors had a medium or large effect including the number of cases (η2 = .88) and the variation in the level-2 errors (η2 = .08)

Table 3 (continued)
Summary of the results for the variance components

Level-2 error standard deviation for shift in slope	• Similar across the two models and both positively biased (Model 1: M=.31; Model 2:M=.30) • No or little difference across the true level-1 error structures (η2 = .0004) • One medium effect (η2 = .07) for the 2-way interaction between the number of cases and the series length per case	• Similar across the two models (Model 1: M=.42; Model 2:M=.41) • No or little difference across the true level-1 error structures (η2 = .0002) • Two of the design factors had a medium or large effect, including the number of cases (η2 = .73) and the series length per case (η2 = .13)	• Over nominal level across the two models (Model 1: M=0.97; Model 2: M=0.98) • Some difference across the true level-1 error structures (η2 = .05) • Two of the design factors and one interaction effect that had a medium or large effect, including the series length per case (η2 = .65), the type of model (η2 = .06), and the 2-way interaction between the variation in the level-2 errors and the true level-1 error structure (η2 = .07)	• Model 1 had a smaller CI width than Model 2. (Model 1: M=1.62; Model 2: M=2.16). • No or little difference across the true level-1 error structures (η2 = .0002) • Three main effects that had a medium or large effect, including the number of cases (η2 = .70), variation in the level-2 errors (η2 = .10), and the series length per case (η2 = .07)

Table 3 (continued)
Summary of the results for the variance components

Level-1 error standard deviation	• Similar across the two models and both positively biased (Model 1: M=.05; Model 2:M=.04)	• Model 2 had a smaller RMSE value than Model 1 (Model 1: M=.22; Model 2:M=.18)	• Substantial difference across the two models (η2 = .3). Under nominal level and more variability for Model 1. Over nominal level for Model 2 (Model 1: M=.85; Model 2: M=.97)	• Model 1 had a smaller CI width than Model 2 (Model 1: M=0.47; Model 2:M=0.81)
	• Substantial differences across the true level-1 error structures (η2 = .22). For the homogeneous or moderately heterogeneous error structure, more biased when estimated by Model 2 than Model 1. For the severely heterogeneous error structure, more biased when estimated by Model 1 than Model 2.	• Substantial differences across the true level-1 error structures (η2 = .62). For the homogeneous error structure, larger when estimated by Model 2 than Model 1. For the heterogeneous error structures, smaller when estimated by Model 2 than Model 1.	• Substantial differences across the true level-1 error structures (η2 = .3). For the homogeneous error structure, over the nominal level for the both models. For the heterogeneous structures, under the nominal level for Model 1 and either at or slightly over the nominal level for Model 2.	• Little difference across the true level-1 error structures (η2 = .02). Generally, the CI width for Model 1 was smaller than the CI width for Model 2.

Table 3 (continued)

Summary of the results for the variance components

	• Three main effects and one 2-way interaction had a medium or large effect, including the series length per case (η2 = .25), the variation in the level-2 errors (η2 = .19), the number of cases (η2 = .11), and the 2-way interaction between the type of model and the true level-1 error structure (η2 = .10)	• One main effect and one interaction effect had a medium or large effect, including the series length per case (η2 = .11) and the 2-way interaction between the type of model and the true level-1 error structure (η2 = .16)	• One main effect and one interaction effect had a medium or large effect, including the series length per case (η2 = .06), the 2-way interaction between the type of model and the true level-1 error structure (η2 = .19)	• Three of the design factors had a medium or large effect, including the type of model (η2 = .44), the number of cases (η2 = .38), and the series length per case (η2 = .11)
Autocorrelation	• Similar across the two models and both negatively biased (Model 1: M=-.10; Model 2:M=-.09)	• Similar across the two models (Model 1: M=.26; Model 2:M=-.25)	• Substantial difference across the two models (η2 = .2). Under nominal level and more variability for Model 1. Close to the nominal level for Model 2 (Model 1: M=.81; Model 2: M=.94).	• Model 1 had a smaller CI width than Model 2 (Model 1: M=0.74; Model 2:M=1.10)

Table 3 (continued)

Summary of the results for the variance components

	• Substantial differences across the true level-1 error structures (η2 = .22). More biased for the heterogeneous error structures than the homogeneous error structure, regardless of the type of model.	• Substantial differences across the true level-1 error structures (η2 = .62). For the homogeneous error structure, larger when estimated by Model 2 than Model 1. For the heterogeneous error structures, smaller when estimated by Model 2 than Model 1. Larger for the heterogeneous error structures than the homogeneous error structure, regardless of the type of model	• Substantial differences across the true level-1 error structures (η2 = .3). Over nominal level for the homogeneous error structure. For the heterogeneous error structures, Model 1 was severely under the nominal level, while Model 2 was either close to the nominal level or slightly under the nominal level.	• No or little difference across the true level-1 error structures (η2 = .0003). Generally, the CI width for Model 1 was smaller than the CI width for Model 2.
	• One large effect for the true level-1 error structure (η2 = .88)	• Two of the design factors had a medium or large effect, including the true level-1 error structure (η2 = .83) and the series length per case (η2 = .06).	• Two interaction effects had a medium effect, including the 2-way interaction between the type of model and the true level-1 error structure (η2 = .07), and the 2-way interaction between the series length per case and the true level-1 error structure (η2 = .07)	• Three of the design factors had a large effect, including the series length per case (η2 = .47), the type of model (η2 = .34), and the number of cases (η2 = .17)

Table 4

Summary of the results for the fixed treatment effects by over-, under-, and correct-specification of the level-1 error structure

Parameter estimate	Bias	RMSE	CI coverage	CI width
Shift in level	• Close to 0 for all three types of specifications, ranging from 0 to .003	• Similar across the types of specifications, ranging from .68 to .69	• Over nominal level across all three types of specifications, ranging from .983 to .985	• Similar across the types of specifications, ranging from 4.94 to 4.99
Shift in slope	• Close to 0 for all three types of specifications, ranging from .001 to .009. The under-specified type had the smallest bias value and variability among the three types of specifications	• Similar across all three types of specifications, ranging from .22 to .23	• Over nominal level across all three types of specifications, .985 for all three types of specifications	• Similar across all three types of specifications, ranging from 1.68 to 1.71

Table 5

Summary of the results for the variance components by over-, under-, and correct-specification of the level-1 error structure

Parameter estimate	Bias	RMSE	CI coverage	CI width
Level-2 error standard deviation for shift in level	• Similar and positively biased for all three types of specifications, ranging from .85 to .86.	• Similar across all three types of specifications, ranging from 1.21 to 1.23	• Similar and over nominal level across all three types of specifications, ranging from .967 to .974 • More conservative CI coverage for the over-specified type than other types of specifications	• Similar across the three types of specifications, ranging from 6.43 to 6.48
Level-2 error standard deviation for shift in slope	• Similar and positively biased for all three types of specifications, ranging from .30 to .31.	• Similar across the three types of specifications, ranging from .41 to .42	• Slightly different and over nominal level across all three types of specifications, ranging from .970 to .982 • More conservative CI coverage for the over-specified type than other types of specifications	• Slightly different across the three types of specifications, ranging from 1.61 to 2.20 • The widest average CI width for the over-specified type among the three types of specifications.

Table 5 (continued)

Summary of the results for the variance components by over-, under-, and correct-specification of the level-1 error structure

				• For the correctly-specified models, the homogeneous error structure had smaller CI width than the heterogeneous error structures
Level-1 error standard deviation	• Different and positively biased for all three types of specifications, ranging from .034 to .056 • The over-specified type had the smallest bias and variability, and the under-specified type had the largest bias value	• Different across all three types of specifications, ranging from .14 to .27 • The over-specified type had the smallest average RMSE value and variability, and the under-specified type had the largest average RMSE value and variability	• Different, and under the nominal level for the under-specified models (M = .79), close to the nominal level for the correctly-specified models (M = .96), and over the nominal level for the over-specified models (M = .99) • The under-specified type had substantially larger variability than the other types of specifications	• Different for all three types of specifications, ranging from .47 to .74 • The under-specified type had a smaller average CI width value than the other types of specifications

Table 5 (continued)

Summary of the results for the variance components by specifications of the level-1 error structure

	• For the under-specified models, the average bias increased as the degree of heterogeneity increased. For the correctly-specified, the homogeneous error structure had the smallest average bias	• For the under-specified models, the average RMSE increased as the degree of heterogeneity increased. For the correctly-specified, the homogeneous error structure had the smallest average RMSE	• For the under-specified models, the average CI coverage tended to be under the nominal level for both heterogeneous error structures, and as the degree of heterogeneity increased, the average CI coverage value substantially decreased. For the correctly-specified models, the average CI coverage tended to be close to the nominal level, regardless of the true level-1 error structure	• For the Under-specified models, the average CI width was similar across the true level-1 error structures. For the correctly-specified models, the homogeneous error structure had the smallest average CI width
Autocorrelation	• Different and negatively biased for the both under-specified and correctly-specified models (M = -0.16, M = -0.09, respectively), but close to 0 for the over-specified models (M = 0.01)	• Different across all three types of specifications, ranging from .18 to .30	• Different, and under the nominal level for the under-specified models (M = .74), close to the nominal level for the correctly-specified (M = .94), and over the nominal level for the over-specified models (M = .99)	• Different across all three types of specifications, ranging from .73 to 1.09

Table 5 (continued)

Summary of the results for the variance components by specifications of the level-1 error structure

	• The over-specified type had the smallest bias and variability	• The over-specified type had the smallest average RMSE values and variability, and the under-specified type had the largest variability	• The under-specified type had substantially larger variability than the other types of specifications	• The under-specified type had smaller average CI width than the other types of specifications.
	• For the under-specified models, the average bias values were similar across the heterogeneous error structures. For the correctly-specified models, the average bias value was substantially smaller for the homogeneous error structure than the heterogeneous error structures	• For the under-specified models, the average RMSE values increased as the degree of heterogeneity increased. For the correctly-specified models, the homogeneous error structure had the smallest average RMSE value	• For the under-specified models, the average CI coverage tended to be under the nominal level for the heterogeneous error structures, and as the degree of heterogeneity increased, the average CI coverage decreased. For the correctly-specified models, the average CI coverage decreased as the degree of heterogeneity increased	• For the under-specified models, the average CI width was similar across the true level-1 error structures. For the correctly-specified models, the homogeneous error structure had the smallest average CI width

Based on these findings, it seemed worthwhile to explore if these results can be generalized to other important conditions that had not been covered in the main study, such as different degrees of the average level of autocorrelation, and the way of generating heterogeneity in the level-1 error structure. Therefore, two small follow up studies with fewer conditions, Study 2 and Study 3, were conducted for further exploration. More detailed information and results of Study 2 and Study 3 are provided in following sections.

Follow-Up Study: Study 2

In terms of the average autocorrelation, the main study had one level of autocorrelation (0.2), and this is the typical average autocorrelation value found in behavior studies (Shadish & Sullivan, 2011). However, other simulation work done in this area used the levels of autocorrelation 0.1, 0.2, 0.3 and 0.4 (Ferron et al., 2009; Ferron, Farmer, & Owens, 2010) which covered the possible autocorrelation values commonly found in behavior research (Huitema, 1985; Matyas & Greenwood, 1996). Thus, it was decided to examine one more level of autocorrelation, .04 in this study. More specifically, average level of autocorrelation .4 with standard deviation of .1 was selected to be examined along with selected conditions used in the main study. The autocorrelation .4 with standard deviation of .1 was selected because it creates a distribution that 99% of the autocorrelation values fall between .1 and .7 that covers the autocorrelation values typically found in behavior research (Huitema, 1985; Matyas & Greenwood, 1996; Shadish & Sullivan, 2011). Thus, in Study 2, there were 6 conditions simulated using the two factors. These factors were (1) the true level-1 error structure (homogeneous, moderately heterogeneous, and severely heterogeneous); (2) the analysis methods modeling level-1 error structure (not modeling between case variation (Model 1), and modeling between case variation (Model 2). Autocorrelation was fixed as .4 and all other factors used in the main study were also fixed; (1) the number of cases, 4 ; (2) the series length per case, 10; (3) the variation in the level-2 errors, most of the variance at level-1 (.5, .05). This yielded a 2x3 factorial design.

Results of the study

The outcomes of all of the simulated conditions for the fixed treatment effects and variance components are provided in Tables 6 and 7.

The results of the fixed treatment effects and the variance components from the main study that used 0.2 as the average autocorrelation value were very similar to the results of the fixed treatment effects and variance components from Study 2 that used 0.4 as the average autocorrelation value. The different modeling methods in level-1 error structure had little to no impact on the estimates of the fixed treatment effects, but substantial impacts on the estimates of the variance components, especially the level-1 error standard deviation and the autocorrelation parameters.

The fixed treatment effects were not biased for both Model 1 where between case variation was not modeled and Model 2 where between case variation was modeled. The average RMSE values for the fixed treatment effects were similar across the models. The interval coverage for the fixed treatment effects tended to be over the nominal level for both models. The interval width values were similar across the two models. In addition, different types of specifications (i.e., over-, under-, and correct-specification) in the level-1 error structure had little to no impact on the estimates of the fixed treatment effects.

Table 6

The results of the fixed treatment effects for Study 2

	Homogeneous		Moderately hetero		Severely Hetero	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Bias						
Intercept	-0.011	-0.010	0.015	0.014	-0.001	-0.003
Phase	-0.035	-0.040	-0.009	-0.008	0.036	0.036
Time	0.004	0.004	0.007	0.007	-0.004	-0.003
Interaction	-0.009	-0.008	-0.006	-0.006	0.011	0.011
RMSE						
Intercept	0.608	0.607	0.512	0.513	0.541	0.524
Phase	0.736	0.742	0.727	0.725	0.722	0.699
Time	0.211	0.212	0.197	0.197	0.198	0.191
Interaction	0.325	0.326	0.282	0.281	0.296	0.284
CI coverage						
Intercept	0.995	0.996	0.999	0.999	0.999	0.999
Phase	0.998	0.997	0.999	1.000	1.000	1.000
Time	0.996	0.996	0.998	0.999	0.998	0.999
Interaction	0.997	0.997	1.000	1.000	0.999	0.999
CI width						
Intercept	5.165	5.128	4.473	4.483	4.617	4.574
Phase	6.617	6.650	6.323	6.345	6.499	6.379
Time	1.669	1.657	1.552	1.549	1.595	1.550
Interaction	2.862	2.845	2.598	2.601	2.654	2.599

For the variance components, the different modeling methods in level-1 error structure had little to no impact on the estimates of the level-2 error standard deviations for phase (the shift in level) and the interaction (the shift in slope). Unlike the level-2 error standard deviations, the level-1 error standard deviation and autocorrelation show some differences in terms of the results across Model 1 and 2. The average bias and RMSE values were similar across the models, but the average CI coverage values were substantially different across the two models. The coverage was substantially under the nominal level when estimated by Model 1, but close to the nominal level when estimated by Model 2. The interval width was smaller when estimated by Model 1 than Model 2. In addition, different types of specifications in the level-1 error structure had a substantial impact on the estimates of the level-1 error standard deviation and the autocorrelation. For the average bias and RMSE values of the level-1 error standard deviation and the autocorrelation, the over-specified models had the smallest bias and RMSE values, and for the CI coverage of the level-1 error standard deviation and the autocorrelation, the correctly-specified models had coverages closest to the nominal level.

Table 7
The results of the variance components for Study 2

		Homogeneous		Moderately hetero		Severely Hetero	
		Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Bias							
Level-2 error standard deviation	Intercept	1.101	1.091	0.851	0.844	0.900	0.881
	Phase	1.542	1.550	1.409	1.408	1.468	1.424
	Time	0.325	0.324	0.288	0.285	0.300	0.287
	Interaction	0.716	0.716	0.626	0.624	0.643	0.623
Level-1 error standard deviation		0.006	0.016	0.038	0.059	0.075	0.067
Autocorrelation		-0.058	-0.116	-0.376	-0.356	-0.365	-0.331
RMSE							
Level-2 error standard deviation	Intercept	1.363	1.349	1.090	1.067	1.160	1.126
	Phase	1.806	1.804	1.650	1.636	1.733	1.670
	Time	0.396	0.395	0.363	0.355	0.376	0.357
	Interaction	0.799	0.800	0.705	0.696	0.732	0.705
Level-1 error standard deviation		0.144	0.179	0.225	0.227	0.356	0.265
Autocorrelation		0.223	0.221	0.456	0.431	0.491	0.449
CI coverage							
Level-2 error standard deviation	Intercept	0.969	0.977	0.986	0.992	0.977	0.985
	Phase	0.964	0.968	0.978	0.984	0.979	0.984
	Time	0.989	0.990	0.981	0.990	0.991	0.994
	Interaction	0.985	0.984	0.981	0.990	0.983	0.993
Level-1 error standard deviation		0.978	0.991	0.845	0.973	0.617	0.953
Autocorrelation		0.989	0.999	0.784	0.939	0.743	0.916
CI width							
Level-2 error standard deviation	Intercept	7.189	7.156	6.182	6.223	6.358	6.350
	Phase	9.198	9.245	8.736	8.780	8.972	8.850
	Time	2.247	2.242	2.075	2.077	2.134	2.088
	Interaction	2.247	3.924	2.075	3.580	2.134	3.581
Level-1 error standard deviation		0.652	0.976	0.679	1.057	0.695	1.125
Autocorrelation		1.078	1.359	1.081	1.424	1.069	1.430

These results imply that the degree of the autocorrelation had little to no impact on the estimates of the fixed treatment effects and the variance components.

Follow-Up Study: Study 3

In terms of the method of generating heterogeneity in the level-1 error structure, in the main study, data having the heterogeneous level-1 error structure had been generated in a way that every case included in the study had a unique value of the level-1 error standard deviation and autocorrelation within a specified range. However, it is possible that the values will not be evenly spread out within a specified range in a real dataset. Instead, one or more cases can have a substantial difference of the level-1 error standard deviation and the autocorrelation. For example, Baek, Petit-Bois, Van den Noortgate, Beretvas, and Ferron (2014) found that in a real dataset from a single-case study, one of the cases had a substantially larger variance compared with the other cases, which can lead to differences in the level-1 error variance and the autocorrelation. Therefore, in Study 3, data were generated in a way that one case had a substantial difference in the level-1 error variance and the autocorrelation compared to the other cases (extremely heterogeneous error structure). More specifically, one case had a 16 times bigger level-1 error variance than the other cases, and an autocorrelation that was either half or twice as large as the other cases (either .2 and .4, or .4 and .2). All other cases were generated to have a same level-1 error variance (1) and autocorrelation value (either .2 or .4). This extreme condition in which one case had 16 times the level-1 error variance of the others was selected based on the finding from Baek and Ferron (2013). They found that when they allowed the level-1 error variance to vary across cases in real datasets, the largest level-1 error variance ranges up to 16 times the smallest. Thus, in Study 3, there were 8 conditions simulated using the three factors. These factors were (1) the analysis method for modeling level-1 error structure (not modeling between case variation (Model 1), and modeling between case variation (Model 2); (2) the combination of number of cases and series length per case (4, 10 or 8, 20); and (3) the combination of level of autocorrelation for the extreme case and the rest of the cases (.2, .4 or .4, .2). All other factors used in the main study were fixed; (1) the true level-1 error structure, extremely heterogeneous; (2) the variation in the level-2 errors, most of the variance at level-1(.5, .05). This yielded a 2x2x2 factorial design.

Results of the study

The outcomes of all of the simulated conditions for the fixed treatment effects and variance components are provided in Tables 8 and 9. The results of the fixed treatment effects and the variance components from this study were different from the main study. Unlike the main study that shows the different modeling methods for the level-1 error structure had little to no impact on the estimates of the fixed treatment effects, this study found that the different modeling methods for the level-1 error structure had some impact on the estimates of the fixed treatment effects. The average bias and RMSE values were generally smaller when estimated by Model 2 where between case variation was modeled. In addition, unlike the main study that showed the different modeling methods for the level-1 error structure had little to no impact on the estimates of the level-2 error standard deviation, this study found that the different modeling methods for the level-1 error structure had some impact on the estimates of the level-2 error standard deviation. Since this study only has one type of true level-1 error structure, extremely heterogeneous, Model 1 represents the under-specified condition in that the model assumed a homogeneous level-1 error structure but the data had a heterogeneous level-1 error structure, and

Model 2 represents the over-specified condition in that the model assumed a heterogeneous level-1 error structure where everyone had their unique value of the level-1 error variance and the autocorrelation value, but the data had a heterogeneous level-1 error structure where only one case had a different level-1 error variance and autocorrelation value than others.

Table 8

The results of the fixed treatment effects for Study 3

Series length per case	Number of cases	Variation in the level-2 errors	Intercept		Phase		Time		Interaction	
			Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
			Bias							
10	4	Extreme(.2, 4)	0.025	0.020	0.030	0.008	0.006	0.010	-0.008	-0.008
		Extreme(.4, 4)	-0.003	-0.021	-0.033	-0.032	0.003	0.009	0.023	0.008
20	8	Extreme(.2, 4)	0.013	0.006	-0.005	-0.006	-0.008	-0.007	0.007	0.005
		Extreme(.4, 4)	-0.029	-0.017	-0.008	-0.006	0.005	0.002	-0.003	0.002
			RMSE							
10	4	Extreme(.2, 4)	0.920	0.702	1.440	1.029	0.342	0.255	0.626	0.410
		Extreme(.4, 4)	0.877	0.668	1.411	0.989	0.349	0.256	0.642	0.419
20	8	Extreme(.2, 4)	0.435	0.362	0.585	0.446	0.104	0.093	0.148	0.113
		Extreme(.4, 4)	0.439	0.351	0.587	0.442	0.106	0.095	0.140	0.111
			CI coverage							
10	4	Extreme(.2, 4)	0.999	1.000	0.999	1.000	0.996	0.998	0.999	1.000
		Extreme(.4, 4)	0.999	0.998	1.000	1.000	0.998	1.000	1.000	1.000
20	8	Extreme(.2, 4)	0.982	0.965	0.993	0.973	0.968	0.971	0.977	0.971
		Extreme(.4, 4)	0.977	0.971	0.990	0.989	0.959	0.964	0.976	0.968
			CI width							
10	4	Extreme(.2, 4)	7.213	6.157	11.307	9.329	2.318	1.942	5.465	4.354
		Extreme(.4, 4)	7.147	6.087	11.278	9.244	2.333	1.947	5.536	4.382
20	8	Extreme(.2, 4)	2.053	1.748	2.910	2.279	0.495	0.459	0.711	0.575
		Extreme(.4, 4)	2.066	1.766	2.942	2.328	0.491	0.459	0.684	0.569

Table 9

The results of the variance components for Study 3

Series length per case	Numb er of cases	Variation in the level-1 errors	Level-2 error standard deviation								Level-1 error standard deviation		Autocorrelation	
			Intercept		Phase		Time		Interaction		Model 1	Model 2	Model 1	Model 2
			Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2				
			Bias											
10	4	Extreme (.2, 4)	1.657	1.331	2.944	2.299	0.459	0.380	1.568	1.139	0.326	0.035	-0.374	-0.300
		Extreme (.4, 4)	1.616	1.305	2.927	2.260	0.461	0.384	1.591	1.151	0.340	0.031	-0.265	-0.197
20	8	Extreme (.2, 4)	0.189	0.125	0.459	0.251	0.038	0.038	0.103	0.051	0.292	-0.009	-0.387	-0.354
		Extreme (.4, 4)	0.204	0.137	0.493	0.286	0.035	0.038	0.082	0.045	0.290	-0.006	-0.236	-0.205
			RMSE											
10	4	Extreme (.2, 4)	2.013	1.587	3.461	2.620	0.526	0.435	1.962	1.335	1.424	0.661	0.494	0.367
		Extreme (.4, 4)	1.960	1.554	3.422	2.568	0.525	0.447	2.007	1.373	1.430	0.655	0.416	0.321
20	8	Extreme (.2, 4)	0.474	0.377	0.797	0.500	0.101	0.096	0.214	0.129	1.054	0.357	0.432	0.382
		Extreme (.4, 4)	0.487	0.382	0.818	0.505	0.100	0.098	0.190	0.123	1.052	0.359	0.301	0.258
			CI coverage											
10	4	Extreme (.2, 4)	0.946	0.988	0.927	0.974	0.986	0.998	0.893	0.985	0.055	0.932	0.674	0.972
		Extreme (.4, 4)	1.000	0.993	0.938	0.985	0.988	0.995	0.907	0.982	0.057	0.935	0.748	0.955
20	8	Extreme (.2, 4)	0.977	0.970	0.933	0.980	0.955	0.959	0.918	0.967	0.000	0.935	0.173	0.582
		Extreme (.4, 4)	0.976	0.971	0.931	0.978	0.954	0.945	0.939	0.979	0.002	0.932	0.407	0.848
			CI width											
10	4	Extreme (.2, 4)	9.707	8.549	15.197	12.950	2.907	2.556	2.907	6.199	1.283	2.099	0.970	1.436
		Extreme (.4, 4)	9.598	8.447	15.182	12.851	2.922	2.564	2.922	6.239	1.287	2.091	0.964	1.429
20	8	Extreme (.2, 4)	2.037	1.834	2.949	2.462	0.440	0.422	0.440	0.614	0.414	0.965	0.381	0.767
		Extreme (.4, 4)	2.051	1.853	3.000	2.536	0.437	0.422	0.437	0.609	0.413	0.967	0.382	0.771

Figure 150 illustrated that the average bias values of the treatment effect for phase were minimal and similar across the two models, and Model 2 (over-specified) had less variability of the bias values than Model 1(under-specified). One of the data factors, the combination of the autocorrelation of the extreme case and the autocorrelation for the others, had an impact on the average bias of the shift in level. When the extreme case had an autocorrelation of .2, which indicated that the rest of cases had an autocorrelation of .4, the average bias value for Model 1 was positive, but the average bias value for Model 2 was close to 0. In addition, Model 2 had substantially less variability of bias values than Model 1. However, when extreme case had an autocorrelation of .4, which indicated that the rest of cases had an autocorrelation of .2, the average bias values for Model 1 and Model 2 were both negative.

Similarly, Figure 151 illustrated that the average bias values of the treatment effect for the interaction were minimal and similar across the two models, but Model 2, which was the over-specified model, had less variability of the bias values than Model 1. The factor of the autocorrelation of the extreme case and others also had an impact on the average bias of the shift in slope. When extreme case had an autocorrelation of .2, which indicated that the rest of cases had an autocorrelation of .4, the average bias value for both Model 1 and 2 were close to 0, and Model 2 had less variability than Model 1. However, when the extreme case had an autocorrelation of .4, which indicated that the rest of cases had an autocorrelation of .2, the average bias values for Model 1 and Model 2 were both positive, and Model 1 had substantially larger variability than Model 2.

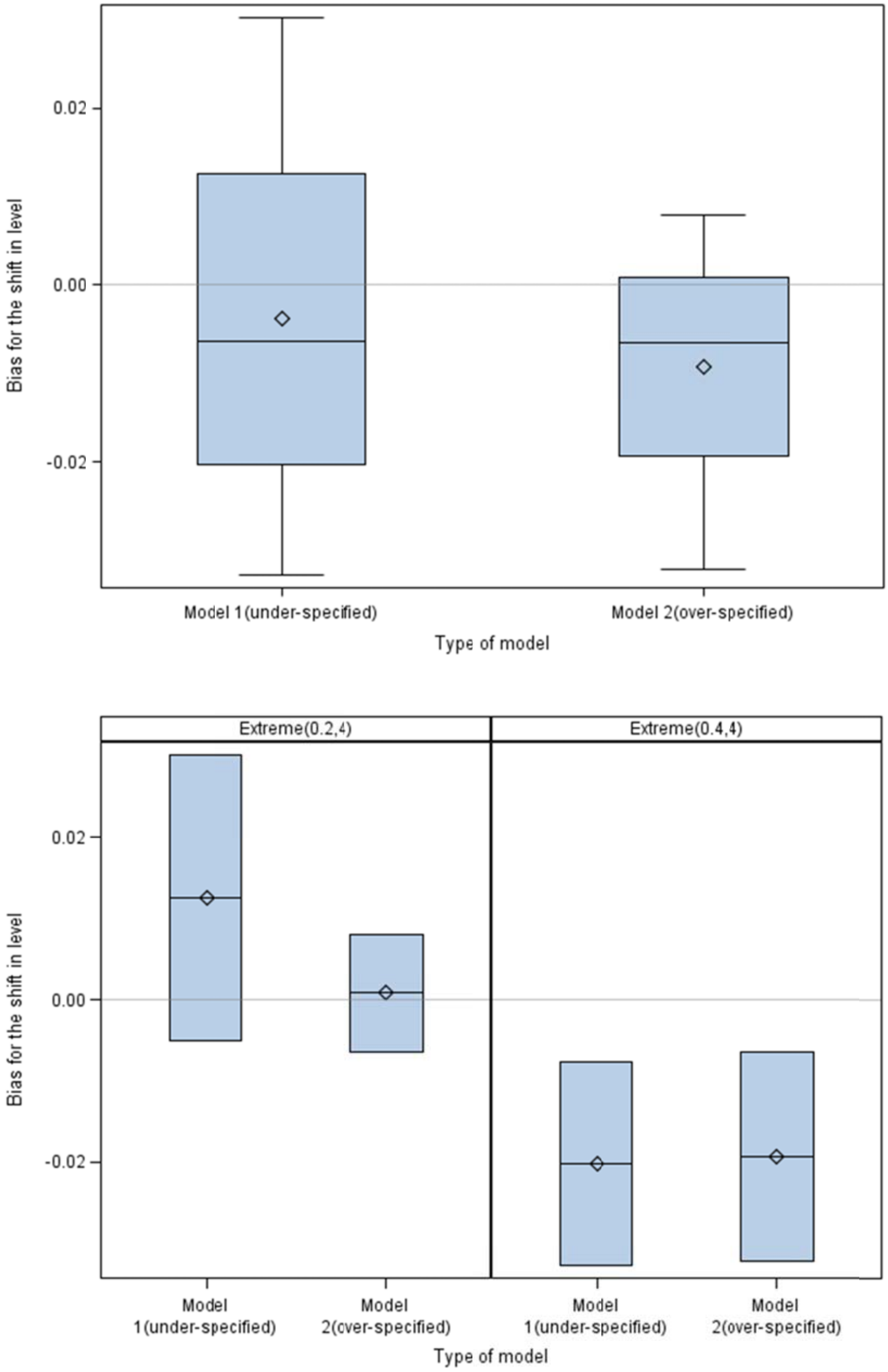


Figure 150. Box plots depicting the estimated bias of the treatment effect for shit in level

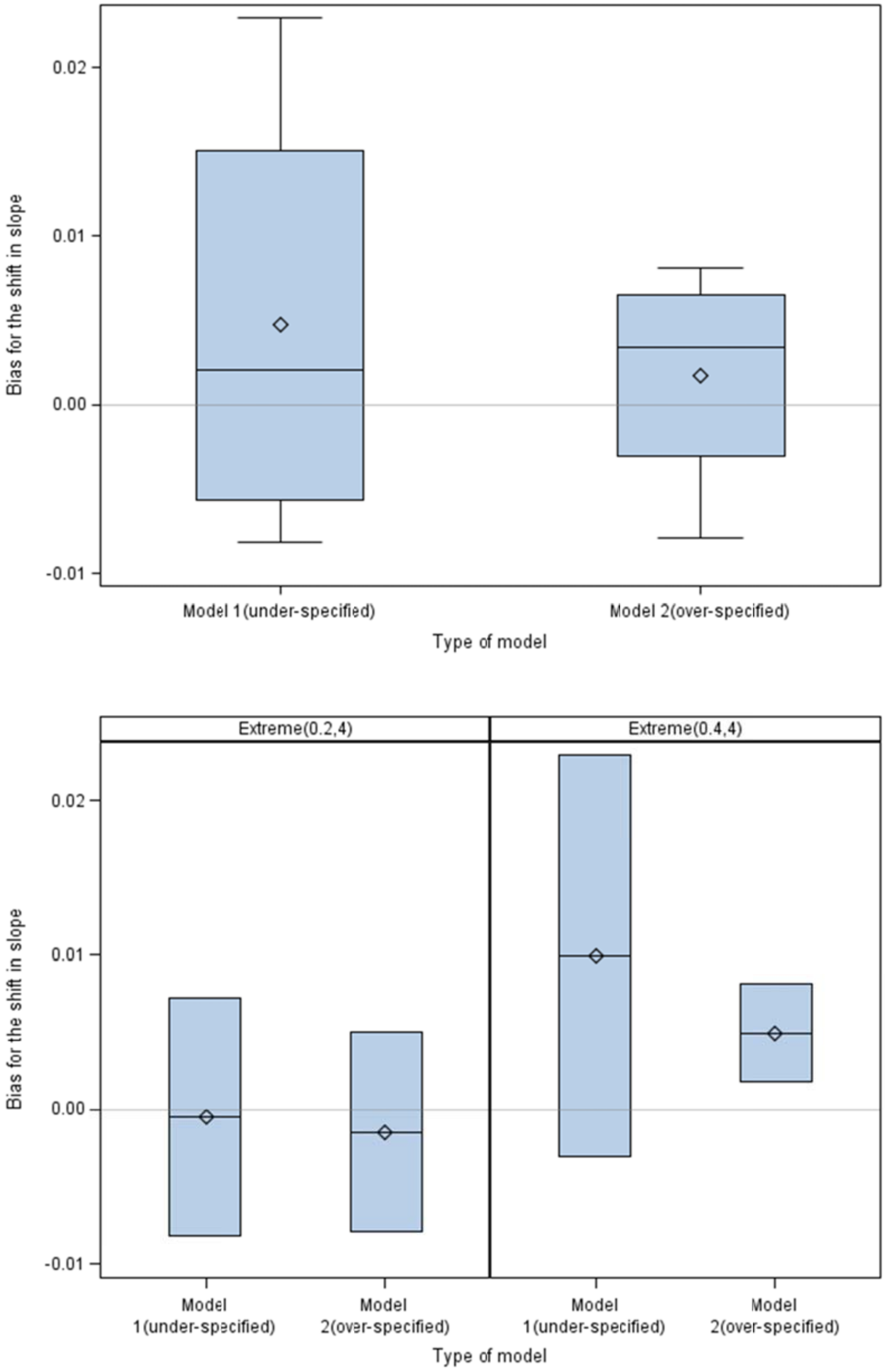


Figure 151. Box plots depicting the estimated bias of the treatment effect for shit in slope

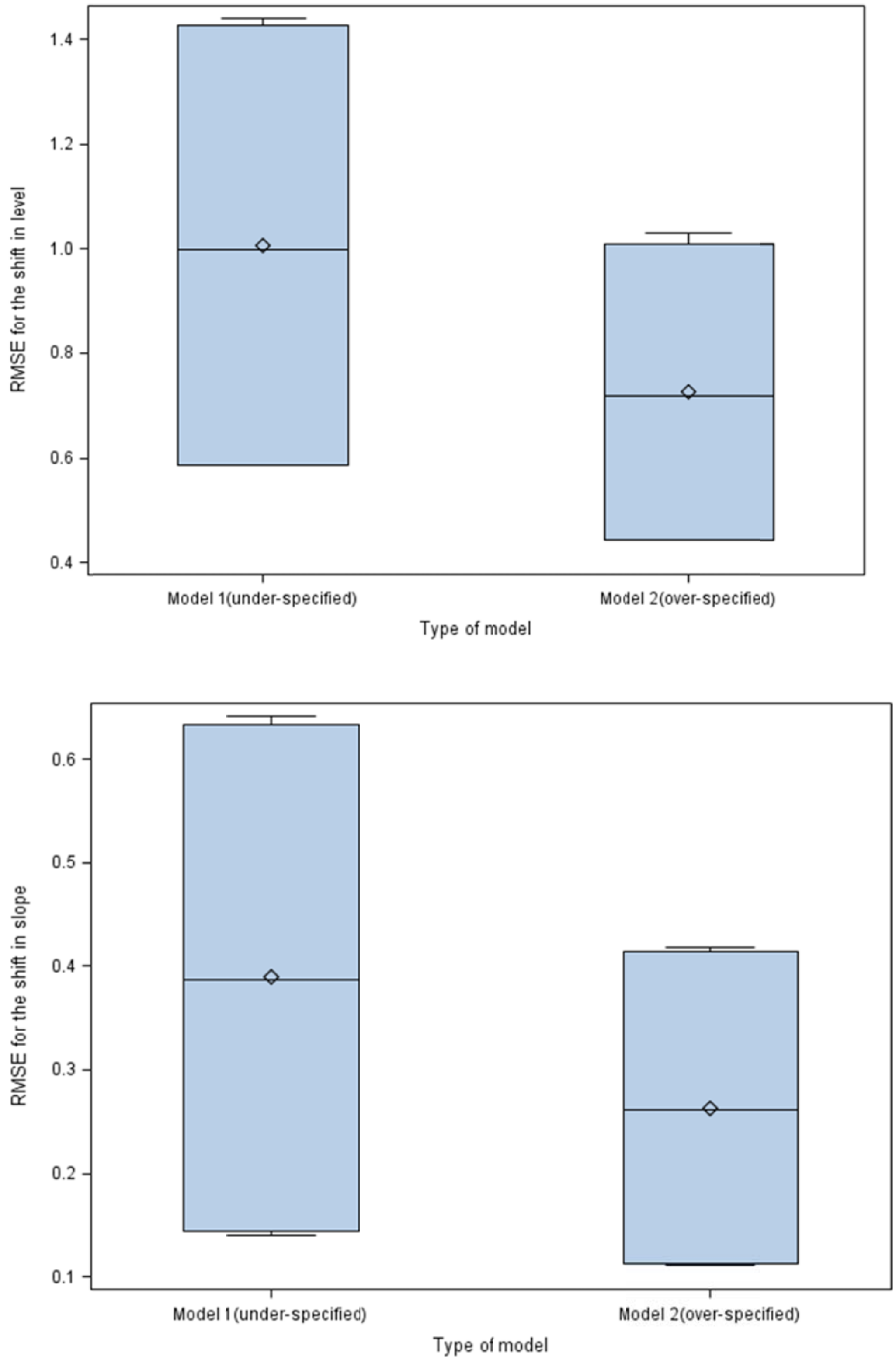


Figure 152. Box plots depicting the estimated RMSE of the treatment effect for shift in level and shift in slope

Figure 152 portrays that the average RMSE values of the treatment effect for phase and the interaction were different across the two models. Model 2 had smaller average RMSE values and less variability of the RMSE values than Model 1. These results were consistent regardless of the pairing of the autocorrelation of the extreme case and others. The rest of the outcomes, the CI coverage and the width had similar results with the results from the main study. The interval coverage for the fixed treatment effects tended to be overly conservative for both models, and the interval width values were similar across the two models.

In terms of the variance components, the average bias and RMSE values of the level-2 error standard deviation for phase and the interaction were similar across the two models. However, Model 2 had generally smaller average bias and RMSE values than Model 1. These results of the average bias and RMSE values of the level-2 error standard deviation for phase and the interaction are illustrated in Figures 153 and 154. These results were consistent regardless of the different pairings of the autocorrelation of the extreme case and others.

In addition, the CI coverage of the level-2 error standard deviation for phase and the interaction were substantially different across the two models. As illustrated in Figure 155, the CIs under covered when estimated by Model 1 for both the level-2 error standard deviation for phase and the interaction, but over covered when estimated by Model 1 for both the level-2 error standard deviation for phase and the interaction. The CI width values of the level-2 error standard deviation for phase and the interaction were similar across the two models.

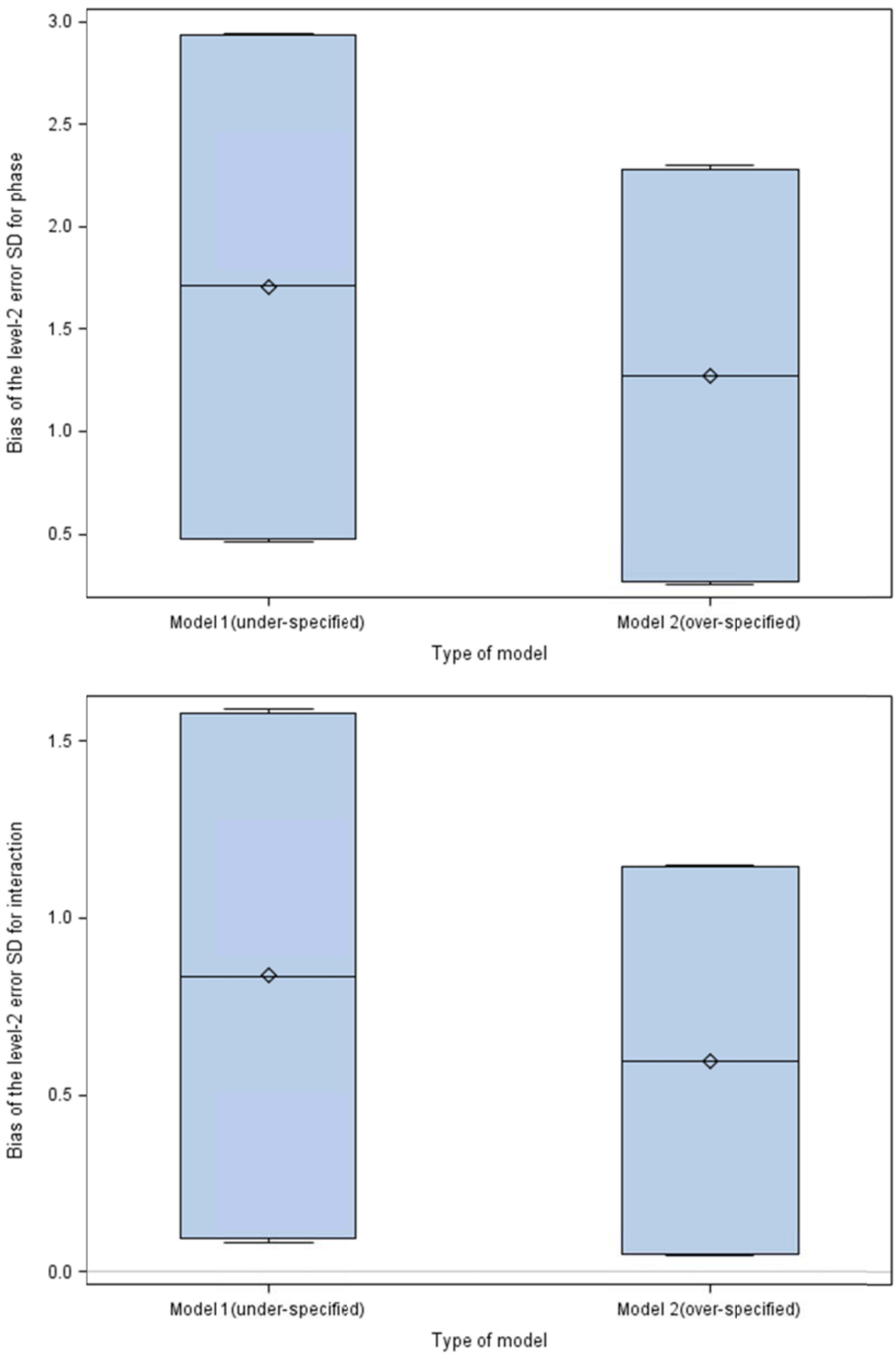
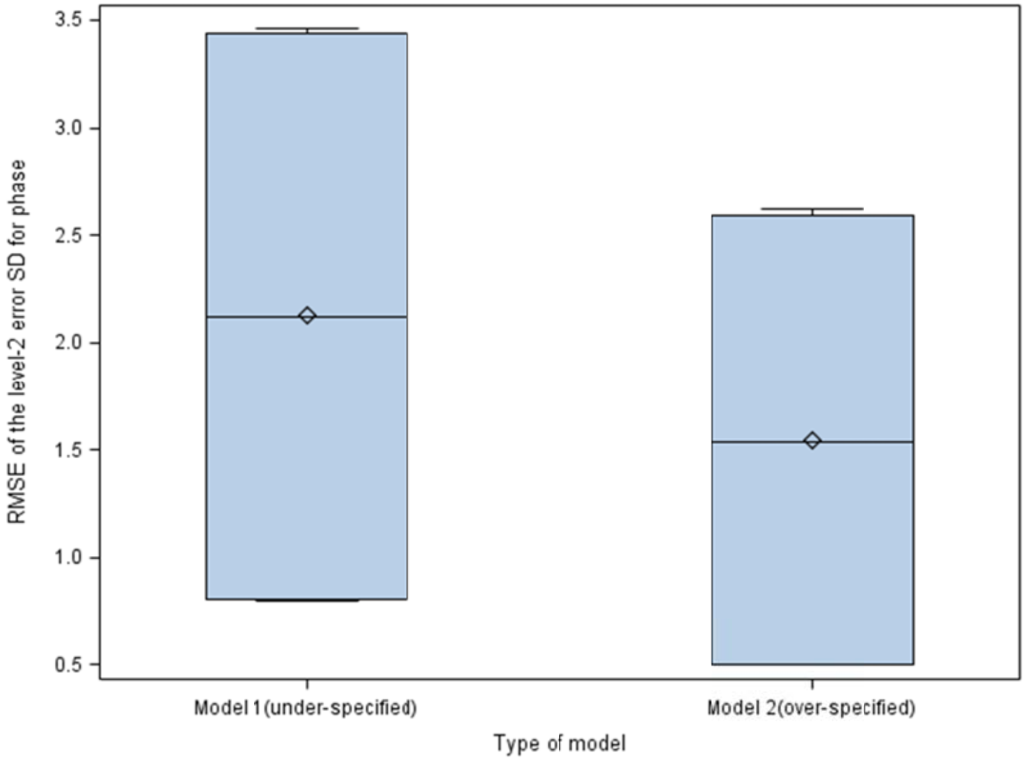


Figure 153. Box plots depicting the estimated bias of the level-2 SD for shift in level and shift in slope



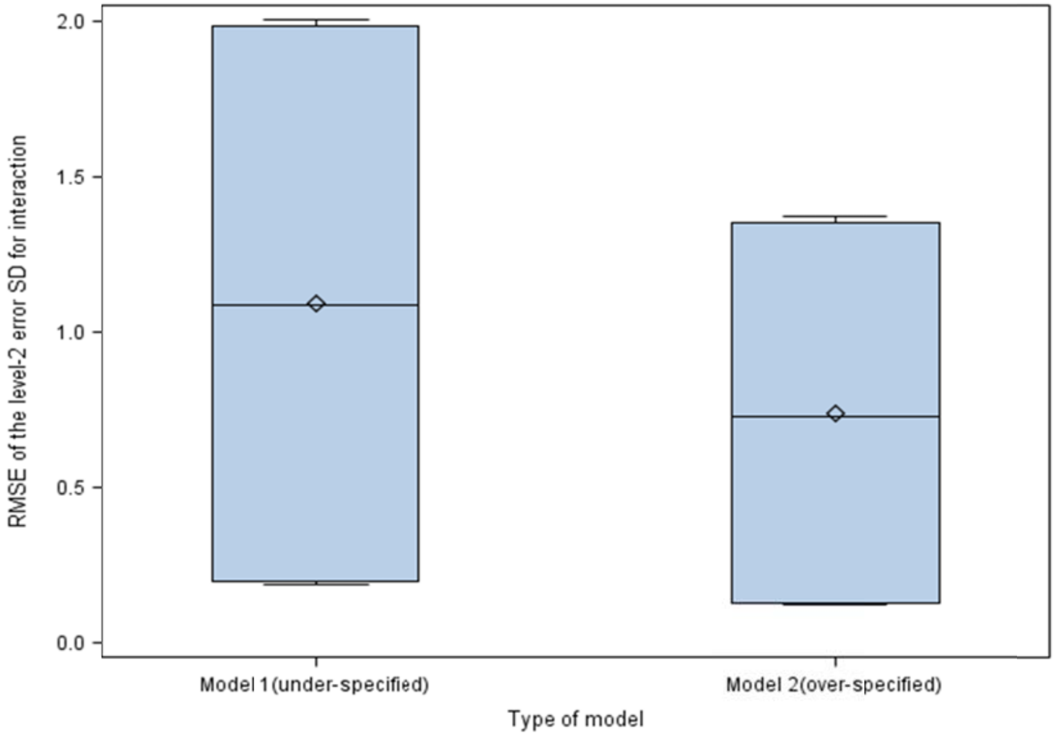


Figure 154. Box plots depicting the estimated RMSE of the level-2 SD for shift in level and shift in slope

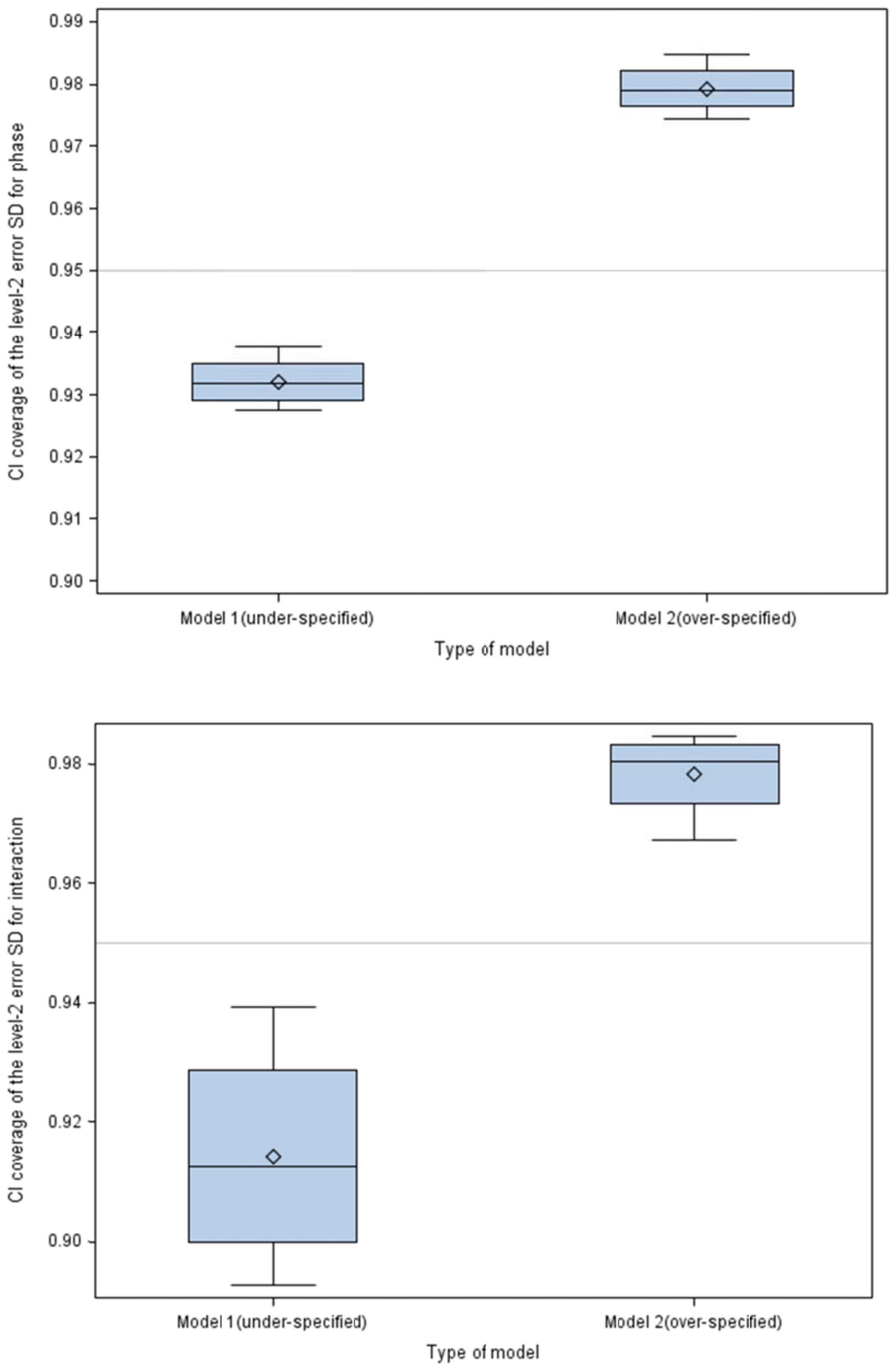


Figure 155. Box plots depicting the estimated RMSE of the level-2 SD for shift in level and shift in slope

Similar to the results from the main study, the different modeling methods in level-1 error structure had substantial impacts on the estimates of the level-1 error standard deviation and the autocorrelation. Figure 156 illustrated that the average bias values of the level-1 error standard deviation and the autocorrelation were substantially different across the two models. Model 2 had smaller average bias values than Model 1 for both the level-1 error standard deviation and the autocorrelation.

Similarly, the average RMSE values of the level-1 error standard deviation and the autocorrelation were also different across the two models. Model 2 had smaller average RMSE values than Model 1 for both the level-1 error standard deviation and the autocorrelation.

In addition, the average CI coverage of the level-1 error standard deviation and the autocorrelation were substantially different across the two models. As illustrated in Figure 157, the CIs substantially under covered when estimated by Model 1 for both the level-1 error standard deviation and the autocorrelation, but provided coverages close to the nominal level for the level-1 error standard deviation and slightly under the nominal level for the autocorrelation when estimated by Model 2.

Lastly, the interval width was smaller when estimated by Model 1 than Model 2. These results imply that the nature of the heterogeneity in the data (i.e., an outlying case versus an even spread of level-1 error variances) might impact the effect of heterogeneity on the estimates of the fixed treatment effects and the variance components.

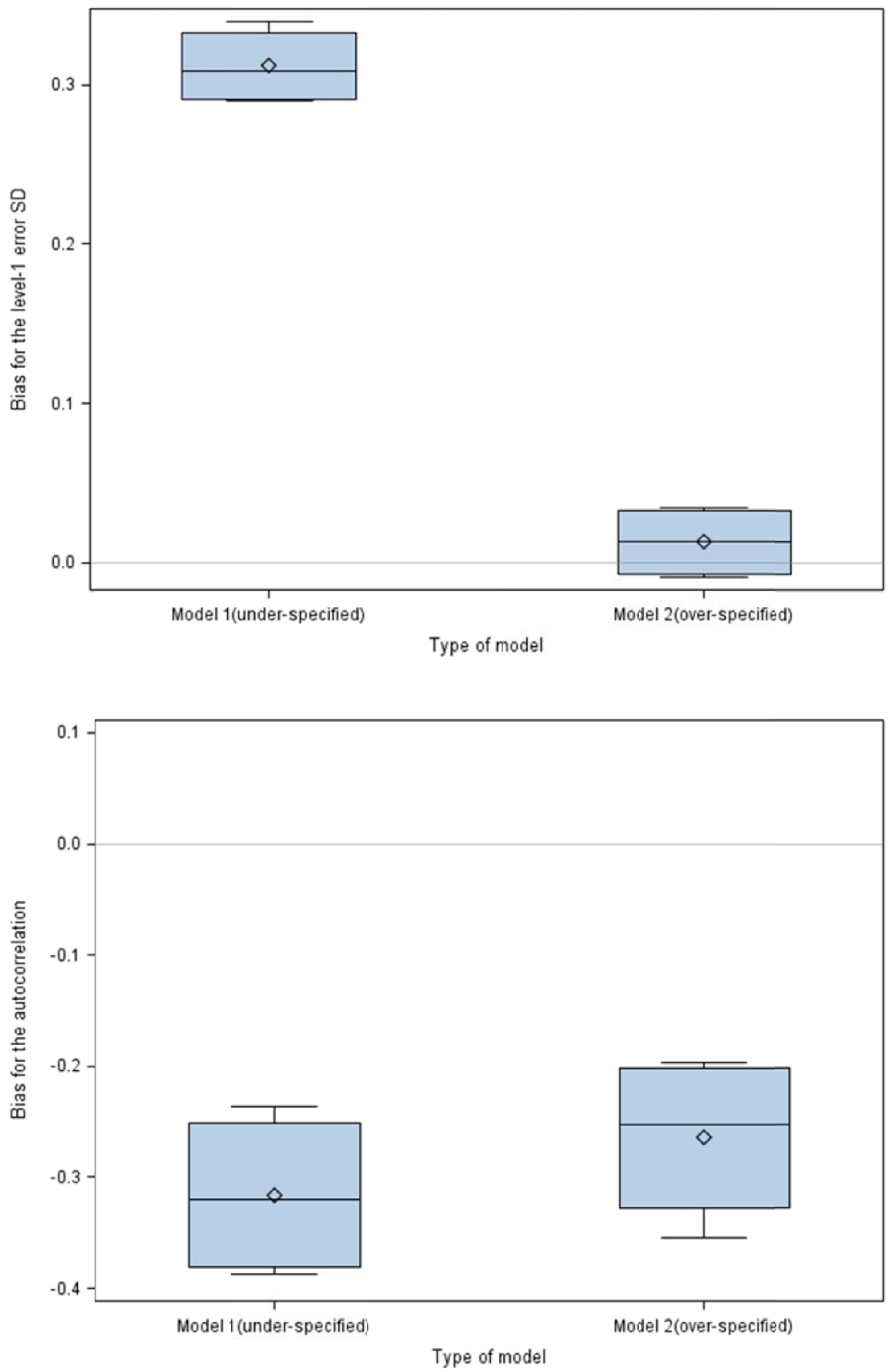


Figure 156. Box plots depicting the estimated bias of the level-1 error standard deviation and the autocorrelation

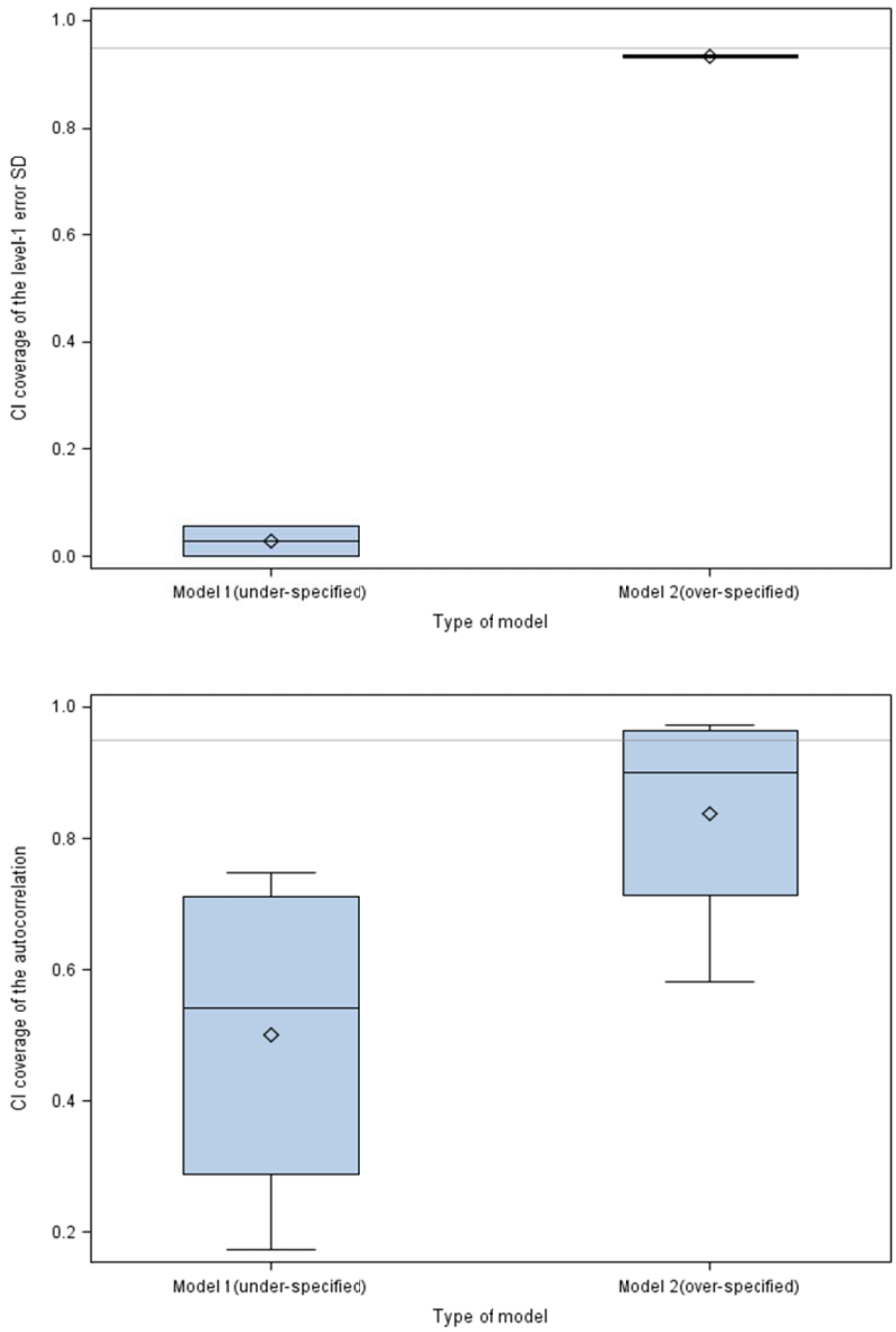


Figure 157. Box plots depicting the estimated CI coverage of the level-1 error standard deviation and the autocorrelation

CHAPTER FIVE: DISCUSSION

This chapter provides a summary of the studies and results, along with a discussion of the findings, limitations of the studies, and implications for applied single-case researchers and methodologists.

Summary of the Studies

Purpose

The purpose of the studies was to extend the MLM modeling in single-case design to allow between case variation in the level-1 error structure such that the estimated level-1 error variance and autocorrelation varies across cases, and to identify the consequences of not modeling and modeling between case variation in the level-1 error structure for single-case studies using Bayesian estimation.

The Main Study

Research questions. Research questions for the main study are following:

1. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **fixed treatment effects** in single-case design?

1) to what extent are the ***bias and RMSE for the fixed treatment effects*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation in the level-2 errors)?

2) to what extent are the ***credible interval coverage and width for the fixed treatment effects*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation in the level-2 errors)?

2. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **variance components** in single-case design?

    1) to what extent are the ***bias and RMSE for the variance components*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation in the level-2 errors)?

    2) to what extent are the ***credible interval coverage and width for the variance components*** impacted as a function of design factors (number of cases and series length per case), and data factors (true level-1 error structure and variation in the level-2 errors)?

Method. Monte Carlo simulation methods were used to address the research questions. In the study, multiple data, design and analysis factors were manipulated. This study used a 2x2x3x2x2 factorial design. These factors were the (1) number of cases (4 and 8); (2) series length per case (10 and 20); (3) true level-1 error structure (homogeneous, moderately heterogeneous, and severely heterogeneous); (4) variation in the level-2 errors (most of the variance at level-1 and most of the variance at level-2); (5) analysis methods for modeling level-

1 error structure (not modeling between case variation (Model 1), and modeling between case variation (Model 2). For each of the 48 conditions, 1,000 data sets were generated using SAS IML (SAS Institute Inc., 2008). These data sets were then analyzed using WinBUGS software.

This study first examined the convergence criteria by using a sample of simulated data sets to test convergence and to make decisions about the number of iterations, and the burn-in period. Secondly, this study examined the fixed effects (i.e., average treatment effect for phase and the interaction) and the variance components (i.e., level-2 error standard deviation for phase and the interaction, level-1 error standard deviation, and autocorrelation) in a multilevel model.

Follow-Up Study: Study 2

Research questions. Research questions for Study 2 are following:

1. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **fixed treatment effects** in single-case design?
    1) to what extent are the ***bias and RMSE for the fixed treatment effects*** impacted as a function of the true level-1 error structure when the average level of autocorrelation is .4?
    2) to what extent are the ***credible interval coverage and width for the fixed treatment effects*** impacted as a function of the true level-1 error structure when the average level of autocorrelation is .4?
2. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **variance components** in single-case design?

1) to what extent are the ***bias and RMSE for the variance components*** impacted as a function of the true level-1 error structure when the average level of autocorrelation is .4?
2) to what extent are the ***credible interval coverage and width for the variance components*** impacted as a function of the true level-1 error structure when the average level of autocorrelation is .4?

Method. Monte Carlo simulation methods were used to address the research questions. In the study, multiple data and analysis factors were manipulated. This study used a 2x3 factorial design. These factors were (1) true level-1 error structure (homogeneous, moderately heterogeneous, and severely heterogeneous); (2) analysis method for modeling the level-1 error structure (not modeling between case variation (Model 1), and modeling between case variation (Model 2)). Autocorrelation was fixed as .4 and all other factors used in the main study were also fixed; (1) the number of cases, 4 ; (2) the series length per case, 10; (3) the variation in the level-2 errors, most of the variance at level-1 (.5, .05). For each of the 6 conditions, 1,000 data sets were generated using SAS IML (SAS Institute Inc., 2008). These data sets were then analyzed using WinBUGS software.

This study examined the fixed effects (i.e., average treatment effect for phase and the interaction) and the variance components (i.e., level-2 error standard deviation for phase and the interaction, level-1 error standard deviation, and autocorrelation) in a multilevel model.

Follow-Up Study: Study 3

Research questions. Research questions for Study 3 are following:

1. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **fixed treatment effects** in single-case design?

    1) to what extent are the ***bias and RMSE for the fixed treatment effects*** impacted as a function of the pairing of the number of cases and series length per case (4, 10 or 8, 20), and the pairing of the level of autocorrelation for the extreme case and the rest of the cases (.2, .4 or .4, .2) when the true level-1 error structure is characterized as having one case with variance that is 16 times the variance of the other cases (extremely heterogeneous)?

    2) to what extent are the ***credible interval coverage and width for the fixed treatment effects*** impacted as a function of the pairing of the number of cases and series length per case (4, 10 or 8, 20), and the pairing of the level of autocorrelation for the extreme case and the rest of the cases (.2, .4 or .4, .2) when the true level-1 error structure is characterized as having one case with variance that is 16 times the variance of the other cases (extremely heterogeneous)?

2. What are the consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the **variance components** in single-case design?

    1) to what extent are the ***bias and RMSE for the variance components*** impacted as a function of the pairing of the number of cases and series length per case (4, 10 or 8, 20), and the pairing of the level of autocorrelation for the extreme case and the rest of the cases (.2, .4 or .4, .2) when the true level-1 error structure is

characterized as having one case with variance that is 16 times the variance of the other cases (extremely heterogeneous)?

2) to what extent are the ***credible interval coverage and width for the variance components*** impacted as a function of the pairing of the number of cases and series length per case (4, 10 or 8, 20), and the pairing of the level of autocorrelation for the extreme case and the rest of the cases (.2, .4 or .4, .2) when the true level-1 error structure is characterized as having one case with variance that is 16 times the variance of the other cases (extremely heterogeneous)?

Method. Monte Carlo simulation methods were used to address the research questions. In the study, multiple data and analysis factors were manipulated. This study used a 2x2x2 factorial design. These factors were (1) analysis method for modeling the level-1 error structure (not modeling between case variation (Model 1), and modeling between case variation (Model 2)); (2) the pairing of the number of cases and series length per case (4, 10 or 8, 20); (3) the pairing of the level of autocorrelation for the extreme case and the rest of the cases (.2, .4 or .4, .2). All other factors used in the main study were fixed; (1) the true level-1 error structure, extremely heterogeneous such that one case has variance that is 16 times the variance of the other cases; (2) the variation in the level-2 errors, most of the variance at level-1(.5, .05). For each of the 8 conditions, 1,000 data sets were generated using SAS IML (SAS Institute Inc., 2008). These data sets were then analyzed using WinBUGS software.

This study examined the fixed effects (i.e., average treatment effect for phase and the interaction) and the variance components (i.e., level-2 error standard deviation for phase and the interaction, level-1 error standard deviation, and autocorrelation) in a multilevel model.

Discussion of the Studies Results

Convergence

As the complexity of the model increased, such that the model required more parameters to be estimated, a longer iteration run was required. Therefore, when the data were analyzed by Model 2, it required more iterations than when the data were analyzed by Model 1. In addition, the parameters that required the most iterations to meet the convergence criteria were the level-2 error standard deviation parameters, especially the level-2 error standard deviation of the phase parameter. Based on the pilot test of a sample of simulated data sets, this study used a burn-in of 2,000 iterations and ran an additional 500,000 iterations, but only used 50,000 samples of the 500,000 iterations to form the posterior distribution for all analyses by thinning at every 10th of the iterations.

All convergence criteria that were used in this study (i.e., trace and history plots, Kernel density plots, Brooks–Gelman–Rubin (BGR) plots) were met for all tested datasets and for all estimated parameters. The computational accuracy index, MC error, was also satisfied in that it was less than .05 for all tested datasets and all estimated parameters. Convergence rates that indicated the number of samples that completely analyzed each condition were over 97% for all 48 conditions.

The Main Study

Fixed treatment effects. The consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the fixed effects were examined in terms of four outcome measures: bias, RMSE, credible interval coverage and width. The results indicated that for the treatment effects, the shift in level and the shift in slope, the average bias values were close to 0, regardless of modeling (Model 2) and not modeling (Model 1) between case variation in the level-1 error structure. In addition, there were no design factors that had meaningful effects on the average bias values of the shift in level and shift in slope. The unbiased fixed effect estimates found in the current study are consistent with the previous research regarding the inferences made from the fixed effects in both the two-level and the three-level models (Ferron et al., 2009; Owen, 2011; Merlande, 2014; Ferron, Dailey, & Yi, 2002; Kwok et al., 2007).

Similarly, the average RMSE values for both treatment effects were similar across the two models. However, the average RMSE values were impacted by three of the design factors, the number of cases, the series length per case, and the variation in the level-2 errors. As the number of cases and the series length per case increased, the average RMSE values decreased. As the variation in the level-2 errors shifted from most of the variance at level-2 to most of the variance at level-1, the average RMSE values decreased.

An examination of the credible interval coverage indicated that the average interval coverages tended to be over the nominal level for both models. There were two design factors that had impact on the average credible interval coverage. As the number of cases increased, the average credible interval coverage for both treatment effects approached the nominal level. As the series length per case increased, the average credible interval coverage for the shift in slope approached the nominal level. The analysis of the credible interval width revealed that the average credible interval width was similar across the two models.

These findings from the fixed effects suggest that if possible, researchers should increase their level-2 and level-1 sample sizes (number of cases and series length per case). In addition, these findings are consistent with previous literature related to two level or three level models for single-case data that states larger numbers of upper level units lead to greater accuracy and precision (Ferron et al., 2009; Merlande, 2014; Owen, 2011).

In addition, an exploration of the different types of specifications (under-specified, correctly-specified, and over-specified) in the level-1 error structure revealed that the different types of specifications had little to no impact on the estimates of the fixed effects. The average bias values were close to 0, regardless of the different types of specifications in the level-1 error structure. The average RMSE values were similar across the three types of specifications, the average interval coverages were over the nominal level for all three types of specifications, and the average interval widths were similar across all three types of specifications. These results also supported the findings of previous multilevel modeling and the latent growth curve modeling work which showed that the estimates of the fixed effects appear not to be biased by the misspecification of the level-1 error structure tests of the fixed effects (Ferron, 2002; Kwok et al., 2007; Merlande, 2014; Sivo, Fan & Witta, 2005). One interesting finding is that the interval coverages were consistently over the nominal level across models and across model specifications. This finding is different from studies that have examined REML estimation of multilevel models for single-case data (e.g., Shadish & Rindskopf, 2007; Shadish, Rindskopf, & Hedges, 2008; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008; Ferron et al., 2009; Merlande, 2014; Owen, 2011), where the CI coverage is very close to the nominal level across conditions, but can be explained by an impact of the Bayesian estimation method. Baek, Petit-Bois, and Ferron (2014) found that there was an impact of estimation method (REML versus Bayesian) on estimating multilevel models for single-case studies. Specifically for the average interval coverage, Baek and her colleague found that the average CI coverage rates for the fixed effects tended to be over the nominal level when using the Bayesian estimation method, while they tended to be close to the nominal level or slightly under when using the REML estimation method.

Variance components. The consequences of modeling and not modeling between case variation in the level-1 error structure in terms of estimation of the variance components were examined in terms of four outcome measures: bias, RMSE, credible interval coverage and width. The results indicated that the level-2 error standard deviation estimates for shift in level and shift in slope were positively biased for both Model 1 and 2. Two design factors, the number of cases and the series length per case, had some impact on the estimates of the level-2 error standard deviation. As the number of cases increased, the average bias of the level-2 error standard deviation for shift in level was decreased. The impact of the series length per case on the average bias of the level-2 error standard deviation for shift in slope was dependent on the number of cases. As the number of cases increased, the impact of the series length per case on the average bias of the level-2 error standard deviation for shift in slope decreased. These findings are consistent with the previous studies that had generally found a substantial bias in the variance components across the various conditions (Kwok et al., 2007; Murphy & Pituch, 2009; Ferron et al., 2009; Merlande, 2014; Owen, 2011). These findings also suggest that as the number of upper units increased, the impact of the number of lower units decreased. Thus, if possible, researchers should try to increase their level-2 units sample size. These results were also supported by the previous work that had revealed the variance components were more biased when the number of cases and the series length per case was small (Kwok et al., 2007; Murphy & Pituch, 2009; Ferron et al., 2009). The impact of the upper units sample size on the bias estimate of the variance components, related with the treatment effects, seems to be showing more in the two-level model studies. Previous studies with the three-level single-case models had not found explicitly this relationship between the upper level unit sample size with the bias estimate of the variance components for the treatment effects (Merlande, 2014; Owen, 2011).

Similarly, the average RMSE values of the level-2 error standard deviation for shift in level and shift in slope were similar across the two models. Three of the design factors, the number of cases, the series length per case, and the variation in the level-2 errors had some impact on the estimates of the level-2 error standard deviation. As the number of cases increased, the average RMSE values for both level-2 error standard deviations decreased. As the variation in the level-2 errors shifted from most of the variance at level-2 to most of the variance at level-1, the average RMSE value of the level-2 error standard deviation for shift in level decreased. As the series length per case increased, the average RMSE value of the level-2 error standard deviation for shift in slope decreased.

An examination of the credible interval coverage of the level-2 error standard deviation for shift in level and shift in slope indicated that the credible interval coverages tended to be over the nominal level for both models. Four of the design factors, the variation in the level-2 errors, the type of model, the series length per case, and the true level-1 error structure, had a meaningful impact on the average interval coverage of the level-2 error standard deviations. Generally, Model 2 was more conservative than Model 1 in that it had coverage estimates that were further above the nominal level. As the variation in the level-2 errors shifted from most of the variance at level-1 to most of the variance at level-2, the average credible interval coverage of the level-2 error variance for shift in level approached the nominal level. An impact of the variation in the level-2 errors on the average interval coverage of the level-2 error standard deviation for shift in slope was dependent on the true level-1 error structure. As the variation in the level-2 errors shifted from most of the variance at level-1 to most of the variance at level-2, the average credible interval coverage of the level-2 error variance for shift in slope approached the nominal level across all three true level-1 error structures. However, the average interval coverage in the moderately heterogeneous error structure was impacted the most by the variation in the level-2 errors. The analysis of the credible interval width for the level-2 error standard deviations revealed that the average credible interval widths were similar across the two models. Three of the design factors had some impact on the credible interval width of the level-2 error standard deviations. As the number of cases increased, and the variation in the level-2 errors shifted from most of the variance at level-2 to most of the variance at level-1, the average width of the CIs for the level-2 error standard deviations decreased. As the series length per case increased, the average width of the CIs for the level-2 error standard deviation for shift in slope decreased. Similar to the results of the CI coverages in the fixed effects, the CI coverages of the level-2 standard deviations were over the nominal level. This finding is not consistent with the previous work that had found the CI coverages of the level-2 error variances were under the nominal level. Both of the studies with the two-level models (Ferron et al., 2009) and the three-level models (Owen, 2011; Merlande, 2014) of the single-case data using the REML estimation method had found that the CI coverages of the level-2 error variances were generally under the nominal level. This contradictory finding of the current study could also be explained by an impact of the Bayesian estimation method. Although Baek and her colleagues (2014) had not explicitly looked at the average CI coverage rates for the variance components, given the impact of the Bayesian estimation method on the average CI coverage rates for the fixed effects, it would seem reasonable to assume that there could be an impact of the Bayesian estimation method on the CI coverage rates for the variance components.

In addition, an exploration of the different types of specifications (under-specified, correctly-specified, and over-specified) in the level-1 error structure revealed that the different types of specifications had little to no impact on the estimates of the level-2 error standard deviations. The level-2 error standard deviation for shift in level and shift in slope were similar and positively biased across the three types of specifications in the level-1 error structure. The average RMSE values were similar across the three types of specifications, the average interval coverages were over the nominal level for all three types of specifications, and the average interval widths were similar across all three types of specifications. Although the interval coverages and widths were similar across all three types of specifications, the over-specified type generally had higher coverage probabilities (more conservative) and wider interval widths than the other types of specifications. These results also supported the findings of the previous work, with the three-level model of the single-case data, that had found the bias of the level-2 error variances were comparable across the different types of the specifications (Merlande, 2014). Overall, these findings suggest that the different modeling in the level-1 error structures had no or little impact on the estimates of the level-2 error standard deviations.

Unlike the level-2 error standard deviations, the results for the level-1 error standard deviation and the autocorrelation indicated that different modeling of the level-1 error structure had a substantial impact on the estimates of the level-1 error standard deviation and the autocorrelation. Consistent to the previous research on the two-level and the three-level models, with the single-case data, that had found the level-1 error variance was biased (Ferron et al., 2009; Merlande, 2014; Owen, 2011), the average level-1 error standard deviations of the current study were similar and positively biased for both models. However, there were some differences between Model 1 and Model 2 within the true level-1 error structure. For Model 1, the bias of the level-1 error standard deviation increased constantly as the true level-1 error structure shifted from the homogeneous to the moderately heterogeneous to the severely heterogeneous error structure. However, for Model 2, the bias of the level-1 error standard deviation increased as the true level-1 error structure shifted from the homogeneity to the moderately heterogeneous error structure, but decreased as the true level-1 error structure shifted from the moderately heterogeneous to the severely heterogeneous error structure. The analysis of the average RMSE value indicated that there was a difference across the two models. There were substantial differences between Model 1 and Model 2 within the true level-1 error structure. For the homogeneous error structure, the average RMSE value was larger when estimated by the Model 2 than Model 1, but for the heterogeneous error structures, the average RMSE values were smaller when estimated by Model 2 than Model 1. In addition, as the series length per case increased, the average RMSE value decreased regardless of the type of models.

An examination of the average credible interval coverage revealed that there were substantial differences between the two models across the true level-1 error structures. For the homogeneous error structure, the average credible interval coverage was over the nominal level across the two models. For both heterogeneous error structures, the average credible interval coverage was substantially under the nominal level for Model 1, but either approached the nominal level or was slightly over the nominal level for Model 2. In addition, as the series length per case increased, the interval coverage decreased. Previous studies also had found impact of the series length per case on the CI coverage of the level-1 error variance (Merlande, 2014).

The analysis of the CI width indicated that the CI width for Model 1 was smaller than the CI width for Model 2. Moreover, as the number of cases and the series length per case increased, the CI width decreased. These findings are consistent with the previous work that had found the

CI width of the level-1 error variance decreased as the series length per case and the number of cases increased (Merlande, 2014; Owen, 2011; Ferron et al., 2009).

The results of the autocorrelation were very similar with the results of the level-1 error standard deviation. The autocorrelation values were similar and negatively biased for both models, which is consistent with the previous work that had found the estimate of the autocorrelation was generally biased (Ferron et al., 2009; Merlande, 2014). However, there were substantial differences across the true level-1 error structures. The autocorrelation values were more biased when the true level-1 error structure was one of the heterogeneous error structures than the homogeneous error structure. In addition, the autocorrelation parameter tended to be slightly more biased when estimated by Model 1 than Model 2 for all three types of the level-1 error structures.

Similarly, the analysis of the average RMSE value indicated that the average RMSE values of the autocorrelation were similar across the two models. However, there were substantial differences across the true level-1 error structures. The average RMSE value was larger for the heterogeneous error structures than the homogeneous error structure, regardless of the type of model. In addition, for the homogeneous error structure, the average RMSE value was larger when estimated by Model 2 than Model 1, but for the heterogeneous error structures, the average RMSE values were smaller when estimated by Model 2 than Model 1. Moreover, as the series length per case increased, the average RMSE value decreased regardless of the type of model.

An examination of the average credible interval coverage revealed that there were substantial differences between the two models across the true level-1 error structures. For the homogeneous error structure, the average credible interval coverage was over the nominal level across the two models. For the heterogeneous error structures, the average credible interval coverage was substantially under the nominal level for Model 1, but either approached the nominal level or was slightly over the nominal level for Model 2. In addition, an impact of the series length per case on the CI coverage was dependent on the true level-1 error structure. The impact of the series length per case was smaller when the true level-1 error structure was the homogeneous error structure than one of the heterogeneous error structures, and as the severity of the heterogeneity in the level-1 error structure increased, the impact of the series length per case increased greatly.

The analysis of the CI width indicated that the average CI width for Model 1 was smaller than the average CI width for Model 2. Moreover, as the number of cases and the series length per case increased, the CI width decreased.

These findings from the level-1 error standard deviation and autocorrelation indicated that Model 2 provides better estimates of some of the variance components when analyzing data that are severely heterogeneous. These findings also suggest that researchers should model between case variation in the level-1 error structure when they analyze data that have a severely heterogeneous level-1 error structure.

In addition, an exploration of the different types of specifications (under-specified, correctly-specified, and over-specified) of the level-1 error structure revealed that the different types of specifications had substantial impacts on the estimates of the level-1 error standard deviation and the autocorrelation. The level-1 error standard deviation was different and positively biased across the three types of specifications in the level-1 error structure. The over-specified type had the smallest bias and variability, and the under-specified type had the largest bias value among the three types of specifications. Similarly, the average RMSE values were different across the three types of specifications. The over-specified type had the smallest average RMSE value and variability, and the under-specified type had the largest average RMSE value and variability. The average interval coverage was under the nominal level for the under-specified type, close to the nominal level for the correctly-specified type, and over the nominal level for the over-specified type. The average interval width was smaller for the under-specified type than other types of specifications. These findings indicate that the estimates of the level-1 error standard deviation are better when the level-1 error structure is either correctly-specified or over-specified, rather than under-specified. These findings were also consistent with the findings from the previous work which showed that the correctly-specified, and the over-specified, level-1 error structures tended to work better than the under-specified level-1 error structure, in terms of the estimates and inferences of the variance components in a multilevel model (Kwok et al., 2007; Merlande, 2014; Sivo, Fan & Witta, 2005).

The results for the autocorrelation were very similar with the results for the level-1 error standard deviation. The autocorrelation was negatively biased for both the under-specified and the correctly-specified, but was close to 0 for the over-specified type. Similarly, the over-specified type had the smallest average RMSE value and variability, and the under-specified type had the largest average RMSE value and variability. The average interval coverage was under the nominal level for the under-specified type, close to the nominal level for the correctly-specified type, and over the nominal level for the over-specified type. The average interval width was smaller for the under-specified type than other types of specifications. These findings indicate that the estimates of the autocorrelation are better when the level-1 error structure is either the correctly-specified or over-specified, as opposed to under-specified. These findings were also consistent with the findings, from the previous work, which showed that the correctly-specified and the over-specified level-1 error structures, tended to work better than the under-specified level-1 error structure, in terms of the estimates and inferences of the variance components in a multilevel model (Kwok et al., 2007; Merlande, 2014; Sivo, Fan & Witta, 2005).

These findings from the level-1 error standard deviation and the autocorrelation also suggest that researchers should try to select either a correctly-specified or over-specified level-1 error structure rather than an under-specified level-1 error structure when they run a multilevel modeling for single-case data.

Follow-Up Study: Study 2

The results of the fixed effects and the variance components from the main study that used 0.2 as the average autocorrelation value were very similar to the results of the fixed effects and variance components from Study 2 that used 0.4 as the average autocorrelation value. The different modeling methods for the level-1 error structure had little to no impact on the estimates of the fixed effects, but had a substantial impact on the estimates of the variance components, especially the level-1 error standard deviation and the autocorrelation parameters.

Fixed treatment effects. Fixed effects were analyzed in terms of bias, RMSE, credible interval coverage and widths. The estimates of the shift in level and the shift in slope were not biased for either Model 1 or Model 2. The average RMSE values for the shift in level and the shift in slope were similar across the models. The confidence intervals for the shift in level and the shift in slope tended to be overly conservative for both models, producing coverage probabilities above the nominal level. The interval widths were similar across the two models. In addition, different types of specifications in the level-1 error structure had little to no impact on the estimates of the shift in level and the shift in slope.

Variance components. Variance components were also analyzed in terms of bias, RMSE, credible interval coverage and widths. For the variance components, the different modeling methods in the level-1 error structure had little to no impact on the estimates of the level-2 error standard deviations for phase and the interaction. Unlike the level-2 error standard deviations, the level-1 error standard deviation and autocorrelation showed some differences in terms of the results across Model 1 and Model 2. The average bias and RMSE values were similar across the models, but the average CI coverage values were substantially different across the two models. The coverage probabilities were substantially under the nominal level when estimated by Model 1, but close to the nominal level when estimated by Model 2. The interval width was smaller when estimated by Model 1 than estimated by Model 2. In addition, different types of specifications of the level-1 error structure had a substantial impact on the estimates of the level-1 error standard deviation and the autocorrelation. For the average bias and RMSE values of the level-1 error standard deviation and the autocorrelation, the over-specified type had the smallest bias and RMSE values. For the CI coverage of the level-1 error standard deviation and the autocorrelation, the correctly-specified type works the best, in that it was the closest to the nominal level. These findings imply that the degree of the autocorrelation had little to no impact on the relative performance of the two models regarding the estimates of the fixed effects and the variance components.

Follow-Up Study: Study 3

The results of the fixed effects and the variance components from this study were different from the main study. Unlike the main study that showed the different modeling methods for the level-1 error structure had little to no impact on the estimates of the fixed effects, this study found that the different modeling methods in the level-1 error structure had some impact on the estimates of the fixed effects. The average bias and RMSE values were generally smaller when estimated by Model 2 than Model 1. Unlike the main study that showed the different modeling methods for the level-1 error structure had little to no impact on the estimates of the level-2 error standard deviations, this study found that the different modeling methods for the level-1 error structure had some impact on the estimates of the level-2 error standard deviations, along with the level-1 error standard deviation and the autocorrelation.

Fixed treatment effects. Fixed effects were analyzed in terms of bias, RMSE, confidence interval coverage and widths. The average bias values for the shift in level and the shift in slope were minimal and similar across the two models, but unlike the results from the main study, Model 2 (over-specified) had substantially less variability of the bias values than Model 1(under-specified). One of the design factors, the pairing of the autocorrelation of the extreme case and others, had an impact on the average bias of the shift in level and shift in slope. For the shift in level, when the extreme case had an autocorrelation of .2, which indicated that the rest of cases had an autocorrelation of .4, the average bias value for Model 1 was positive, but the average bias value for Model 2 was close to 0. In addition, Model 2 (over-specified) had substantially less variability of the bias values than Model 1(under-specified). However, when the extreme case had an autocorrelation of .4, which indicated that the rest of the cases had an autocorrelation of .2, the average bias values for Model 1 and Model 2 were both negative.

Similarly, the average bias values for the shift in slope were minimal and similar across the two models, and Model 2 (over-specified) had less variability of the bias values than Model 1(under-specified). For the shift in level, when the extreme case had an autocorrelation of .2 which indicated that the rest of cases had an autocorrelation of .4, the average bias value for both Model 1 and 2 were close to 0, and Model 2 (over-specified) had less variability than Model 1 (under-specified). However, when the extreme case had an autocorrelation of .4, which indicated that the rest of cases had an autocorrelation of .2, the average bias values for Model 1 and Model 2 were both positive, and Model 1 (under-specified) had substantially larger variability in bias estimates than Model 2 (over-specified).

Unlike the results of the main study that showed the similar average RMSE values across the two models, the average RMSE values of the treatment effect for the shift in level and the shift in slope were different across the two models. Model 2 (over-specified) had a smaller average RMSE value and less variability of the RMSE values than Model 1 (under-specified). These results were consistent regardless of the different pairings of the autocorrelation of the extreme case and others. The rest of the outcomes, the CI coverage and the width had similar results with the results from the main study. The interval coverage for the fixed effects tended to be over the nominal level for both models, and the interval width values were similar across the two models. These findings indicate that the different modeling methods in the level-1 error structure had substantial impact on the estimates of the fixed effects when the level-1 error structure is the extremely heterogeneous level-1 error structure (i.e., one case has 16 times the variance of the other cases). Generally, Model 2 (over-specified) that models between case variation in the level-1 error structure worked better than Model 1 (under-specified) that does not model between case variation in the level-1 error structure.

Variance components. Variance components were also analyzed in terms of bias, RMSE, confidence interval coverage, and widths. The average bias and RMSE values of the level-2 error standard deviation for the shift in level and the shift in slope were similar across the two models, but Model 2 (over-specified) had a generally smaller average bias and smaller RMSE values than Model 1 (under-specified). These results were consistent, regardless of the different pairings of the autocorrelation of the extreme case and others. Unlike the results from the main study, the CI coverage of the level-2 error standard deviations for both treatment effects were substantially different across the two models. The average coverage probabilities were under the nominal level when estimated by Model 1 (under-specified), but over the nominal level when estimated by Model 2 (over-specified). The CI widths were similar across the two models.

Similar to the results from the main study, the different modeling methods in the level-1 error structure had substantial impacts on the estimates of the level-1 error standard deviation and the autocorrelation. The average bias values of the level-1 error standard deviation and the autocorrelation were substantially different across the two models. Model 2 (over-specified) had smaller average bias values than Model 1 (under-specified) for both the level-1 error standard deviation and the autocorrelation. Similarly, the average RMSE values of the level-1 error standard deviation and the autocorrelation were also different across the two models. Model 2 (over-specified) had smaller average RMSE values than Model 1 (under-specified) for both the level-1 error standard deviation and the autocorrelation. In addition, the average CI coverage of the level-1 error standard deviation and the autocorrelation were substantially different across the two models. The average CI coverage was substantially under the nominal level when estimated by Model 1 (under-specified) for both the level-1 error standard deviation and the autocorrelation, but close to the nominal level for the level-1 error variance, and slightly under the nominal level for the autocorrelation when estimated by Model 2 (over-specified). Lastly, the interval width was smaller when estimated by Model 1 (under-specified) than Model 2 (over-specified). These findings indicate that the different modeling methods in the level-1 error structure had substantial impact on the estimates of the variance components when the level-1 error structure was the extremely heterogeneous level-1 error structure. Generally, Model 2 (over-specified) that models between case variation in the level-1 error structure worked better than Model 1 (under-specified) that does not model between case variation in the level-1 error structure.

These results from Study 3 also imply that the form of heterogeneity in the data (i.e., one extreme case versus a more even spread of the level-1 variances) might have some impact on relative effectiveness of the two models for estimating the fixed effects and the variance components. In addition, these results suggest that researchers should try to model between case variation in the level-1 error structure when they analyze data that have the extremely heterogeneous structure showing one or more cases have substantially different variability than other cases.

Limitations of the Study

Since this study was conducted using the Monte Carlo simulation method, there are generalizability limitations regarding this study. Although the Monte Carlo method used in this study allowed the investigation of how various design factors can impact the parameter estimates, specific conditions (design factors) used in the study limit the generalizability of the study. The conditions were chosen based on a review of single-case literature and applied studies that were done using two-level models to analyze single-case data. The specific conditions chosen for this study, however, are only some of the possible options that could have been included in the study. Specifically, the follow up studies (Study 2 and Study 3) used only a few conditions. Therefore, the results of this study can only be generalized to studies with the same or similar conditions. Any conclusions beyond the observed conditions should be interpreted with caution.

Another limitation is related to the model specification and the types of outcome measure. First, this study assumed the outcome variable is continuous. There are various types of outcomes that are commonly used in single-case studies, such as binary, ordinal, or count outcomes which require different types of assumptions using a different distribution (e.g., Beta distribution and Poisson distribution) (Shadish & Rindskopf, 2007; Shadish et al., 2008).

In addition, the two level model used in this study only included linear trends. However, there are more complex trends (e.g., non-linear trends) that are also used in models to investigate single-case data (Shadish & Rindskopf, 2007). Moreover, this study only investigated the first-order autoregressive level-1 error structure (AR(1)). As previously mentioned, there are various complex level-1 error structures that assume the errors to be autocorrelated, such as compound symmetry, second order autoregressive, banded toeplitz, or moving average. The benefit of choosing the AR(1) model is that it is one of the simplest autocorrelated level-1 error structures, and is the most commonly studied and applied the correlated error structure for the time series data (Velicer & Fava, 2003; West & Hepworth, 1991), and, therefore, the most logical for an initial study into modeling between case variation in the level-1 error structure

Implications of the Study

Although single-case researchers have recognized the misspecification effect of level-1 error structures on statistical inferences of multilevel models, researchers have overlooked how they have made a critical homogeneity assumption about the level-1 error structure in their studies. This study provides insight about how not modeling and modeling between case variation in the level-1 error structure, a misspecification issue of the level-1 error structure, impacts statistical inferences, an issue that had not previously been systematically explored. The results lead to various implications for applied single-case researchers who are conducting intervention studies, as well as for the methodologists who seek precise methods for determining intervention effects when analyzing single-case research.

Implications for the Applied Single-Case Researchers

The findings from this study provide a few recommendations for researchers who conduct single-case studies. The results of this study confirm that single-case researchers should feel comfortable interpreting the overall average treatment effects (shift in level and shift in slope) when they have data that show no between case variation of data, and furthermore, that the overall average treatment effects can also be comfortably interpreted when there is some between case variation in the variance (evenly spread out up to a variance ratio of 16), regardless of whether the heterogeneity has been explicitly modeled. However, researchers should be cautious to interpret overall treatment effects from a model that assumes homogeneity when they have data that show one or more cases that have substantially different variability than other cases. In the real world, single-case data that show one of the cases have a substantially larger amount of variability compared with the other cases exist (e.g., Harris, Friedlander, Saddler, Frizzelle, & Graham, 2005). The results of this study indicate that if researchers had this kind of data, but they failed to correctly model or specify the level-1 error structure, then the results of the treatment effects would be inaccurate. Therefore, findings from this study suggest that researchers need to carefully inspect their data, and if they have data that show one or more cases that have a large amount of variability compared to the other cases, then they should try to model between case variation in the level-1 error structure to obtain more accurate and precise average treatment effects.

Generally, variance components were biased in multilevel modeling of single-case data analysis. The results from this study were consistent with this previous finding. However, this study suggests that accuracy and precision of the variance components can be improved by modeling between case variation in the level-1 error structure. Specifically, for researchers that have data regardless of showing or not showing between case variation, modeling between case variation can be beneficial to improve accuracy and precision of the estimates of the variation within cases and the autocorrelation. For researchers that have data that show one case that has a substantially larger amount of variability compared with the other cases, modeling between case variation can be beneficial to improve accuracy and precision of the estimates of all variance parameters, including variation in the treatment effects across cases, and variation within cases, and autocorrelation.

In addition, it was found that the design factors that continued to impact parameter estimates were the number of cases and the series length per case. As the number of cases and the series length per case increased, the accuracy and precision of the parameter estimates increased. This conclusion suggests that researchers should try to increase the number of participants or cases as well as the number of time points in their studies whenever possible.

Particularly, increasing the number of participants or cases can be more beneficial since the impact of the number of time points can be reduced if the number of participants or cases increases.

Lastly, this study also provides a way to model between case variation in the level-1 error structure using WinBUGS, and makes these created codes accessible to applied researchers for use in their own research (Appendix D).

Implications for Methodologists

This study provides a few implications for methodologists who use a multilevel modeling to conduct single-case data analyses. Since this study only used the simplest correlated level-1 error structure, AR(1), methodologists may want to look at more complex correlated level-1 error structures to investigate if the results from this study can be replicated with other error structures. Similarly, further research can be done using different types of outcomes, such as binary, ordinal, or count outcomes. This would be reasonable because many of the outcomes used in single-case research are not continuous outcomes.

In addition, more simulation work can be done with data having an extremely heterogeneous error structure. The results of Study 3 indicate that the different modeling methods in the level-1 error structure can have a substantial impact on both fixed effects and variance components when analyzing data having the extremely heterogeneous error structure. This finding is particularly distinguished from previous works that have investigated the misspecifications of the level-1 error structures on the single-case research and other research on the longitudinal analysis. The previous studies have found that the fixed effects are generally robust to misspecifications of the level-1 error structure (Ferron et al., 2009; Ferron et al., 2002;

Kwok et al, 2007; Merlande, 2014; Owen, 2011). However, Study 3 found that the misspecification of the level-1 error structure can have a substantial impact on both fixed effects and variance components. Therefore, these finding can be meaningful and beneficial for both researchers who are interested in average treatment effects as well as researchers who are interested in variation in the treatment, variation within cases, and autocorrelation, if it can be generalized to a broad range of the conditions. Because Study 3 only included a few conditions, further work should include more conditions that would allow for a thorough investigation of the impact of different models of the level-1 error structure on the estimates of multilevel models used with heterogeneous single-case data.

Additionally, this study can be expended to more general growth curve studies or meta-analysis studies using multilevel modeling. For those studies, it is possible that the level-1 error structure may vary across upper levels (e.g., classes or schools) or studies. Further work needs to be done to explore if the level-1 error structure varies across different studies or upper levels in real data sets, and if so, methodologist may want to examine if different methods of modeling level-1 error structure have some impact on the results of those studies.

Furthermore, further research should be done to find the alternative estimation approaches on estimating variance components. This study indicated that the variance components are generally biased, especially the level-2 error variance. Thus, it would be worth investigating if the observed bias in the variance components can be reduced by using different approaches, such as different choices of priors (e.g., the use of more informative priors) in the Bayesian framework.

Finally, this study focused on only the Bias, RMSE, the CI coverage, and the width outcomes of the parameter estimates. It would be interesting to investigate the impact of the different modeling in the level-1 error structure on Type I error rates and the power estimates of the treatment effects. Some previous research on the misspecification of the multilevel growth curve models had found that the under specified models showed within the nominal alpha level (.05) of Type I error rates but the low statistical power of the fixed effects (Kwok et al., 2007; Ferron et al., 2002). There are few single-case studies that looked at the Type I error rate and the power estimates. Previous work with the three-level models on the single-case data (Merlande, 2014) had found that the Type I error rates tended to be close to the nominal level which is consistent with the previous studies of the multilevel growth curve models. In addition, Merlande (2014) had found that the variability at the upper levels had substantial impact on the power estimates of the fixed effects. Since few studies were done on the Type I error rates and the power estimates in the multilevel modeling frame work on the single-case data analyses, it would be worthwhile to investigate these outcomes of the parameter estimates.

REFERENCES

Armitage, P., Berry, G., & Matthews, J. N. (2002). *Statistical Methods in Medical Research, 4th ed.* Oxford: Blackwell Science.

Baek, E., & Ferron, J. M. (2013). Multilevel Models for Multiple-Baseline Data: Modeling Across Participant Variation in Autocorrelation and Residual Variance. *Behavior Research Method. 45,* 65-74.

Baek, E., & Ferron, J. M. (2011, April). *Multilevel models for multiple-baseline data: Modeling Between-Case Variation in Autocorrelation.* Structured poster presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Baek, E.K., Moeyaert,M., Petit-Bois, M., Beretvas, S.N., Van den Noortgate, W., & Ferron, J.M.(2013). The use of multilevel analysis for intergrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation: An International Journal*, doi: 10.1080/09602011.2013.835740.

Baek, E., Petit-Bois, M., Van den Noortgate, W., Beretvas, T., & Ferron, J. (2014). Using visual analysis to evaluate and refine multilevel models of single-case studies. *Journal of Special Education*. Advance online publication. DOI: 10.1177/0022466914565367.

Baek, E., Petit-Bois, M., & Ferron, J. M. (2014, April). *A comparison of Bayesian restricted maximum likelihood approach in multilevel models for single-case data.* Structured poster presented at the Annual Meeting of the American Educational Research Association. Philadelphia, PA.

Baek, E., Petit-Bois, M., Ferron, J. M. (2012, April). *The effect of error structure specification on the meta-analysis of single-case studies of reading fluency interventions.* Structured poster presented at the Annual Meeting of the American Educational Research Association. Vancouver, BC, Canada.

Baek, E., Petit-Bois, M., Ferron, J. M. (2013, April). *A feasible way to vary level-1 error structure across participants in multilevel models for single-case data.* Structured poster presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single Case Experimental Designs: Strategies for Studying Behavior Change* (3rd ed.). Boston, MA: Pearson.

Baldwin, S.A. & Fellingham, G.W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods, 18,* 151-164.

Beretvas, S. N., & Chung, H. (2008). An evaluation of modified R2-change effect size indices for single-subject experimental designs. *Evidence-based Communication Assessment and Intervention, 2*, 120-128.

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 3,* 385-402.

Berger, J. O., & Strawderman, W. E. (1996). Choice of hierarchical priors: Admissibility in estimation of normal means. *Annals of Statistics, 24,* 931-951.

Brooks, S. P., & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7,* 434–455.

Brooks, S. P., & Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, *8,* 319-335.

Browne, W. (2008). *MCMC estimation in MLwinN*. Bristol, England: Centre for Multilevel Modeling.

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

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

Chen, M. H., & Shao, Q. M. (1999). Monte Carlo estimation of Bayesian credible and HPD Intervals. *Journal of Computational and Graphical Statistics*, *8,* 69-92.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Lawrence Erlbaum Associates.

Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association, 44,* 32–61.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91,* 883-904.

Daniels, M.J. & Kass, R.E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models, *Journal of the American Statistical Association, 94,* 1254–1263.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573-579.

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5,* 235–251.

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association 70,* 311-319.

Farmer, J. L., Owens, C.M., Ferron, J.M., & Allsopp, D.H. (2010, April). *A methodological review of single-case meta-analyses.* Paper presented at the American Educational Research Association. Denver, CO.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making Treatment Effect Inferences from Multiple-Baseline Data: The Utility of Multilevel Modeling Approaches. *Behavior Research Methods, 41*, 372-384.

Ferron, J. M., Dailey, R., & Yi. Q. (2002). Effects of misspecifying the first level error structure in two-level models of change. *Multivariate Behavioral Research, 37*, 379-403.

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

Ferron, J. & Jones, P. K. (2006). Test for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education, 75,* 66-81.

Ferron, J. M. & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *Journal of Experimental Education, 64*, 231-239.

Ferron, J. M. & Rendina-Gobioff, G. (2005). Interrupted time series design. In B. Everitt & D. Howell (Eds.), *Encyclopedia of Behavioral Statistics* (Vol. 2, pp. 941-945). West Sussex, UK: Wiley & Sons Ltd.

Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *Journal of Experimental Education, 63,* 167-178.

Fisch, G.S. (2001). Evaluating data from behavioral analysis: visual inspection or statistical models? *Behavior Processes, 54*, 137-154.

Fisher,W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36,* 387–406.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1997). *Design and Analysis of Single-case Research.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gamerman, D., & Lopes, H.(2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference, second edition.* Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A. (2002). Prior distribution. *Encyclopedia of Environmetrics*, *3,* 1634–1637.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1,* 515–533.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis (2nd Ed).* Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Gilks, W.R., Richardson, S., & Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in practice.* Boca Raton, FL: Chapman and Hall/CRC.

Greenwood, K. M., & Matyas, T. A. (1990). Problems with the application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12,* 355-370.

Goldstein, H. (1995). *Multilevel statistical models* (2nd Ed.). New York: Wiley.

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *3,* 403-420.

Goldstein, H., Healy, M., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, *13*, 1643-1655.

Harris, K.R., Friedlander, B.D., Saddler, B., Frizzelle, R., & Graham, S. (2005). Self-monitoring of attention versus self-monitoring of academic performance: Effects among students with ADHD in the general education classroom. *The Journal of Special Education, 39,* 145-156.

Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis, 30,* 313-326.

Heitjan, D. F. & Sharma, D. (1997). Modeling repeated-series longitudinal data. *Statistics in Medicine*, *16*, 347-355.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5,* 315-332.

Hox, J. (1998). Multilevel modeling: When and why? In I. Balderjahn, R. Mathar & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147-154). New York: Springer.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7,* 107-118.

Huitema, B.E., & McKean, J.W. (1991). Autocorrelation estimation and inference with small samples. *Pscyhological Bulletin, 110*, 291-304.

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. doi: 10.1177/00131640021970358.

Huitema, B. E., McKean, J. W., & McKnight, S.D. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement, 59,* 767-786.

Ittenbach, R. F., & Lawhead, W. F. (1997). Historical and philosophical foundations of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 13–39). Mahwah, NJ: Erlbaum.

Jeffreys, H. (1961). *Theory of Probability (3rd Ed)*. London, England: Oxford University Press.

Jennrich, R. I. & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, *42*, 805-820.

Johnston, J. (1984). *Economic methods* (3rd ed). New York: McGraw-Hill.

Jones, R.R., Weinrott, M.R., & Vaught, R.S. (1978). Effects of serial D=dependency on the agreement between visual analysis and statistical inference. *Journal of Applied Behavior Analysis, 11*, 277-283.

Kazdin, A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings* (2nd ed.). New York: Oxford University Press.

Kenward, M. G. & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983 – 997.

Kenward, M. G. & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis, 53*, 2583 – 2595.

Kratochwill, T. R. (1985). Case study research in school psychology. *School Psychology Review, 14,* 204–215.

Kratochwill, T., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arntson, P., McMurray, N., Hempstead, J., & Levin, J. (1974). A further consideration in the application of an analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 7*, 629-633.

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299-312.

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.

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,* 557-592.

Lindley, D. V., & Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological), 34,* 1-41.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*: New York, NY: Springer. doi:10.1007/978-0-387-71265-9.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30,* 598-617.

Maggin, D.M, Swaminathan, H., Rogers, H., O'Keeffee, B. Sugai, G., & Horner, R. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49,* 301-321.

Matyas, T.A. & Greenwood, K.M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Matyas, T.A., & Greenwood, K.M. (1996). Serial dependency in single-case time series. In R.D. Franklin, D.B. Allison, & B.S. Gorman, *Design and analysis of single-case research* (pp. 215-243). Mahwah, New Jersey: Lawrence Erlbaum Associates.

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods*, *5*, 87-101.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013a). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research*, *48*, 719-748.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2013b). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education, 82,* 1-21.

Morgan, D.L. & Morgan, R.K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist, 56,* 119 – 127.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association 78,* 47–65.

Murphy, D.L., & Pituch, K.A. (2009). The Performance of Multi-level Growth Curve Models Under an Auto regressive Moving Average Process. *The Journal of Experimental Education, 77(3),* 255-282.

Neter, J., Wassermann, W., & Kutner, M. H. (1990). *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs* (3rd ed.). Homewood, IL: Irwin.

Nugent, W. (1996). Integrating single-case and group comparison designs for evaluation research. *Journal of Applied Behavioral Science, 32*, 209-226.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical Journal of Pain, 21*, 56-68.

Owens, C. M. (2011). Meta-analysis of single-case data: A monte carlo investigation of a three level model. (Unpublished doctoral dissertation). University of South Florida: Tampa, FL.

Owens, C. M., & Ferron, J. M. (2012). Synthesizing single-case studies: A Monte Carlo examination of a three-level meta-analytic model. *Behavior Research Methods, 44,* 795-805.

Paris, S.J., & Winsten, C. B. (1954). *Trend estimators and serial correlation.* Unpublished discussion paper for the Cowles Commission, Chicago (Stat. No. 383). (Availble from the Department of Business, Univeristy of Chicago, Chicago, IL 60637)

Parker, R.I., Hagan-Burke, S., & Vannest, K. (2007). Percent of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40*, 194-204.

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

Parker, R. I., Vannest, K. J. , Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy,* 10.1016/j.beth.2010.08.006.

Parsonson, B. S.,& Baer, D. M. (1986). The graphic analysis of data. In A. Poling & W. R. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157-186). New York: Plenum Press.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometri1ka 58,* 545-554.

Petit-Bois, M. (2012, April). *Consequences of misspecification of growth trajectories when meta-analyzing single-case data using a three level model***.** Poster presented at the American Educational Research Association. VanCouver, CA.

Petit-Bois, M. (2014). Monte Carlo study: Consequences of the misspecification of the level-1 error structure when meta-analyzing single-case designs using a three-level model. (Unpublished doctoral dissertation). University of South Florida: Tampa, FL.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Shadish, W. R., Kyse, E. N., & Rindskopf (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18,* 385-405.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113*, 95-109.

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

Sidman, M. (1960). *Tactics of scientific research.* Boston: Authors Cooperative, Inc.

Solanas, A., Manolov, R., & Sierra, V. (2010). Lag-one autocorrelation in short series: Estimation and hypotheses testing. *Psicologica, 31,* 357-381.

Scheffe, H. (1959). *The Analysis of Variance.* New York: John Wiley & Sons.

Scruggs, T.E., Mastropieri, M.A., & Castro, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24-33.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113,* 95-109.

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

Shadish, W.R. & Sullivan, K.J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43, 971-980*. Doi: 10.3758/s13428-011-01111-y.

Simonton, D. K. (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin, 84*, 489-502.

Sivo, S. A., Fan, X., & Witta, L. (2005). The biasing effects of unmodeled ARMA time series processes on latent growth curve model estimates. *Structural Equation Modeling, 12,* 215–231.

Sivo, S. A., & Willson, V. L. (2000). Modeling causal error structures in longitudinal panel data: A Monte Carlo study. *Structural Equation Modeling, 7,* 174–205.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS User Manual 1.4.* Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22,* 1701–1762.

Toothaker, L. E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). N=1 designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 4*, 289-309.

Ugille, M., Moeyaert, M., Beretvas, T., Ferron, J., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Res.*doi: 10.3758/s13428-012-0213-1.

Van den Noortgate, W. & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18,* 325-346.

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

Van den Noortgate, W. & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today, 8*, 196-209.

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

Velicer, W. F., & Fava, J. L. (2003). Time series analysis. In J. A. Schinka & W. F. Velicer (Eds.), Handbook of psychology. Vol. 2: *Research methods in psychology* (pp. 581–606). Hoboken, NJ: Wiley.

Wacker, D. P., Steege, M., & Berg, W. K. (1988). Use of single-case designs to evaluate manipulable influences on school performance. *School Psychology Review, 17,* 651–657.

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inference. *Behavioral Assessment, 3*, 71-92.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, *39*, 95-101.

Watson, T. S., Meeks, C., Dufrene, B., & Lindsay, C. (2002). Sibling thumb sucking: Effects of treatment for targeted and untargeted siblings. *Behavior Modification, 26,* 412-423.

West, S. G., & Hepworth, J. T. (1991). Data analytic strategies for temporal data and daily events. *Journal of Personality, 59,* 60-9-662.

Whalen, C., Schreibman, L., &Ingersoll, B. (2006). The collateral effects of joint attention training on social initiations, positive affect, imitation, and spontaneous speech for young children with autism. *Journal of Autism and Developmental Disorders, 36,* 655-664.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation*, *22*, 1079-1106.

Yang, M. & Goldstein, H. (1996). Multilevel models for longitudinal data. In U. Engel & J. Reinecke (Eds.), *Analysis of change: Advanced techniques in panel data analysis* (pp. 191-220). New York: Walter de Gruyter.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods, 14,* 301–322.

APPENDIX A: TABLES OF RELATIVE BIAS VALUES

Table A1

Relative bias for the fixed treatment effects

Number of cases	Series length per case	Variation in the level-2 errors	Shift in level						Shift in slope					
			Homo		Moderately hetero		Severely hetero		Homo		Moderately hetero		Severely hetero	
			Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
4	10	Level-2	-0.004	-0.005	0.030	0.027	-0.033	-0.036	0.094	0.090	0.018	0.011	0.033	0.002
		Level-1	0.001	-0.002	-0.013	-0.012	0.014	0.014	-0.001	-0.001	0.063	0.055	-0.062	-0.076
	20	Level-2	0.012	0.009	-0.008	-0.007	0.000	-0.001	0.038	0.040	-0.018	-0.021	-0.027	-0.029
		Level-1	-0.011	-0.011	-0.005	-0.007	0.007	0.008	0.052	0.056	0.020	0.021	0.051	0.051
8	10	Level-2	0.004	0.003	0.020	0.017	0.010	0.007	0.122	0.118	-0.030	-0.024	0.012	0.009
		Level-1	0.007	0.008	-0.003	-0.002	0.001	0.003	0.040	0.043	0.006	0.004	-0.010	-0.002
	20	Level-2	0.007	0.007	-0.004	-0.005	-0.002	-0.004	0.017	0.021	0.004	0.003	0.027	0.026
		Level-1	0.005	0.005	-0.008	-0.010	-0.003	-0.001	-0.011	-0.012	-0.005	-0.004	-0.016	-0.011

Table A2

Relative bias for the variance components

Number of cases	Series length per case	Variation in the level-2 errors	Level-2 error SD for shift in level						Level-2 SD for shift in slope					
			Homo		Moderately hetero		Severely hetero		Homo		Moderately hetero		Severely hetero	
			Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
4	10	Level-2	1.134	1.129	1.175	1.159	1.257	1.235	1.610	1.604	1.548	1.545	1.678	1.618
		Level-1	2.104	2.099	2.019	2.012	2.122	2.049	2.983	2.975	2.860	2.860	2.885	2.795
	20	Level-2	1.045	1.052	1.056	1.062	1.111	1.104	0.941	0.927	0.987	0.974	1.001	0.983
		Level-1	1.743	1.759	1.654	1.670	1.657	1.636	1.240	1.248	1.227	1.210	1.229	1.199
8	10	Level-2	0.177	0.172	0.182	0.170	0.177	0.174	0.279	0.271	0.217	0.202	0.242	0.212
		Level-1	0.434	0.432	0.423	0.418	0.432	0.384	0.650	0.650	0.589	0.567	0.657	0.590
	20	Level-2	0.164	0.158	0.143	0.133	0.154	0.162	0.173	0.154	0.169	0.153	0.159	0.148
		Level-1	0.362	0.363	0.296	0.287	0.281	0.262	0.192	0.172	0.150	0.134	0.196	0.183

APPENDIX B: TABLES OF ETA-SQUARED VALUES

Table A3

Eta-squared values (η2) for the association of the design factors with the bias for the shift in level parameter

	99%
Variation in the level-2 errors*True level-1 error structure	0.24476
Series length per case*Variation in the level-2 errors*True level-1 error structure	0.20389
Number of cases*Variation in the level-2 errors*True level-1 error structure	0.14989
Series length per case*True level-1 error structure	0.11015
Series length per case*Number of cases*Variation in the level-2 errors*True level-1 error structure	0.10181
Series length per case*Number of cases*True level-1 error structure	0.04305
Number of cases	0.02686
Series length per case*Number of cases	0.02642
Series length per case	0.02012
Series length per case*Number of cases*Variation in the level-2 errors	0.02011
Number of cases*True level-1 error structure	0.01605
True level-1 error structure	0.01196
Number of cases*Variation in the level-2 errors	0.00999
Variation in the level-2 errors	0.00896
Variation in the level-2 errors*Type of model	0.00124
Variation in the level-2 errors*Type of model*True level-1 error structure	0.00087
Series length per case*Variation in the level-2 errors*Type of model*True level-1 error structure	0.00086
Type of model	0.00078
Number of cases*Type of model	0.00043
Series length per case*Variation in the level-2 errors*Type of model	0.00029
Number of cases*Type of model*True level-1 error structure	0.00025
Series length per case*Variation in the level-2 errors	0.00021
Series length per case*Type of model	0.00018
Series length per case*Number of cases*Type of model	0.00014
Number of cases*Variation in the level-2 errors*Type of model	0.00012
Series length per case*Number of cases*Variation in the level-2 errors*Type of model	0.00011
Type of model*True level-1 error structure	0.0001
Series length per case*Number of cases*Type of model*True level-1 error structure	0.00003
Number of cases*Variation in the level-2 errors*Type of model*True level-1 error structure	0.00003
Series length per case*Type of model*True level-1 error structure	0.00002
Total Explained	99%

Table A4

Eta-squared values (η2) for the association of the design factors with the bias for the shift in slope parameter

	η^2
True level-1 error structure	0.24789
Variation in the level-2 errors*True level-1 error structure	0.14349
Series length per case*Number of cases*True level-1 error structure	0.10482
Series length per case*True level-1 error structure	0.09062
Series length per case*Number of cases*Variation in the level-2 errors	0.08815
Series length per case*Variation in the level-2 errors*True level-1 error structure	0.08258
Series length per case*Variation in the level-2 errors	0.06979
Number of cases*True level-1 error structure	0.03448
Variation in the level-2 errors	0.02223
Number of cases*Variation in the level-2 errors	0.0209
Series length per case*Number of cases	0.01767
Series length per case	0.01627
Number of cases	0.00464
Number of cases*Variation in the level-2 errors*True level-1 error structure	0.00158
Number of cases*Type of model	0.00144
Number of cases*Type of model*True level-1 error structure	0.00139
Series length per case*Type of model	0.00119
Series length per case*Number of cases*Type of model	0.00118
Type of model*True level-1 error structure	0.00079
Variation in the level-2 errors*Type of model*True level-1 error structure	0.0007
Type of model	0.00059
Variation in the level-2 errors*Type of model	0.0005
Series length per case*Type of model*True level-1 error structure	0.00047
Series length per case*Variation in the level-2 errors*Type of model	0.00011
Number of cases*Variation in the level-2 errors*Type of model	0.00004
Total Explained	95%

Table A5

Eta-squared values (η2) for the association of the design factors with the RMSE for the shift in level parameter

	η^2
Number of cases	0.48343
Variation in the level-2 errors	0.3801
Series length per case	0.10668
True level-1 error structure	0.00113
Type of model	0.00025
Total Explained	97%

Table A6

Eta-squared values (η2) for the association of the design factors with the RMSE for the shift in slope parameter

	η^2
Series length per case	0.47322
Number of cases	0.27695
Variation in the level-2 errors	0.22127
True level-1 error structure	0.00143
Type of model	0.00013
Total Explained	97%

Table A7

Eta-squared values (η2) for the association of the design factors with the CI coverage for the shift in level parameter

	η^2
Number of cases	0.88425
Variation in the level-2 errors	0.03597
Series length per case	0.01128
True level-1 error structure	0.00842
Type of model	0.00127
Total Explained	94%

Table A8

Eta-squared values (η2) for the association of the design factors with the CI coverage for the shift in slope parameter

	η^2
Number of cases	0.8304
Series length per case	0.08105
Variation in the level-2 errors	0.0166
Series length per case*Number of cases	0.01376
Series length per case*Variation in the level-2 errors	0.00771
Series length per case*True level-1 error structure	0.00529
Variation in the level-2 errors*True level-1 error structure	0.00168
Number of cases*Variation in the level-2 errors	0.00064
True level-1 error structure	0.00051
Series length per case*Type of model	0.00032
Type of model	0.00032
Type of model*True level-1 error structure	0.00019
Number of cases*True level-1 error structure	0.00011
Variation in the level-2 errors*Type of model	0.00006
Number of cases*Type of model	0.00002
Total Explained	96%

Table A9

Eta-squared values (η2) for the association of the design factors with the CI width for the shift in level parameter

shift in level	η^2
Number of cases	0.84039
Variation in the level-2 errors	0.11522
Series length per case	0.01657
True level-1 error structure	0.0002
Type of model	0
Total Explained	97%

Table A10

Eta-squared values (η2) for the association of the design factors with the CI width for the shift in slope parameter

	η^2
Number of cases	0.65402
Series length per case	0.19115
Variation in the level-2 errors	0.0952
True level-1 error structure	0.00046
Type of model	0.00003
Total Explained	94%

Table A11

Eta-squared values (η2) for the association of the design factors with the bias of the level-2 error standard deviation for the shift in level parameter

	η^2
Number of cases	0.95699
Series length per case	0.01168
Variation in the level-2 errors	0.00933
True level-1 error structure	0.00035
Type of model	0.00005
Total Explained	98%

Table A12

Eta-squared values (η2) for the association of the design factors with the bias of the level-2 error standard deviation for the shift in slope parameter

	η^2
Number of cases	0.74042
Series length per case	0.15689
Series length per case*Number of cases	0.06818
Variation in the level-2 errors	0.01321
Number of cases*Variation in the level-2 errors	0.01295
Series length per case*Variation in the level-2 errors	0.00605
Series length per case*True level-1 error structure	0.00046
True level-1 error structure	0.0004
Type of model	0.00017
Variation in the level-2 errors*True level-1 error structure	0.00016
Type of model*True level-1 error structure	0.00007
Number of cases*True level-1 error structure	0.00006
Series length per case*Type of model	0.00001
Variation in the level-2 errors*Type of model	0.00001
Number of cases*Type of model	0
Total Explained	99%

Table A13

Eta-squared values (η2) for the association of the design factors with the bias of the level-1 error standard deviation parameter

	99%
Series length per case	0.25041
True level-1 error structure	0.22317
Variation in the level-2 errors	0.19479
Number of cases	0.11453
Type of model*True level-1 error structure	0.10383
Series length per case*Variation in the level-2 errors	0.04285
Number of cases*Type of model	0.01779
Series length per case*Number of cases	0.01247
Series length per case*Type of model	0.00873
Series length per case*True level-1 error structure	0.00854
Number of cases*True level-1 error structure	0.00729
Type of model	0.00456
Number of cases*Variation in the level-2 errors	0.00016
Variation in the level-2 errors*True level-1 error structure	0.0001
Variation in the level-2 errors*Type of model	0.00008
Total Explained	99%

Table A14

Eta-squared values (η2) for the association of the design factors with the bias of the autocorrelation parameter

	η^2
True level-1 error structure	0.88205
Variation in the level-2 errors	0.04965
Series length per case	0.02232
Type of model	0.0043
Number of cases	0.00322
Total Explained	96%

Table A15

Eta-squared values (η2) for the association of the design factors with the RMSE of the level-2 error standard deviation for the shift in level parameter

	η^2
Number of cases	0.8857
Variation in the level-2 errors	0.08459
Series length per case	0.00822
True level-1 error structure	0.00029
Type of model	0.00015
Total Explained	98%

Table A16

Eta-squared values (η2) for the association of the design factors with the RMSE of the level-2 error standard deviation for the shift in slope parameter

	η^2
Number of cases	0.73172
Series length per case	0.13375
Variation in the level-2 errors	0.05757
Series length per case*Number of cases	0.05416
Number of cases*Variation in the level-2 errors	0.01477
Series length per case*Variation in the level-2 errors	0.00456
Type of model	0.00026
Number of cases*True level-1 error structure	0.00023
True level-1 error structure	0.00023
Series length per case*True level-1 error structure	0.00018
Type of model*True level-1 error structure	0.00011
Variation in the level-2 errors*True level-1 error structure	0.00009
Series length per case*Type of model	0.00003
Number of cases*Type of model	0.00001
Variation in the level-2 errors*Type of model	0
Total Explained	>99%

Table A17

Eta-squared values (η2) for the association of the design factors with the RMSE of the level-1 error standard deviation parameter

	η^2
True level-1 error structure	0.62073
Type of model*True level-1 error structure	0.16449
Series length per case	0.10855
Type of model	0.05149
Number of cases	0.02176
Series length per case*Type of model	0.00789
Number of cases*True level-1 error structure	0.00775
Variation in the level-2 errors	0.0043
Series length per case*Number of cases	0.00315
Number of cases*Type of model	0.00207
Series length per case*Variation in the level-2 errors	0.00181
Series length per case*True level-1 error structure	0.00046
Variation in the level-2 errors*True level-1 error structure	0.00013
Variation in the level-2 errors*Type of model	0.00005
Number of cases*Variation in the level-2 errors	0.00001
Total Explained	99%

Table A18

Eta-squared values (η2) for the association of the design factors with the RMSE of the autocorrelation parameter

	η^2
Series length per case	0.06078
Number of cases	0.04635
Type of model	0.00194
Variation in the level-2 errors	0.00032
Total Explained	94%

Table A19

Eta-squared values (η2) for the association of the design factors with the CI coverage of the level-2 error standard deviation for the shift in level parameter

	η^2
Variation in the level-2 errors	0.29178
Series length per case	0.20843
Series length per case*Number of cases*True level-1 error structure	0.10725
Type of model	0.06713
Series length per case*Variation in the level-2 errors	0.05435
Series length per case*Number of cases*Variation in the level-2 errors*True level-1 error structure	0.04182
Series length per case*True level-1 error structure	0.04138
Series length per case*Variation in the level-2 errors*True level-1 error structure	0.03496
True level-1 error structure	0.0204
Number of cases*Variation in the level-2 errors	0.01805
Number of cases*Variation in the level-2 errors*True level-1 error structure	0.01632
Number of cases	0.0156
Series length per case*Number of cases*Type of model*True level-1 error structure	0.01346
Variation in the level-2 errors*True level-1 error structure	0.01297
Number of cases*True level-1 error structure	0.00824
Number of cases*Type of model	0.00806
Type of model*True level-1 error structure	0.00682
Series length per case*Number of cases*Variation in the level-2 errors	0.0055
Number of cases*Variation in the level-2 errors*Type of model	0.00373
Series length per case*Type of model*True level-1 error structure	0.00367
Number of cases*Variation in the level-2 errors*Type of model*True level-1 error structure	0.00227
Number of cases*Type of model*True level-1 error structure	0.00169
Series length per case*Type of model	0.00136
Series length per case*Number of cases*Type of model	0.00094
Variation in the level-2 errors*Type of model	0.00061
Series length per case*Variation in the level-2 errors*Type of model*True level-1 error structure	0.00027
Variation in the level-2 errors*Type of model*True level-1 error structure	0.00017
Series length per case*Number of cases*Variation in the level-2 errors*Type of model	0.00008
Series length per case*Number of cases	0.00003
Series length per case*Variation in the level-2 errors*Type of model	0.00003
Total Explained	99%

Table A20

Eta-squared values (η2) for the association of the design factors with the CI coverage of the level-2 error standard deviation for the shift in slope parameter

	η^2
Series length per case	0.64808
Variation in the level-2 errors*True level-1 error structure	0.06562
Type of model	0.06035
Variation in the level-2 errors	0.05248
Series length per case*Variation in the level-2 errors	0.04775
True level-1 error structure	0.04525
Number of cases*True level-1 error structure	0.01203
Series length per case*True level-1 error structure	0.00568
Number of cases*Variation in the level-2 errors	0.00503
Number of cases	0.00384
Series length per case*Type of model	0.00283
Type of model*True level-1 error structure	0.00199
Series length per case*Number of cases	0.00057
Variation in the level-2 errors*Type of model	0.0001
Number of cases*Type of model	0
Total Explained	95%

Table A21

Eta-squared values (η2) for the association of the design factors with the CI coverage of the level-1 error standard deviation parameter

	η^2
True level-1 error structure	0.32557
Type of model	0.30486
Type of model*True level-1 error structure	0.18597
Series length per case	0.06123
Series length per case*Type of model	0.04574
Series length per case*True level-1 error structure	0.03272
Number of cases*True level-1 error structure	0.00187
Number of cases	0.00077
Variation in the level-2 errors*True level-1 error structure	0.0004
Number of cases*Type of model	0.00018
Number of cases*Variation in the level-2 errors	0.00014
Series length per case*Variation in the level-2 errors	0.00013
Series length per case*Number of cases	0.00011
Variation in the level-2 errors*Type of model	0.00011
Variation in the level-2 errors	0
Total Explained	96%

Table A22

Eta-squared values (η2) for the association of the design factors with the CI coverage of the autocorrelation parameter

	η^2
True level-1 error structure	0.29609
Type of model	0.22422
Series length per case	0.17723
Type of model*True level-1 error structure	0.07331
Series length per case*True level-1 error structure	0.07218
Number of cases	0.05228
Series length per case*Type of model	0.02828
Number of cases*True level-1 error structure	0.02098
Number of cases*Type of model	0.01599
Variation in the level-2 errors	0.00922
Variation in the level-2 errors*True level-1 error structure	0.00419
Series length per case*Number of cases	0.00185
Variation in the level-2 errors*Type of model	0.0015
Number of cases*Variation in the level-2 errors	0.00053
Series length per case*Variation in the level-2 errors	0.00003
Total Explained	98%

Table A23

Eta-squared values (η2) for the association of the design factors with the CI width of the level-2 error standard deviation for the shift in level parameter

	η^2
Number of cases	0.88199
Variation in the level-2 errors	0.08128
Series length per case	0.00941
True level-1 error structure	0.00015
Type of model	0
Total Explained	97%

Table A24

Eta-squared values (η2) for the association of the design factors with the CI width of the level-2 error standard deviation for the shift in slope parameter

	η^2
Number of cases	0.698
Variation in the level-2 errors	0.09949
Series length per case	0.06895
Type of model	0.04344
Number of cases*Variation in the level-2 errors	0.03027
Series length per case*Number of cases	0.02176
Series length per case*Type of model	0.01841
Number of cases*Type of model	0.01166
Variation in the level-2 errors*Type of model	0.00052
True level-1 error structure	0.00017
Series length per case*Variation in the level-2 errors	0.00008
Type of model*True level-1 error structure	0.00006
Series length per case*True level-1 error structure	0.00004
Variation in the level-2 errors*True level-1 error structure	0.00002
Number of cases*True level-1 error structure	0.00001
Total Explained	99%

Table A25

Eta-squared values (η2) for the association of the design factors with the CI width of the level-1 error standard deviation parameter

	η^2
Type of model	0.43862
Series length per case	0.38414
Number of cases	0.11094
True level-1 error structure	0.0172
Variation in the level-2 errors	0.00949
Total Explained	96%

Table A26

Eta-squared values (η2) for the association of the design factors with the CI width of the autocorrelation parameter

	η^2
Series length per case	0.47031
Type of model	0.33877
Number of cases	0.16732
Variation in the level-2 errors	0.00931
True level-1 error structure	0.00028
Total Explained	99%

APPENDIX C: TABLES AND FIGURES OF INDIVIDUAL ESTIMATES OF OUTCOME VALUES

Table A27

Individual bias, RMSE, CI coverage and width for the fixed treatment effects

	Shift in level				Shift in slope			
	Model 1		Model 2		Model1		Model2	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Bias	-0.001	0.019	-0.002	0.019	0.003	0.008	0.002	0.007
RMSE	0.887	0.149	0.874	0.147	0.281	0.108	0.277	0.105
CI coverage	0.953	0.009	0.958	0.009	0.958	0.013	0.964	0.012
CI width	3.636	0.613	3.654	0.632	1.237	0.549	1.249	0.537

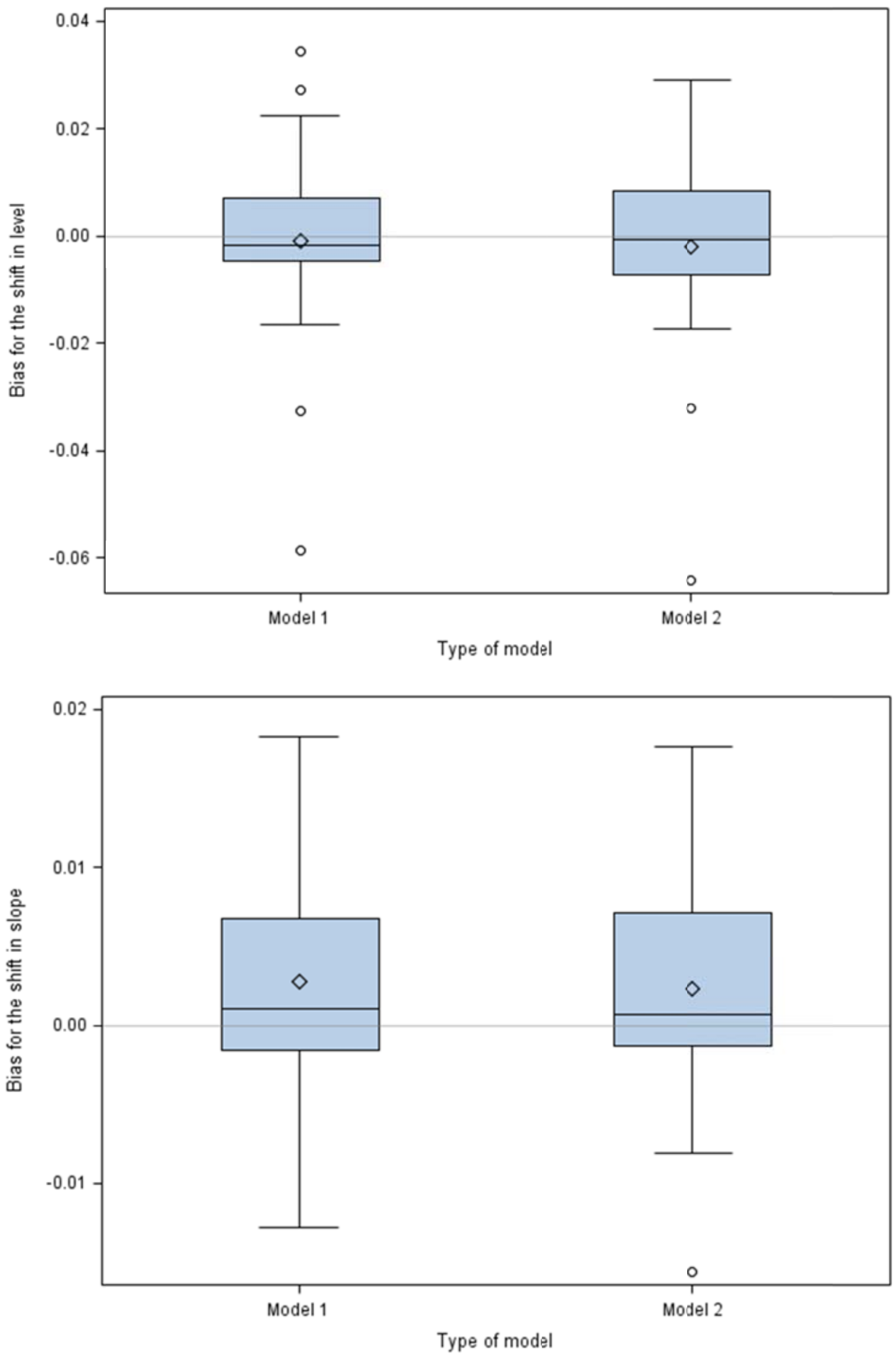


Figure A1. Box plots illustrating the distribution of the bias for the shift in level and the shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

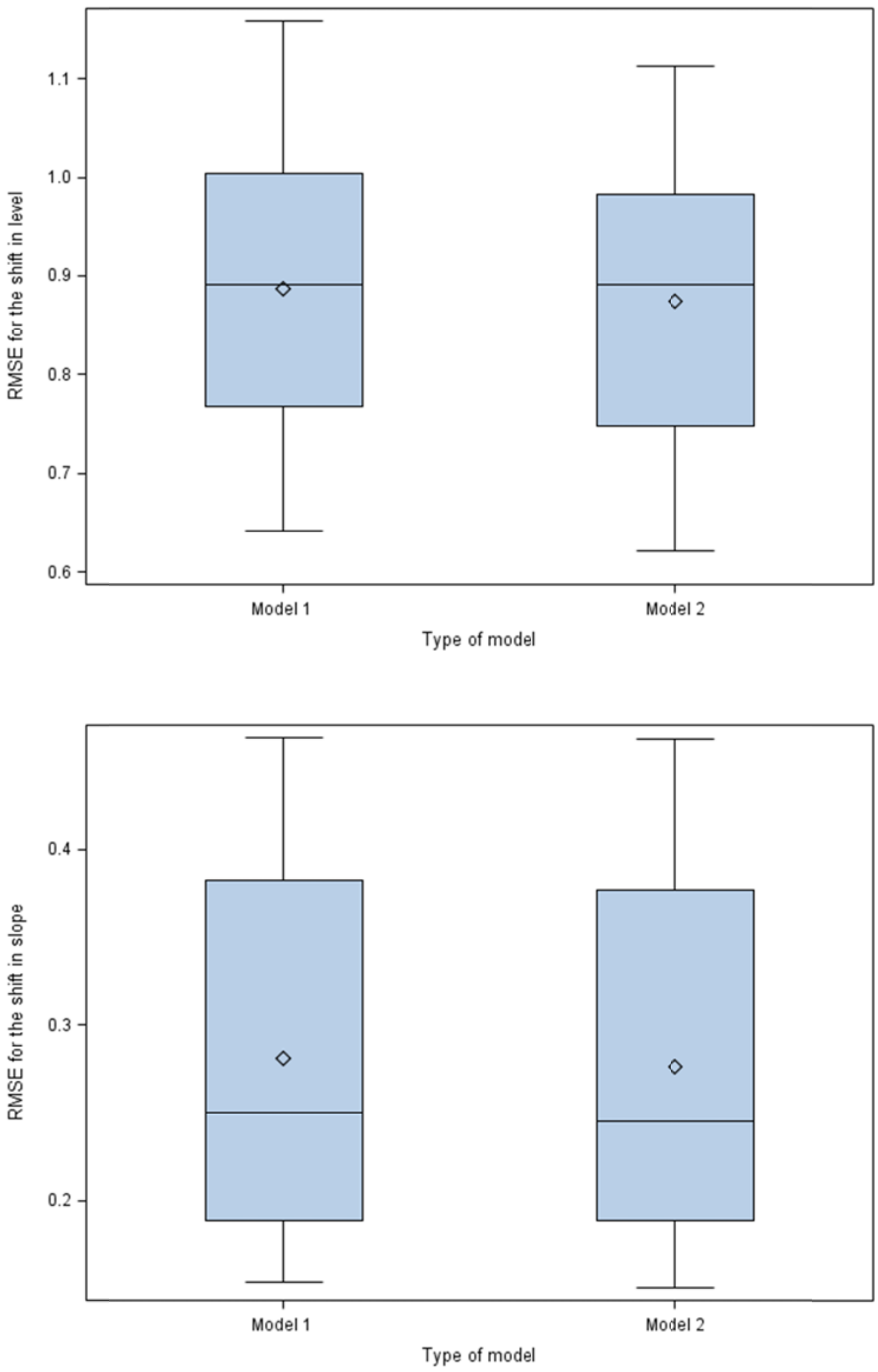


Figure A2. Box plots illustrating the distribution of the RMSE for the shift in level and the shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

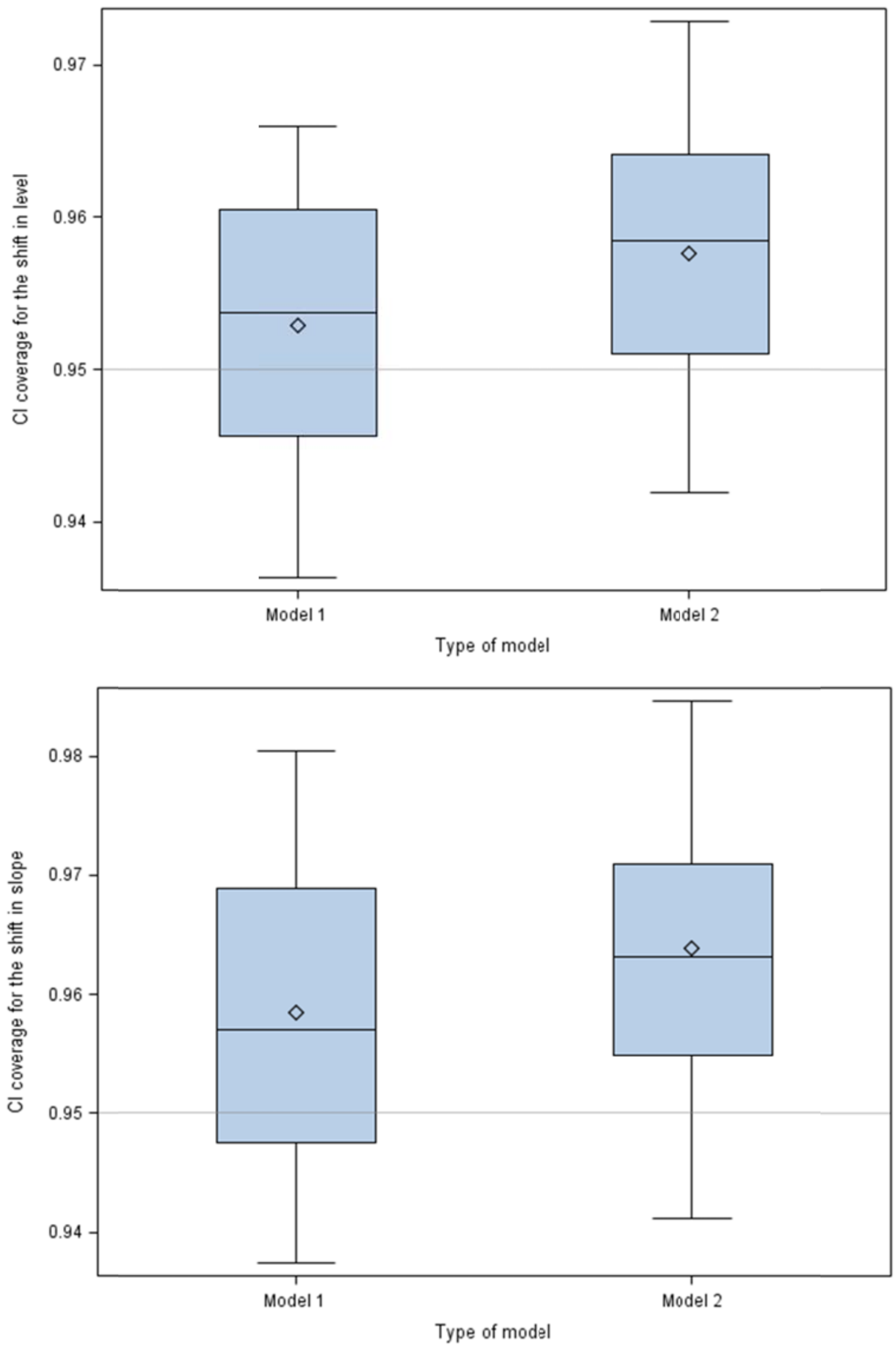


Figure A3. Box plots illustrating the distribution of the CI coverage for the shift in level and the shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

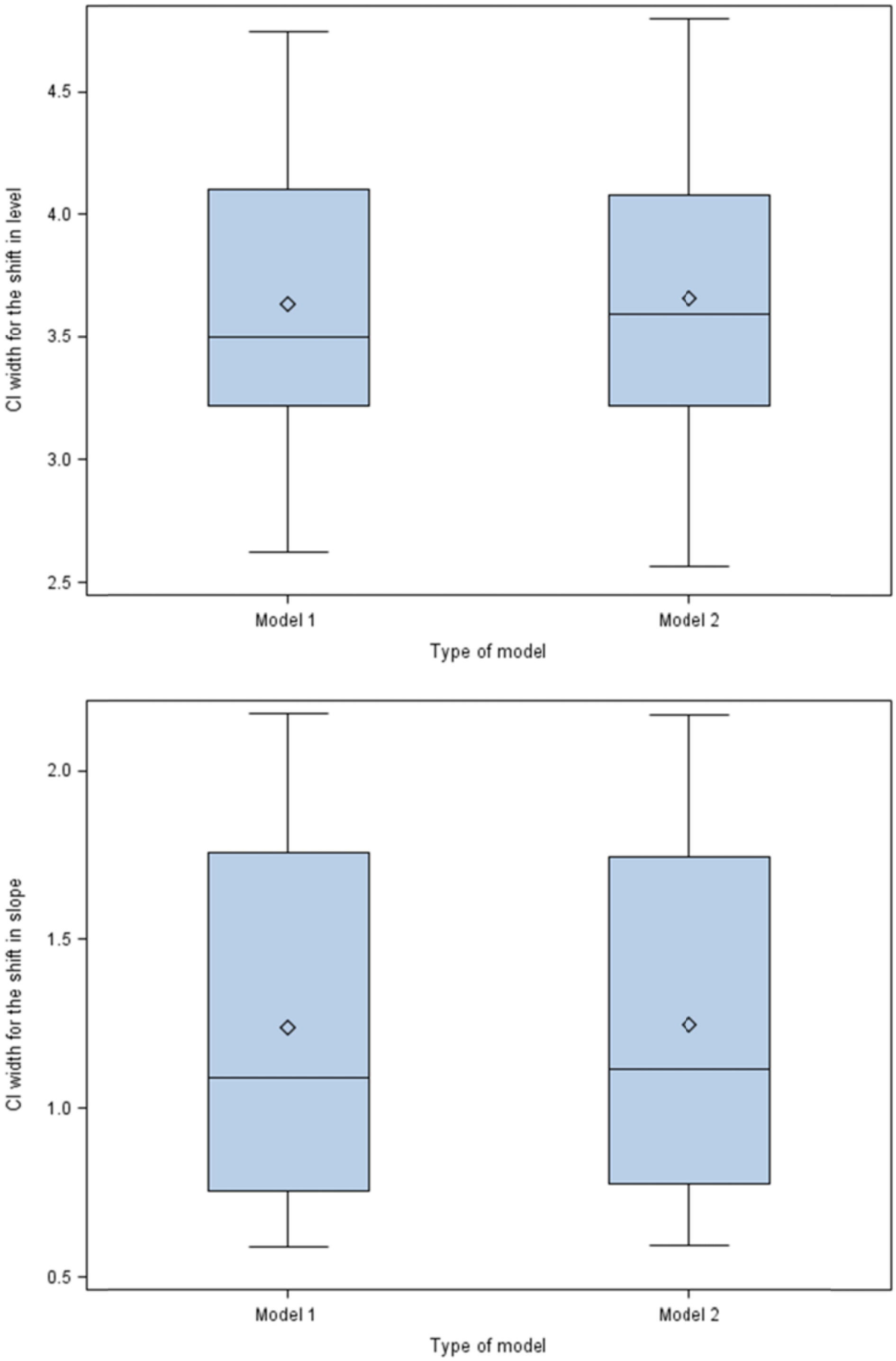


Figure A4. Box plots illustrating the distribution of the CI width for the shift in level and the shift in slope across Model 1 which did not model between case variation, and Model 2 which models between case variation.

APPENDIX D: WinBUGS codes for Model 1 and Model 2

Model 1:

```
Model
{
for( i in 1 : N ) {
for( j in 1 : T ) {
Y[i , j] ~ dnorm(theta[i ,j],tauc)
mu[i , j] <- alpha[i] + beta[i]*step(x[j]-CP[i])+ ca[i] * (x[j]) +
da[i]*(x[j] - CP[i])*step(x[j]-CP[i])
}
theta [i,1]<- mu [i,1]
for ( j in 2 : T) {
theta[i ,j]<-mu[i ,j]+tgamma*(Y[i ,j-1]-mu[i ,j-1])
}
alpha[i] ~ dnorm(alphac,alphatau)
beta[i] ~ dnorm(betac,betatau)
ca[i] ~ dnorm(cac,catau)
da[i] ~ dnorm(dac,datau)
}
alphac ~ dnorm(0.0,1.0E-6)
betac ~ dnorm(0.0,1.0E-6)
cac ~ dnorm(0.0,1.0E-6)
dac ~ dnorm(0.0,1.0E-6)
sigmaalpha~ dunif(0,100)
sigmabeta~ dunif(0,100)
sigmaca~ dunif(0,100)
sigmada~ dunif(0,100)
alphatau<-1/(sigmaalpha*sigmaalpha)
betatau<-1/(sigmabeta*sigmabeta)
catau<-1/(sigmaca*sigmaca)
datau<-1/(sigmada*sigmada)
tgamma~dnorm(0.0,1.0E-6)I(-0.99999,0.99999)
tsigma~ dunif(0,100)
tauc<- 1 / (tsigma*tsigma)
}
```

Model 2:

```
Model
{
for( i in 1 : N ) {
for( j in 1 : T ) {
Y[i , j] ~ dnorm(theta[i ,j],tauc[i])
mu[i , j] <- alpha[i] + beta[i]*step(x[j]-CP[i])+ ca[i] * (x[j]) +
da[i]*(x[j] - CP[i])*step(x[j]-CP[i])
}
theta [i,1]<- mu [i,1]
for ( j in 2 : T) {
theta[i ,j]<-mu[i ,j]+tgamma[i]*(Y[i ,j-1]-mu[i ,j-1])
}
alpha[i] ~ dnorm(alphac,alphatau)
beta[i] ~ dnorm(betac,betatau)
ca[i] ~ dnorm(cac,catau)
da[i] ~ dnorm(dac,datau)
tgamma[i]~dnorm(simge,gr)I(-0.99999,0.99999)
tsigma[i] ~ dunif(sa,sb)
tauc[i] <- 1 / (tsigma[i]*tsigma[i])
}
alphac ~ dnorm(0.0,1.0E-6)
betac ~ dnorm(0.0,1.0E-6)
cac ~ dnorm(0.0,1.0E-6)
dac ~ dnorm(0.0,1.0E-6)
sigmaalpha~ dunif(0,100)
sigmabeta~ dunif(0,100)
sigmaca~ dunif(0,100)
sigmada~ dunif(0,100)
alphatau<-pow(sigmaalpha, -2)
betatau<-pow(sigmabeta, -2)
catau<-pow(sigmaca, -2)
datau<-pow(sigmada, -2)
simge~dnorm(0.0,1.0E-6)
simgr~ dunif(0,100)
gr <- pow(simgr, -2)
sa~ dunif(0,100)
sb~ dunif(sa,100)
tmsig<-mean(tsigma[])
tmgamma<-mean(tgamma[])
smsig<- (sa+sb)/2
svsig<- sqrt((pow((sb-sa), 2))/12)
}
```