Performance of Contextual Multilevel Models for Comparing Between-Person and Within-Person Effects

by

Ingrid Carlson Wurpts

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy.

Approved May 2016 by the
Graduate Supervisory Committee

David P. MacKinnon, Chair
Stephen G. West
Kevin J. Grimm
Hye Won Suk

ARIZONA STATE UNIVERSITY

August 2016

ABSTRACT

The comparison of between- versus within-person relations addresses a central issue in psychological research regarding whether group-level relations among variables generalize to individual group members. Between- and within-person effects may differ in magnitude as well as direction, and contextual multilevel models can accommodate this difference. Contextual multilevel models have been explicated mostly for cross-sectional data, but they can also be applied to longitudinal data where level-1 effects represent within-person relations and level-2 effects represent between-person relations. With longitudinal data, estimating the contextual effect allows direct evaluation of whether between-person and within-person effects differ. Furthermore, these models, unlike single-level models, permit individual differences by allowing within-person slopes to vary across individuals. This study examined the statistical performance of the contextual model with a random slope for longitudinal within-person fluctuation data.

A Monte Carlo simulation was used to generate data based on the contextual multilevel model, where sample size, effect size, and intraclass correlation (ICC) of the predictor variable were varied. The effects of simulation factors on parameter bias, parameter variability, and standard error accuracy were assessed. Parameter estimates were in general unbiased. Power to detect the slope variance and contextual effect was over 80% for most conditions, except some of the smaller sample size conditions. Type I error rates for the contextual effect were also high for some of the smaller sample size conditions. Conclusions and future directions are discussed.

ACKNOWLEDGMENTS

I would like to thank my committee—Dr. Stephen West, Dr. Kevin Grimm, and Dr. Hye Won Suk—for their insightful feedback on this project. Particularly, I thank my graduate mentor Dr. David MacKinnon for his encouragement, advice, and sense of humor during my time in the Research in Prevention Lab. Thanks to my graduate student friends and colleagues who have shared with me their books, code, coffee, frozen meals, and sympathetic ears. Deep gratitude to my parents Ken and Carolyn who are the best cheerleaders for any long-distance endeavor. Finally, I could not have finished this dissertation without the love and support of my husband Christopher, who has been the happy bookend before and after every long day of graduate school.

This dissertation is derived in part from an article published in Multivariate Behavioral Research December 30, 2015, copyright Taylor & Francis, available online: http://www.tandfonline.com/10.1080/00273171.2015.1120176

TABLE OF CONTENTS

Page

CHAPTER Page

APPENDIX

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Longitudinal data are crucial for developing and testing theories in psychology. Not only do longitudinal data allow for establishing temporal precedence (West & Hepworth, 1991), increasing power (Muthén & Curran, 1997), and reduction of alternative explanations of cross-sectional effects (MacKinnon, 2008), but also they allow for simultaneous examination of both within-person change and between-person differences. Curran and Bauer (2011) argued that although there has been a recent emphasis on collecting and analyzing longitudinal data, models that separate between- and within-person effects in longitudinal data have not been used to their full potential in psychology research. This is unfortunate, because such models allow development and testing of interesting and complex psychological processes that may not be equal across different levels of analysis.

For example, Tennen, Affleck, Armeli, and Carney (2000) described how longitudinal data may contain differing between- and within-person effects. They collected data from 93 moderate- to heavy-drinking men and women for 60 consecutive nights. When examining the between-person data, they found that higher average daily drinking was associated with lower average daily emotion-focused coping (an adaptive coping strategy). However, when they estimated this relation within persons, they found that participants consumed more alcohol on days when they also used emotion-focused coping strategies. In a separate study, they found that daily alcohol consumption was related to a reduction in nervousness. They concluded that the within-person relation of drinking resulting in more positive coping strategies and decreased anxiety had a reinforcing effect such that people with higher anxiety were then more likely to drink more alcohol compared to those with lower anxiety. Importantly, although the findings at between and within levels contradict each other, they are still both valid, and both of these findings have implications for public health and the development of interventions. Furthermore, it would be erroneous to make generalizations from one level to the other, such as claiming that the more anxiety a person has, the less likely he or she is to drink. In an example such as this, there is a complex relation between alcohol and anxiety. Simply examining the relation at one level and ignoring the other would be incomplete, yet many studies in psychology focus on either within-person relations or between-person relations, to the exclusion of the other.

Because many longitudinal studies in psychology have not properly separated between- and within-person effects, it is difficult to know exactly how often between- and within-person effects diverge in real data. However, methodologists who study longitudinal data structures concur that between-person and within-person effects often differ (Hoffman, 2015; Snijders & Bosker, 2012; Bolger & Schilling, 1991). Furthermore, Molenaar (2004; 2008) has mathematically demonstrated the stringent conditions under which between-person and within-person relations will be equal, and he has claimed that many psychological processes inherently will not satisfy these stringent conditions. Therefore, if many psychological processes have differing relations at the between- and within-person levels, there is a need for methods that can accurately uncover these differences.

This paper discusses how and why relations found at the between-group and within-group levels often differ, particularly for longitudinal data where the grouping level is persons. First, the history of between- versus within-group effects is discussed. Then these issues are extended to longitudinal data where relations can differ between-person and within-person. Contextual multilevel models are described as way to properly analyze longitudinal data that include both between- and within-person effects. An extension of the contextual multilevel model that allows for individual differences in the relation between predictor and outcome variable (i.e., a random slope) is described. Next, daily diary methods are discussed, as they are an area to which contextual multilevel models are particularly well suited. This is followed by a discussion of why person-mean centering is an intuitive method of centering in contextual multilevel models and reasons why between- and within-person relations may differ. Previous simulation studies on contextual multilevel models are then reviewed. Based on this theoretical foundation, Chapters 2 and 3 describe an empirical simulation study to assess the performance of contextual multilevel models with random slopes for longitudinal data that fluctuates over time. Chapter 4 concludes with a discussion of the simulation study results.

Between- versus Within-Group Effects

Although this paper focuses on within-person and between-person effects that occur in longitudinal psychology data, other disciplines have been grappling with analogous issues for decades. Longitudinal data are one specific type of nested (or multilevel) data. The individual scores in multilevel data can be grouped in some way, and scores within a group are generally not independent from one another. That is, in multilevel data, individuals within a group are likely to have more similar scores than compared with individuals from different groups. For example, if the academic achievement of students in a school is measured, it is likely that the students' classroom has an effect on achievement. Students in a class with a good teacher might have higher achievement scores, whereas students in a class that is very rowdy might have lower achievement scores, even if both classes have students with all levels of academic ability. In this example, students are nested within classes, and students' academic achievement will be affected by their class membership. In the case of longitudinal data, the person is the grouping variable, and the repeated measures across time are nested within persons. Multilevel data violate the independence assumption of ordinary least squares (OLS) regression and should be analyzed with a model that accounts for the dependencies among the data.

The challenges of multilevel data have long been discussed in fields such as education (e.g., students grouped within classrooms), epidemiology (e.g., persons grouped within neighborhoods), and sociology (e.g., persons grouped within countries). When collecting and analyzing multilevel data, it is important to consider what causes and effects exist at each level of analysis. These issues surrounding levels of analysis occur with any nested data structure, including repeated measurements nested within persons and persons nested within groups. As methods to test between- and within-person relations with longitudinal psychology data have not yet been widely developed, it is useful to consider the work that has been done to address these issues in other disciplines, beginning in the 1950s.

Robinson (1950) was the first to demonstrate how "ecological correlations," or between-group correlations, cannot be used as substitutes for "individual correlations," or within-group correlations. This incorrect use of ecological correlations as substitutes for individual correlations was later described as the ecological fallacy (Selvin, 1958).

Likewise, an atomistic fallacy occurs when individual behavior is used to draw false conclusions about population behavior (Riley, 1963). More broadly, the ecological fallacy has been described as any fallacy that occurs when incorrectly applying inferences across any levels of aggregation—not just individuals (Firebaugh, 2001). For example, the individual unit could represent classrooms or schools. Susser (1994) suggested that we should examine links between all possible levels of analysis, as infinite levels of organization exist both between and within individuals.

Robinson defined individual correlations as those where the statistical object is indivisible, and he defined ecological correlations as those where the statistical object is a group of persons or objects. When his article was published, ecological correlations were prevalent in influential studies. Although the studies used ecological correlations, this was not necessarily because the authors were interested in relations at the ecological (or group) level. Rather, many researchers at that time used ecological studies to discover information about individuals. Robinson (1950) claimed, "Ecological correlations are used [in current literature] simply because correlations between the properties of individuals are not available" (p. 337). Robinson sought to show, mathematically, the discrepant relations between individual and ecological correlations.

To illustrate how ecological studies could incorrectly apply ecological results to the individual level, Robinson (1950) examined the relation of race/nativity and illiteracy. Using data from the 1930 U.S. Census, he computed the individual-level correlation between illiteracy (illiterate versus literate) and ethnicity (black versus white) as 0.203. He then computed ethnicity at the state level as the percentage of the state population 10 years and older which was black, and illiteracy as the percentage of the state population

10 years and older which was illiterate. When considered at the state level, the correlation between ethnicity and illiteracy was 0.773. Then he computed ethnicity and illiteracy as proportions at the level of nine geographic divisions, according to the U.S. Census Bureau. The correlation between illiteracy and ethnicity at this level was 0.946. When calculating the correlation between illiteracy and nativity (native-born or foreign born), the correlation went from 0.118 at the individual level to -0.526 at the state level. However, Robinson (1950) went a step beyond simply showing this inequality with sample data and claimed that ecological correlations cannot be used as substitutes for individual correlations, as the conditions under which the two correlations will be equal are far removed from reality. Furthermore, he implied that his paper was successful if it prevented future researchers from calculating "meaningless" ecological correlations (p. 341).

Alker (1969) used Robinson's (1950) basic statistical tools to look for any other potential inferential fallacies, beyond the ecological fallacy. Given *N* units, *R* groups, and *T* times, one can define averages across units, groups, or times. Similar relations (correlations or covariances) can occur at different levels of analysis, but inference from one level to another is more complicated. Incorrectly inferring from lower to higher levels can be called the aggregation fallacy (Alker, 1969). The individualistic fallacy occurs under the same conditions, but the individual-level relation is misapplied to the group. Incorrectly inferring from higher to lower levels can be called the disaggregation or decomposition fallacy. Alker (1969) described the ecological fallacy as when the effects of the grouping variable and within-group covariation interfere with the relation between ecological and individual effects. This explains why Robinson's (1950) individual-level correlation of ethnicity and literacy was 0.2, while the same correlation was 0.95 when aggregating the data by geographic region. Alker (1969) concluded that the research literature would be vastly different if all researchers were aware of these aggregation and spurious correlation fallacies.

Robinson's article was quite influential on social science research. Before the article was published, ecological correlations were often used to assess individual-level relations, but after the article was published, the use of ecological correlations sharply decreased (Firebaugh, 2001). Robinson's study discouraged the use of ecological-level analysis for many years, as he suggested that researchers only use ecological data when preferable individual-level data were unavailable. Recently, however, there has been a reinterest in ecological research and the problems that can occur from only examining individual-level relations (Pearce, 2000). In fact, some mechanisms actually do operate at the population level, and vital risk factors for diseases such as cancer and asthma have been discovered through population-level studies (Pearce, 2000). Singular consideration of individual-level effects may obscure important ecological factors that affect individual behavior above and beyond individual factors (Blakely & Woodward, 2000). For example, the combination of group norms for eating (ecological cause), as well as individual hunger (individual cause), is responsible for individuals' eating behavior. The same principle can be applied to longitudinal data as well: the combination of a person's trait stress, as well as their daily stress level, is responsible for their behavior.

Furthermore, individual variables that have been aggregated at the group level may measure a different construct than the corresponding individual variable. For example, average age of a sample may measure something different from individual age.

Also, the effect of an individual-level variable may differ depending on group-level factors: poverty is an individual factor determined by social context (Pearce, 2000). Poor people in a wealthy nation or city might face risk factors such as social exclusion and lack of access to fresh fruits and vegetables, whereas people of a similar income in impoverished nations may not face these same risk factors. Contrary to Robinson’s claim, individual-level variables are not always the most important predictors to consider in a study, and group-level variables are important to include as well. Given there is evidence that variables have effects that operate at multiple levels, simultaneous consideration of both individual- and group-level effects through multilevel analysis is ideal.

Subramanian, Jones, Kaddour, and Krieger (2009) demonstrated this by reanalyzing Robinson’s (1950) data using a binomial multilevel logistic model in order to show how using only single-level analyses, particularly individual-level analyses, could still result in a misspecified model. They found that state-level predictors had important effects on individual literacy and thus demonstrated the pitfalls of Robinson’s (1950) assertion that only individual-level analyses are “meaningful” (p. 341). Based on this understanding that individual-level studies can also be incomplete, many researchers have advocated that multilevel modeling can avoid both ecological and individualistic fallacies. Ultimately, multilevel models may provide the most complete way to avoid both aggregate and disaggregate fallacies, as data can be collected and analyzed simultaneously at multiple levels of inference.

Multilevel Modeling

Much of the literature on multilevel models considers cross sectional nested data where persons can be categorized into groups based on similar context. For example, students can be grouped by classrooms or schools, jury members can be grouped by the trial on which they serve, and patients can be grouped by the psychotherapist from which they receive treatment. Multilevel data such as these can be analyzed at two levels: the first level is the individual level (level-1) where observations from each individual are analyzed within the group. The aggregate level (level-2) relations are calculated by first forming some summary measure of the observations within each group (such as the mean, median, or standard deviation) and then estimating the relations among these aggregate measures called "contextual variables." It is important to consider that variables may operate at multiple levels, because if a level is left out of the analysis, variance associated with that level will be redistributed to the next lower and next higher levels (Snijders & Bosker, 2012). Erroneous standard errors may be obtained for coefficients of variables that are defined on this level—tests of such variables will be unreliable (Snijders & Bosker, 2012).

Much of the recent literature on the ecological fallacy has suggested that multilevel models should be used when individual-level data are available, so that effects at both individual and aggregate levels can be examined simultaneously. The limitations of single-level models (including the problem of ecological inference) can be overcome by estimating both levels in one model (Diez-Roux, 2000). Specifically, contextual multilevel models can separate level-1 predictor variables into level-1 and level-2 orthogonal (non-correlated) components. This separation of level-1 and level-2 effects not only aids in interpretability, but also it allows testing whether the effect of a predictor on an outcome is different at each level.

Persons as Contexts

Contextual multilevel models have been explicated mostly for cross-sectional data, but they can also be useful for longitudinal data. Although Enders and Tofighi (2007) argued that extending multilevel contextual analysis models to longitudinal data is easy, Hoffman and Stawski (2009) argued that it is not intuitive at all, hence the need for thorough demonstration of how to use existing contextual multilevel analyses to model persons as contexts in longitudinal data. However, despite several demonstrations of how to apply these models to longitudinal data, there has not yet been a thorough investigation of how these models perform with longitudinal data structures and what factors affect their performance. Longitudinal data can be analyzed at two levels: the first level is within each person, and the second level is between persons (Hoffman, 2015; Hoffman & Stawski, 2009). Thus, each person becomes a context, just as group membership or geographic location may be a context in cross-sectional nested data. With longitudinal data, estimating the contextual effect allows direct evaluation of whether between-person and within-person effects differ.

In longitudinal data, level-1 predictors are called “time-varying predictors.” These variables are not constant across time points in a study. Whereas predictors such as gender or ethnicity are generally considered to be stable no matter when they are measured, other predictors, like mood or tiredness, can vary over weeks, days, or even hours. Some time-varying predictors are expected to change stably over time. For example, children are expected to show a general increase in their language skills while in school. Other time-varying predictors, like mood, are generally expected to fluctuate over time without showing any stable growth. This paper focuses on the latter example, for reasons that are discussed below. Note that whether a predictor is expected to grow stably or fluctuate may depend on the window of data collection. Although some studies have shown that personality traits may change over the course of a lifetime, one would not expect them to show stable change within the course of a 30-day study.

Effects of time-varying predictors, whether they show stable change or fluctuate, can be separated into two separate constructs: trait (chronic) effects that reflect stable attributes of the person and state (acute) effects that represent short-term deviation from stable attributes. People may vary in their amounts of chronic stress due to personality differences, family situations, or career choice. However, there will still be days that individuals have either more or less stress than usual due to vacations, work deadlines, or interpersonal conflict.

For example, researchers designing a health intervention may ask participants to record their daily stress and daily food intake every day for a month. In this case, the researchers may be interested in whether higher stress (a time-varying predictor) is related to more calories consumed from sugar. When a time-varying predictor relates to an outcome at both within-person and between-person levels, the effect of the raw predictor is a blend of effects at both levels. A multilevel contextual model, however, is able to separate the effect of a time-varying predictor into level-1 and level-2 components. A multilevel contextual model could be used to determine whether, for individuals, more stress is related to more sugar consumption—a within-person (level-1) effect—or whether people who have a higher average amount of stress also eat a higher average amount of sugar—a between-person (level-2) effect. Just as many have shown that effects may not be equal across levels of a cross-sectional study, between- and within-person effects may not be equal as well.

Hoffman (2015) suggested that it is rare to find a time-varying predictor with variation at only level 1; in fact, she said it is a rule rather than an exception that between-person and within-person effects of a time-varying predictor will differ in magnitude or even direction. Snijders and Bosker (2012) likewise claimed that differing within-group and between-group regression effects were more common than not. Bolger and Schilling (1991) also suggested that it is common for between- and within-person effects in a multilevel model to differ. Contextual multilevel models can accommodate differing between- and within-person effects.

Generally, though, researchers collect longitudinal data because they are more interested in within-person effects than between-person effects. Indeed, Bolger and Laurenceau (2013) asserted that, although both between-person and within-person effects should be included in a multilevel model for proper estimation, between-person effects could often be ignored (p. 32). Although longitudinal data are often collected primarily to provide information about within-person relations, the data also provide important information about cross-sectional, between-person relations as well (Hoffman & Stawski, 2009; Curran & Bauer, 2011). If the effects of stable individual differences are ignored in longitudinal analysis, effects at the within-person level are confounded with the between-person level. Separating out the between- and within-person level effects in longitudinal data has three benefits: 1) there is no confounding of effects at within- and between-person levels, 2) one can examine the effects of between-person differences in addition to primary examination of within-person effects, and 3) one can examine a contextual effect that measures the discrepancy of between- and within-person effects.

Contextual Multilevel Models

Consider again the example of a study which measured *i* participants over *t* days in order to determine how strongly daily stress (X) was related to daily sugar consumption (Y). Again, note that daily stress and daily sugar consumption are not expected to change in a stable manner over time.[1] In multilevel model notation, the level-1 relation between stress and sugar consumption is:

$$Y_{ti} = \beta_{0i} + \beta_{1i}(X_{ti} - \bar{X}_i) + r_{ti}. \tag{1}$$

Y_{ti} is the number of calories from sugar consumed by person i on day t. β_{0i} is the intercept (mean) number of calories from sugar consumed by person i, and β_{1i} is the regression slope showing the effect of stress on calories from sugar for person i. Note that stress has been centered within person: each person's mean level of stress ($\bar{X}_i$) has been subtracted from their daily stress scores X_{ti}, so that $(X_{ti} - \bar{X}_i)$ represents daily deviations in stress from a person's mean level of stress. Centering individual predictor scores based on the person mean also implies that the intercept β_{0i} can be interpreted as the predicted amount of sugar calories consumed on a day with a mean amount of stress. r_{ti} is the level-1

[1] For analyzing real data, it is important to include elapsed time as a predictor, even if this is not a focal interest, as it may reflect other effects such as boredom or habituation to the measures (Iida et al., 2012). However, for simplicity, I will assume in the example here and in the simulated data that all effects due to time, like any other confounder, have been removed.

residual. Now each person's regression slope β_{1i} and intercept β_{0i} can be expressed as dependent variables at level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}\bar{X}_i + u_{0i} \tag{2}$$

$$\beta_{1i} = \gamma_{10} \tag{3}$$

In these level-2 equations, γ_{00} is the mean intercept across all people, and γ_{10} is the mean slope across all people. γ_{01} is the level-2 slope that indicates the effect of each person's mean stress on their mean sugar consumption, and u_{0i} is the level-2 residual. Now the equations from both levels can be combined into one model:

$$Y_{ti} = \gamma_{00} + \gamma_{10}(X_{ti} - \bar{X}_i) + \gamma_{01}\bar{X}_i + u_{0i} + r_{ti} \tag{4}$$

In this model, each person has a different regression intercept that characterizes the relation between daily stress and sugar consumption—all of the regression slopes across people are constrained to be equal. (Whether or not this is a tenable assumption is discussed later.) According to this model, there are two regression coefficients relating stress to sugar consumption: the within-person (level-1) regression coefficient γ_{10} and the between-person (level-2) regression coefficient γ_{01}. Note that the person-specific residuals u_{0i} and r_{ti} are not explicitly estimated by the model, but rather the variance of the level-1 residual is estimated Var(r_{ti}) = σ^2, and the variance of the level-2 residual is estimated Var(u_{0i}) = τ^2_{00}. A contextual effect is present if the within-person effect γ_{10} and the between-person effect γ_{01} are significantly different.

The comparison of between- versus within-person relations addresses a central issue in psychological research about whether group-level relations among variables generalize to individual group members (Molenaar, 2008). However, even if the between-person effect has been disentangled from the within-person effect, the model may still fail to account for some important between-person differences.

Idiographic Methods

For most of past century, psychology has operated with the assumption that the relations of variables found among a population of people can generalize to a population member's life trajectory (Molenaar, 2008). This variation between individuals, called interindividual variation (IEV), can be conceptualized as the covariance matrix formed by observing *V* variables for *N* subjects at one time point. Time-dependent variation within a single participant's time series, called intraindividual variation (IAV), can be conceptualized as the covariance matrix formed by observing *V* variables over *T* times for one subject in the population (Molenaar, 2004). If it is true that the variability among people can generalize to the variability over time for one person, then the variable by person covariance matrix for the population will be asymptotically equal to the variable by time covariance matrix for one member of the population (Molenaar, 2004).

Although it is often assumed that findings regarding the variation between people (IEV) apply to the variation within an individual (IAV), this generalization requires that the data meet a set of strict assumptions called ergodicity. The first aspect of ergodicity is homogeneity, meaning that all of the same members of a population follow the same laws of behavior and are essentially exchangeable units (Gu, Preacher, & Ferrer, 2014). For example, homogeneity states that the number of factors and factor loadings for a factor analysis of variable by time data are invariant across subjects (Molenaar, 2004). The second aspect of ergodicity is stationarity, meaning that the between- subjects variable by person data matrix has constant statistical parameters, such as mean, variance, or factor loadings, over time (Molenaar, 2004). Stated another way, ergodicity means that the statistical parameters (such as mean, variance, etc.) of a single person's trajectory over time are equal to the statistical parameters (such as mean, variance, etc.) of a group of trajectories at a single point in time (Gu et al., 2014). Gu et al., (2014) used a space-state model to look at time-related sequences among variables. Because this type of modeling allows estimating distinct models for all individuals, it allows the development of homogenous subpopulations.

Molenaar (2004) argued that few psychological processes meet the conditions for ergodicity. In fact, some processes, such as developmental or learning processes, by their very nature cannot be stationary because the moments of such data are necessarily time-varying (Molenaar, 2008). However, even for processes that do not involve stable change over time and meet the stationarity assumption, the homogeneity assumption is very strong and often violated. Furthermore, standard structural equation modeling of IEV does not detect even strong violations of the homogeneity assumption (Molenaar, 2008). Velicer (2010) presented several examples of how idiographic analyses revealed very different conclusions than corresponding nomothetic analysis. Thus with psychological data, between-subjects results are often not generalizable to individuals, and most analyses will not be able to detect this deficiency.

So far, this paper has discussed two main issues that can arise as the result of data aggregation. The first issue is the disaggregation or aggregation fallacy: variable relations found at one level of analysis are not necessarily equal to those found at another level of analysis. For example, does the relation between *X* and *Y* within person equal the relation between *X* and *Y* between person? This can be addressed by using a contextual multilevel model to estimate the relations at both levels simultaneously. The second issue is ergodicity, and specifically homogeneity: relations among variables may not be the same for all people. For example, is the relation between *X* and *Y* the same for each individual?

So in order to avoid committing an aggregation or disaggregation fallacy, a statistical model (such as a contextual multilevel model) should account for differing relations at both individual and aggregate levels. Also, in order to account for non-homogeneous processes, it should allow for individual differences. This can be done by allowing regression parameters to vary across individuals. Even in a multilevel model without predictors, the intercept will vary across individuals, but with at least one linear level-1 predictor, another parameter can be added that allows the level-1 linear regressions (within-person effects) to vary across individuals. Similarly, if quadratic level-1 predictors are added to the model, the degree of curvature can be allowed to vary across individuals.

However, individual-specific scores on the intercept and slope are not estimated, but rather the mean and variance of the intercept's and slope's distributions are estimated. Thus people can vary quantitatively in both intercept and slope. Still, the model is restricted such that each person still has the same functional form, i.e., one person cannot have a linear relation between *X* and *Y* while another person is allowed a exponential relation between *X* and *Y*. Thus, the model is not truly idiographic in the sense that a separate longitudinal model is estimated for each individual. However, estimating random slope variation helps ensure that at least some individual differences are accounted for. In order to meet the ergodicity conditions that allowed inference between within-person and between-person levels, daily diary data would have to be collected from a homogeneous sample, and the data would have to be stationary over time.

Random Slopes in Contextual Multilevel Models

A contextual analysis model with random slopes not only allows the comparison of within- and between-person effects, but also allows the estimation of between-person differences. Furthermore, random slopes allow estimation and documentation of variability, without needing to know the sources of the variability (Bolger & Laurenceau, 2013). Equation 5 includes a level-1 predictor divided into level-1 and level-2 effects, as well as a random slope of the level-1 predictor.

$$Y_{ti} = \gamma_{00} + \gamma_{10}(X_{ti} - \bar{X}_i) + \gamma_{01}\bar{X}_i + u_{0i} + u_{1i}(X_{ti} - \bar{X}_i) + r_{ti} \qquad (5)$$

Compared to the contextual multilevel model without a random slope as in equation 4, γ_{10} now indicates the average within-person effect of daily stress on daily sugar consumption, and u_{1i} indicates person i's random deviation from the average within-person regression slope γ_{10}. Just as person i's intercept u_{0i} is not explicitly estimated by the model, but rather the variance of all intercepts Var(u_{0i}) = τ^2_{00} is estimated, the variance of all person-specific slopes Var(u_{1i}) = τ^2_{11} is estimated, and the variance of person-level residuals Var(r_{ti}) = σ^2 is estimated. The covariance of slope and intercept Cov(u_{0i} , u_{1i}) = τ^2_{01} is also estimated. The other model parameters are interpreted the same as in equation 4, and γ_{01} - γ_{10} still indicates the contextual effect: the difference between the between-person effect and the average within-person effect. It is important to note that in the presence of large slope variation, the implication of the average within-person effect (and subsequently, the contextual effect) is more nuanced, as persons may deviate greatly from the average within-person effect. Furthermore, finding such large variation in within-person relations is ideally proceeded by examining potential predictors of such variation (Raudenbush & Liu, 2000). Although this is an important topic for future research, the current study remains only concerned with the contextual model as seen in equation 5, excluding any predictors of slope variation.

Daily Diary Studies

It is important to develop contextual multilevel models for longitudinal data, as many currently popular research designs depend on longitudinal data with time-varying predictors. One such type of research design, the daily diary study, includes both within-person and between-person data. Daily diary studies are used to collect repeated measurements on participants in a natural environment in real time (Bolger & Laurenceau, 2013). Such studies are becoming increasingly widespread in the social sciences, including social, personality, clinical, developmental, organizational, and health psychology (Iida, Shrout, Laurenceau, & Bolger, 2012). From 2009 to 2012, more than 250 journal articles per year used daily diary results (Iida et al., 2012). In comparison to data collected in a laboratory, or questionnaires that ask participants to recall the events and behaviors of the past, daily diary data are collected in a natural environment; they can also reduce retrospective bias, as the data can be collected closer in time to when the events and behaviors of interest occur (Iida et al., 2012).

Very basic daily diary studies can involve the collection of a single measure that is expected to change or vary over time. However, in many cases, psychologists are interested in the antecedents, correlates, and consequences of daily experience (Bolger, Davis, & Rafaeli, 2003). So, a daily diary study can not only help determine the within-person processes of the average person, but also a diary study can reveal how much people vary in their processes and what predicts this variation (Bolger et al., 2003). Some within-person processes might be expected to exhibit systematic and durable change over time, such as language acquisition in children. However, other within-person processes are assumed to vary over time, even if they do not exhibit systematic change over time, such as the effect of stress on relationship intimacy. Daily diary studies can address the questions of what predicts steady growth in language acquisition and what predicts daily variation in relationship intimacy. Much of psychology research has focused on the measurement of stable traits or steady states, but many of these constructs are not perfectly stable over time. For example, even a construct such as ethnic identity can vary from day to day (Yip & Fuligni, 2002). Many researchers may be interested in whether these variations are due to measurement error or meaningful within-person variability.

However, models for within-person change over time have been well explicated, but models for within-person fluctuation have not been so heavily explored (Hoffman & Stawski, 2009). Furthermore, longitudinal predictors that change systematically over time require more complex parameterization than those that simply fluctuate over time (Hoffman & Stawski, 2009). Curran and Bauer (2011) demonstrated how, in a model with a time-varying predictor that shows no growth and no cycling (as is discussed in this paper), the person mean centering strategy (explained below) provides a valid estimate of the within-person effect. However, it is more complex to extract the pure within-person effect from predictors that show growth over time, and person-mean centering does not provide a valid estimate in this case. For these reasons, this paper focuses on longitudinal data where variables are expected to fluctuate, rather than show steady change, over the interval of data collection.

Daily diary studies are an ideal application of multilevel contextual models because time-varying predictors almost always include both between- and within-person variability (Hoffman 2015). Partitioning between- and within-person variability is important even if one is only interested in examining effects at one level, and this step is often overlooked in analyses of intensive longitudinal data (Bolger & Laurenceau, 2013). Without separating predictor variables into between-person and within-person components, the effect of the predictor confounds both sources of variability, and the results cannot be properly interpreted to apply at either level (Enders & Tofighi, 2007; Hoffman & Stawski, 2009; Hoffman, 2015). Contextual multilevel models offer an intuitive way to analyze such longitudinal data where within-person fluctuation (rather than systematic change) is expected. These models also directly address the question of whether within-person or between-person relations are equal and generalizable to the other level.

Centering in Longitudinal Multilevel Models

When applying multilevel models to real data, the choice of where to center predictor variables is important. Time-varying predictors are generally composed of two sources of variation, so they are essentially two variables in one. Whether or not these dual sources of variation is viewed as interesting or a nuisance is important to consider. Proper specification and interpretation of time-varying predictors is complex. If the contextual effect of a person-mean predictor is significant, then the composite effect should not be used (Hoffman & Stawski, 2009, p. 108). Predictor variables can be left as raw scores, they can be centered at the grand mean, or they can be centered at the cluster mean. In grand mean centering, the mean of the predictor across all observations is subtracted from each score. Thus, the predictor now becomes a measure of deviation from the overall mean. In cluster-mean centering, the cluster mean is subtracted from each score in the cluster. In this case, the predictor is now a measure of deviation from the group mean. This applies to cross-sectional data, where the cluster is some grouping of persons, and to longitudinal data, where the cluster comprised of multiple observations for each person. As this project focuses on the latter case, cluster-mean centering is referred to as "person-mean centering" for clarity. The centering of level-1 predictors in studies of individual change affects the definition of the intercept, intercept variance, and intercept-slope covariance, as all of these parameters are interpreted at the zero value of the predictor. The centering of level-1 predictors also affects possible biases in studying time-varying covariates, as well as the estimation of variance parameters (Raudenbush & Bryk, 2012).

In multilevel models without a random slope, person-mean centering and grand-mean centering produce equivalent estimates of the contextual effect. This means that the fixed effect parameter estimates (including the contextual effect) from a grand-mean centered model can be algebraically equated to those from a person-mean centered model, even though the actual estimates are not exactly the same (Kreft, de Leeuw, & Aiken, 1995; Enders & Tofighi, 2007). Thus, the choice of centering method is less important as parameter estimates obtained from either method provide the same information about the size of the within-person effect, the size of the between-person effect, and the difference between these estimates (the contextual effect). Person-mean centering of time-varying predictors is an intuitively appealing option, as effects at both levels are directly represented; however, grand-mean centering is still more commonly used (Hoffman, 2015).

However, this equivalence no longer holds once a random slope is introduced into the model, and so far, there have not been many empirically based recommendations for which centering method should be used when a contextual model includes random slopes. Under grand-mean-centering, the random effect is based on all variation, whereas under person-mean-centering, the random effect is based on within-person variation (Hoffman & Stawski, 2009). Snijders and Bosker (2012) advised taking substantive theory and model fit into consideration, but ultimately advocated against using person-mean centering for random slopes unless there is clear theory that deviation scores are related to the outcome rather than raw scores. There are not many straightforward recommendations in the multilevel literature to indicate which method of centering is preferable in a model with a random slope. Either substantive reasoning, such as the case where daily deviations from a person's mean level of stress are the focus of interest, or empirical criteria, such as the estimated AIC or BIC, could be used to choose which centering method to use when estimating a model with a random slope (Hoffman, 2015). There is some evidence that in grand-mean centered models, random slope variance is downwardly biased compared to person-mean centered models, perhaps due to the models' differences in random intercepts (Hoffman, 2015). Wang and Maxwell (2015) suggested, based on a simulation, that person-mean centering provided the most accurate and precise estimates of fixed and random effects. Although there remains much work to be done on which centering method is preferable for multilevel models with random slopes, this project focuses on person-mean centered predictors, as this method presents a clear and interpretable way to differentiate between orthogonal between-person and within-person variation in a predictor.

Why Contextual Effects Arise

Although many experts on multilevel modeling suggest that differing between- and within-person effects are common (Hoffman, 2015; Snijders & Bosker, 2012; Bolger & Schilling, 1991), it is not as clear why such effects differ. These differences could be due to substantive and theoretical differences at each level, or they could be due to statistical artifacts, such as unreliability in the measure or the effects being measured on different scales.

Contextual effects may arise because some constructs have different meanings at the between-person versus within-person levels, as was discussed for cross-sectional data as well. We might expect that a person's average level of positive mood relates to their quality of life differently than their amount of positive mood on any given day. If a daily-observed variable were an entirely different construct than its corresponding aggregate, then it seems likely that being different constructs, they would have different effects on an outcome. Schwartz (1994) also argued that aggregate measures can be qualitatively different constructs than their corresponding individual-level measures. For example, the proportion of females in a group can affect group eating behavior in a different way than the individual's gender affects his or her eating behavior. Furthermore, the reasons why persons differ from one another may not be the same reasons why any given person varies from time to time.

However, contextual effects may also arise because of statistical reasons: between-person and within-person effects may differ simply because they are unstandardized coefficients estimated on different scales, as shown by Hoffman and Stawski (2009) and Hoffman (2015). To the extent that the variation in the predictors and outcomes is not equally distributed across between- and within-levels, effects at each level will likely differ simply due to this unequal distribution of variation (Hoffman & Stawksi, 2009). One way to ensure that any contextual effect is not simply due to this statistical property is to calculate standardized fixed effects. Unlike in OLS linear regression, the formulae for calculating standardized regression coefficients are not as straightforward. However, one can obtain pseudo-standardized coefficients by multiplying the unstandardized coefficient by its sample standard deviation and dividing it by the residual variance of Y at its level. Equations 6 and 7 demonstrate how to obtain the pseudo-standardized between-person coefficient (γ_{01std}) and within-person coefficient (γ_{10std})

$$\gamma_{01std} = \gamma_{01} * \frac{SD(X_B)}{SD(Y_B)} \tag{6}$$

$$\gamma_{10std} = \gamma_{10} * \frac{SD(X_W)}{SD(Y_W)} \tag{7}$$

where $SD(X_B)$ and $SD(X_W)$ give the standard deviation of X between-person and within-person, respectively. The standard deviation of X at each level can be obtained simply through sample statistics. In the case of the person-mean centered model in equation 5, the standard deviation at each level can be computed as:

$$SD(X_B) = SD(\bar{X}_i) \tag{8}$$

$$SD(X_W) = SD(X_{ti} - \bar{X}_i). \tag{9}$$

However, there are different ways to calculate the standard deviation of Y at the between- and within-levels. Hoffman and Stawski (2009) suggested obtaining the level-2 and level-1 residual variance of *Y* using a means-only multilevel model (that is, an unconditional multilevel model with no predictors) so that

$$SD\ (Y_B) = \sqrt{\tau^2_{\mathrm{E}}} \tag{10}$$

$$SD\ (Y_W) = \sqrt{\sigma^2_{\mathrm{E}}}\ . \tag{11}$$

Applying these formulae to real data can indicate if the contextual effect is simply due to the comparison of unstandardized regression coefficients. If the standardized contextual effect $\gamma_{01\mathrm{std}} - \gamma_{10\mathrm{std}}$ is still non-zero, however, this indicates that there is still a contextual effect.

Previous Research

There is scant literature on the power, bias, and accuracy of contextual effects. Only a few published statistical simulation studies have examined contextual effects or random slopes. Only one published study has used a statistical simulation to test a multilevel model that includes both a contextual effect and a random slope variance component, although this study only considered 144 conditions (Wang & Maxwell, 2015).

Only two published papers have done statistical simulations focusing on the estimation of contextual effects. Both of these studies were done to address the problem of unreliability in the level-2 predictors. When level-2 predictors are created by simply aggregating level-1 observation within each group, there is an implicit assumption that this group mean is perfectly reliable. If the group mean is not perfectly reliable, then the contextual effect and its standard error can be biased. Lüdtke, Marsh, Robitzsch, and Trautwein (2008) and Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov, and Muthén

(2011) introduced a multilevel latent covariate (MLC) model that corrects for this unreliability and provides more unbiased estimates in some conditions.

In particular, Lüdtke et al. (2008) simulated data based on a within-cluster centered contextual model and varied the number of groups, number of observations within each group, ICC_x, and sampling ratio (the proportion of the population of group members that were actually sampled for each group.) For each simulated data set, they estimated a contextual multilevel model using manifest group means (the multilevel manifest covariate, or MMC, approach as seen in equation 4) as well as an MLC model. They found that whereas the MMC approach produced more biased estimates of the contextual effects, the MLC approach produced contextual effects with more variability. However, Lüdtke et al. (2008) provided recommendations for when the MMC versus MLC approaches were appropriate based on the nature of the aggregated construct. For formative constructs, they recommended the MMC approach in most cases, especially with high ICC_x values and large cluster sizes. For reflective constructs, they recommend using the MLC approach.

In the case of daily diary data, it is not always clear-cut whether daily score aggregation represents a formative or reflective construct. Lüdtke et al. (2008) argued that if within-cluster variation in scores were a substantively meaningful group characteristic, then the construct would be considered formative. However, if within-cluster variation in scores were simply unreliability or lack of agreement among individual scores, then the construct would be considered reflective. For daily diary data, fluctuation in a construct over time often may be substantively meaningful, particularly if the construct indicates mood or affect. If large daily fluctuations in negative mood are observed for one person, this variability may be related to other important variables (including average negative mood) and should not necessarily be dismissed as "unreliability" in daily measures of mood. For this reason, it is appropriate to use a multilevel manifest covariate approach to estimate contextual effects with daily diary data, as discussed here. Furthermore, the MMC approach is by far the most commonly used in research practice (Lüdtke et al., 2008), so the findings of this study will be useful to those who collect daily diary data and use manifest group means as between-person predictors.

Lüdtke et al. (2011) compared four different contextual models: those that corrected for either sampling error or measurement error, a model that corrected for both, and a model that corrected for neither. The model that corrected for neither sampling error nor measurement error was equivalent to the random-intercept contextual model in equation 4. The models that corrected for sampling error decomposed observed predictor scores into observed and unobserved components, while the models that corrected for measurement error assumed the independent variable was measured by multiple variables. Lüdtke et al. (2011) showed mathematically that uncorrected or partially corrected models would result in biased estimates of the contextual effect. However, in a simulation study with similar conditions to those used in Lüdtke et al. (2008), Lüdtke et al. (2011) found that partial correction approaches can actually perform better than the full correction approach in some circumstances. Specifically, the partial correction approaches do better than the full correction approach when there are a small number of clusters and a low ICC_x. Still, both of these papers focused on data characteristics that are common in cross-sectional data: including lower ICC_x values. Furthermore, in both simulation studies, the size of the contextual effect was not varied (it was set at 0.5).

Wang and Maxwell (2015) used a contextual multilevel model with a random slope to compare various methods of centering and detrending. However, they only considered three contextual effects (-0.5, 0, 0.5) and found that the person-mean centering approach was preferable to grand-mean centering or no centering in estimating both fixed and random effects.

Curran and Bauer (2011) performed a very small simulation to test the parameter recovery of a contextual multilevel model with a between-person effect $\gamma_{01} = 1.5$ and a within-person effect $\gamma_{10} = -1.0$. With 500 simulated cases and 9 repeated measures, they found that the model estimates were very accurate. Further simulation research is needed to understand the behavior of the multilevel contextual model with contextual effects of all sizes and directions, including zero and negative contextual effects.

Increased power to detect fixed effects can be achieved by increasing the duration of the study, or increasing the frequency of observations. All else being equal, increasing the duration of the study will increase reliability (Raudenbush & Bryk, 2002.) It is important to note that these findings from Raudenbush and Bryk (2002) focused primarily on multilevel models applied to longitudinal data where stable change over time was expected. Bolger and Laurenceau (2013) demonstrated how increasing the number of level-2 units (in this case, persons), was more effective in increasing power to detect a level-2 treatment-by-time slope than increasing the number of level-1 units (repeated measurements per person.)

Research on the performance of significance tests for variance components in multilevel models indicates that in general, it may be difficult to detect non-zero slope variance. Snijders and Bosker (2002) found that there was low power for the significance test of slope variance. Hertzog, Lindenberger, Ghisletta, and von Oertzen (2006) similarly found that slope covariance was difficult to detect in most cases. However, Rast and Hofer (2014) found that longitudinal studies did have adequate power to detect slope variances and covariances.

Hox (2002) showed that fixed effects were estimated more precisely than random effects. It is common practice for researchers to only include a cross-level interaction in the multilevel model if the slope variance is significant (La Huis & Ferguson, 2009). However, La Huis and Ferguson (2009) used real multilevel data to demonstrate that there could be a significant cross-level interaction, even if there was no significant slope variance. Thus, it may not be good practice to include a cross-level interaction only after finding a significant slope variance.

Raudenbush and Liu (2000) found that power for tests of variance components and fixed effects related to the number of groups, group size, and effect size, and the greatest determinant of power to detect significant slope variance was group size. La Huis and Ferguson (2009) tested the power and Type I error rates of three significance tests for slope variance: the chi-square test, the likelihood ratio test, and the corrected likelihood ratio test. They varied the number of groups, group size, and effect size of slope variance. Effect size of slope variance was specified by varying the degree to which the slope was related to a level-2 predictor. La Huis and Ferguson (2009) found that there was not a great difference in performance among the three tests, although the one-tailed likelihood ratio test had the best balance of low Type I errors and high power. They also found that group size and effect size had the biggest influence on the power to detect significant slope variance (La Huis & Ferguson, 2009). However, they found that significance tests for slope variance components do not always correspond directly to the relation of level-2 to the slope. La Huis and Ferguson (2009) suggested that future simulation research should vary ICC values, as well as fixed effect parameter sizes to determine their effect on the power of variance components tests. If a random slope is omitted, the standard errors of cross-level interactions may be incorrectly estimated (Snijders & Bosker, 2012). Snijders and Bosker (2012) advised that in order to precisely estimate variance parameters, at least 30 observations should be used at either level and the independent variable should have enough dispersion within level-2 units.

The Present Study

This study investigates the statistical performance of the contextual model with a random slope for longitudinal within-person fluctuation data. Furthermore, the few existing studies on contextual multilevel models have either not examined a variety of contextual effect sizes, not included a random slope parameter or have focused on data characteristics common in cross-sectional research, rather than those common in longitudinal research. In particular, this study incorporates larger ICC values than those studied by Lüdtke et al. (2008), as well as contextual effects of different sizes and directions. Typical dairy studies have ICC values ranging from .20 to .40 (Bolger & Laurenceau, 2013). Further research has been recommended for elaborating guidelines on how to choose sample size approximately for the estimation of variable parameters in random slope models (Snijders & Bosker, 2012). Curran and Bauer (2011) argued that existing multilevel models are not always appropriate for the wide variety of longitudinal data structures and much quantitative work is needed to make these methods more applicable to longitudinal data.

Incorporating a random slope makes a multilevel model even more flexible to accommodate individual differences, which again relates to Molenaar's admonition that between-person and within-person relations should never be assumed equal. Again, however, the model in the current study is not a truly idiographic model where individuals are allowed to have entirely different models. But assuming that the individuals in the data are homogenous with respect to the functional form of the X-Y relation over time, estimating a random slope allows for any heterogeneity that may be present in the strength or direction of the X-Y relationship. To my knowledge, there has been only one limited simulation study of a contextual model with random slopes (Wang & Maxwell, 2015), and in fact, Lüdtke et al. (2008) specifically suggested this as an area for future research.

The secondary focus of this study is to further explore standardized contextual effects as described by Hoffman and Stawski (2009) and Hoffman (2015). In their example with real data, they showed that even when standardized to the same scale, the between and within effects still differ. Comparing the unstandardized and standardized contextual effects with simulated data can demonstrate how frequently a significant contextual effect is only due to the regression coefficients being on different scales and how frequently the standardized between and within effects are truly different. If seemingly large and significant contextual effects are only due to this problem of different scales, this would undermine the idea that contextual effects are substantively important and meaningful. The goal of this work on standardized contextual effects is to address a potential criticism of contextual effects and thus address a criticism of the more general idea that between and within person effects will often differ.

Hypotheses

Effect Size of Fixed Effects

Larger effect size of the within-person and between-person fixed effects should increase power to detect these effects. However, Lüdtke et al. (2008) showed mathematically that the contextual effect will be underestimated if the between-person effect is larger than the within-person effect in the population. Likewise, the contextual effect should be positively biased in the sample if the between-person effect is smaller than the within-person effect (Lüdtke et al., 2008).

Effect Size of Slope Variance

Smaller slope variance will result in better power to detect the average within-person effect (Bolger & Laurenceau, 2013), and pilot studies I performed have shown that smaller slope variance also increased power for the contextual effect.

Sample Size

Number of persons (*i*) will be more important than the number of observations per person (*t*) in increasing power to detect the within-person effect, between-person effect, and contextual effect (Raudenbush & Liu, 2000). However, the number of observations per person (*t*) will have a greater impact than the number of persons (*i*) on increasing power to detect the slope variance. Increasing the number of observations per person (*t*) should decrease bias of the contextual effect.

Intraclass Correlation (ICC_x)

Larger ICC_x values of the predictor should result in less bias for the contextual effect (Lüdtke et al., 2008). Pilot studies I performed have shown that larger ICC_x values will increase power to detect the contextual effect. These pilot studies also showed an interaction with ICC_x and number of observations per person (*i*) such that larger ICC_x improved coverage and Type I error rates for the contextual effect and slope variance at smaller sizes of *i*, but larger ICC_x corresponded with poorer outcomes at larger sizes of *i*.

CHAPTER 2

METHOD

Simulation Conditions and Procedure

Data were simulated in Mplus 7 according to Equation 5:

$$Y_{ti} = \gamma_{00} + \gamma_{10}(X_{ti} - \bar{X}_i) + \gamma_{01}\bar{X}_i + u_{0i} + u_{1i}(X_{ti} - \bar{X}_i) + r_{ti} \tag{5}$$

where Y_{ti} and X_{ti} represent outcome and predictor scores, respectively, for person i at time t. This study varied the number of observations per person (t), number of persons (i), within-person fixed effect size (γ_{10}), between-person fixed effect size (γ_{01}), random slope variance effect size (τ^2_{11}), and intraclass correlation (ICC_x) of X. The intraclass correlation of X represents the proportion of variation in X that is between-person, relative to the total variation in X. So then:

$$ICCx = \frac{Var(X_B)}{Var(X_W) + Var(X_B)}. \tag{12}$$

For equation 5 that is person-mean centered, the variation within person and between person can easily be obtained from the predictor terms $(X_{ti} - \bar{X}_i)$ and $\bar{X}_i$ as they are orthogonal. So then for a person-mean centered multilevel model:

$$ICCx = \frac{Var(\bar{X}_i)}{Var(X_{ti} - \bar{X}_i) + Var(\bar{X}_i)}. \tag{13}$$

The performance of the contextual effect and the slope variance parameter estimates was assessed by calculating parameter bias, parameter variability, and standard error accuracy. The simulation followed a factorial design with six manipulated data characteristics including number of observations per person (t = 5, 10, 20, 40, 80), number of persons (i = 30, 50, 100, 150, 200), intraclass correlation for the predictor variable (ICC_x = .10, .20, .50, .60), between-person fixed effect size (γ_{01} = -.59, -.14, 0,

.14, .59), within-person fixed effect size (γ_{10} = -.59, -.14, 0, .14, .59), and random slope variance effect size (τ^2_{11} = 0.05, 0.01, 0.15). Contextual effect sizes ($\gamma_{\text{contextual}} = \gamma_{01} - \gamma_{10}$) ranged from -1.18 to 1.18. The other model parameters were specified as: 0 for the intercept, 0 for the intercept-slope covariance, 0 for the mean of *X* at both between and within levels, 1 for the total variance of *X* and *Y*, and 0.2 for the unconditional intraclass correlation of *Y* given by:

$$ICCy = \frac{\tau_E^2}{\tau_E^2 + \sigma_E^2}. \tag{14}$$

In generating data according to equation 5, only the residual variance of *Y* at between and within levels could be specified explicitly. However, Snijders and Bosker (2012) provided an equation (Appendix B) that gave the variance decomposition of *Y* for a multilevel model with random slopes. Based on the simulation parameters in each study condition, the residual variance of *Y* at each level was computed so that *Y* in each population model had a total variance of 1 and an unconditional ICC_y = .2. Mplus 7 was used to generate the data and to analyze correctly specified models using maximum likelihood estimation. Appendix C gives example Mplus code for data generation and estimation of the contextual model and estimation of a means-only model for calculating the standardized contextual effect. All data characteristics were fully crossed. There were 5 (*t*) x 5 (*i*) x 4 (ICC_x) x 5 (γ_{10}) x 5 (γ_{01}) x 3 (τ^2_{11}) = 7,500 conditions, with 1,000 replications generated for each condition. The independent variables and their respective levels were chosen based on previous Monte Carlo studies and common findings in the substantive research.

Much of the other simulation work with multilevel models has focused on their use with cross sectional data, and thus the chosen simulation parameters have reflected the sample sizes and effect sizes commonly seen in cross-sectional data. However, the present study's focus is on repeated observations within persons, so following Ohly, Sonnentag, Niessen, and Zapf, (2012) this study varied the number of observations per person between 5 and 80 and varied the number of persons between 30 and 200. Ohly et al. (2012) suggested that daily diary research published in top journals generally have at least 5 repeated measurements per person and at least 100 participants, although anything smaller than 30 may lead to biased results. Likewise, Lüdtke et al. (2008) chose relatively low ICC_x values for their simulation, based on commonly observed ICC_x values in organizational research. However, ICC_x values chosen for the present study reflect the higher ICC estimates obtained in daily diary research. Typical dairy studies have ICC values ranging from .20 to .40 (Bolger & Laurenceau, 2013), although larger ICC values of .60 are also commonly seen. ICC values of .50 and .60 were chosen to explore how having equal or larger amounts of variation at between-person level affects estimation. Effect sizes were chosen following Cohen's (1988) small and large benchmarks for regression slope parameters, and a slope parameter of 0 was added to test for Type I error rates. In most of the existing simulations of contextual models, effect size was not of primary interest. For example, Lüdtke et al. (2008) and Lüdtke et al. (2011) only considered a within-group regression slope of 0.2 and a between-group regression slope of 0.7 leading to a contextual effect size of 0.5. Curran and Bauer (2011) used a within-group regression slope of -1.0 and a between-group regression slope of 1.5 leading to a contextual effect size of 2.5. This study is one of the first to examine a range of contextual effect sizes, including negative effect sizes, and the effect of a zero contextual effect size on Type I error rates.

Slope variance sizes were chosen following Raudenbush and Liu’s (2000) rules of thumb for small, medium, and large variances of a fixed effect. A small variance of 0.05 implies that the fixed effect has a standard deviation of 0.22, a medium variance of 0.10 implies that the fixed effect has a standard deviation of 0.32, and a large variance of 0.15 implies that the fixed effect has a standard deviation of 0.39.

Dependent Variables

This study assessed the performance of the parameter estimates for the contextual effect and slope variance by calculating parameter bias, parameter variability, and standard error accuracy. The accuracy of a standardized contextual effect was also evaluated (Hoffman & Stawski, 2009; Hoffman, 2015). For each replicated data set, a means-only model for *Y* was estimated, and a standardized contextual effect was calculated as in equations 6 and 7. This standardized contextual effect was compared to the population standardized contextual effect.

Parameter bias was evaluated by calculating raw parameter bias, relative parameter bias, and standardized parameter bias. Raw parameter bias was calculated by subtracting the true value of the parameter from the within-cell average of the simulated parameter estimate. Relative parameter bias was calculated by dividing the raw parameter bias by the parameter’s true value, and standardized parameter bias was calculated by dividing raw parameter bias by the standard deviation of the parameter across all replications in each design cell. Parameter variability was measured by estimating the root mean square error (RMSE) across all observations in a cell.

Standard error accuracy was evaluated by observing the coverage of the 95% Confidence Interval (CI) and standard error (SE) bias. If the true parameter was within the confidence interval, then coverage was given a value of 1; if the true parameter was outside of the confidence interval, then coverage was given a value of 0. The proportion of times the true parameter lies above the CI and the proportion of times the true parameter lies below the CI was also calculated. Raw standard error bias was calculated by finding the difference between the within-cell mean of the standard errors of the parameter estimates across and the within-cell standard deviation of parameter estimates. Statistical power was also be assessed by calculating the proportion of times across all replications that a parameter was significant at the 0.05 alpha level when there was a true non-zero effect. Type I error rates were similarly studied by observing the proportion of times across all replications that a parameter was significant at the 0.05 alpha level when the true value of the effect was 0. All significance tests for the parameters of interest were performed using a Wald chi-square test available in Mplus. Although there have been some indications that a one-tailed likelihood ratio test would be ideal for testing variance components, the chi-square test is still widely used by applied researchers. Furthermore, La Huis and Ferguson (2009) concluded that the chi-square test did not result in any large differences from likelihood ratio tests.

CHAPTER 3

RESULTS

The results section includes four parts. The first part describes the findings from the full simulation proposed in the prospectus meeting. ANOVA was used to determine which study factors affected parameter recovery and whether the hypotheses were supported. For each ANOVA, effects that had an effect size of $\eta^2 \geq .10$ were reported and described. Corresponding tables and figures were included for simulation factors had an effect size of $\eta^2 \geq .10$. Then, for a subset of the original conditions, replication-level logistic regression analyses were conducted to see which factors affected coverage, power, and Type I error at the replication level. Then effect of autocorrelated residuals on parameter recovery was also described for a single condition. Finally, an empirical illustration of how to fit a contextual multilevel model to real daily diary data was described.

Statistical Analyses

The main purpose of this study was to examine the effect of sample size, covariate effect size, number of indicators, and quality of indicators on simulation outcomes using analysis of variance (ANOVA). ANOVAs for the full simulation were conducted at the cell level, where a given outcome measure across all replications in one study condition was computed and treated as one observation. In a simulation study such as this, the large sample size had high statistical power to detect even very small effect sizes. Results therefore focused on the effect size measure $\eta^2 = SS_{effect}/SS_{total}$. Omega squared ($\omega^2 = (SS_{effect} – MS_{error})/(SS_{total} + MS_{error})$) was also calculated for each analysis, but $\omega^2$and η^2 were nearly identical, so only η^2 was reported here. Specifically, factors that showed an effect size of $\eta^2 \geq .10$ were interpreted, independent of other factors in the model. Values of $\eta^2 \geq .10$ are considered to be at least "medium" effect sizes (Cohen, 1988) and using η^2 to determine meaningful effects has been used in other simulation studies (Krull & MacKinnon, 1999). For specific effects that were hypothesized, the benchmark was also used to determine which effects were meaningful.

Using an effect size benchmark of $\eta^2 \geq .10$ was appropriate for this study, as the analyses were performed on cell-level means (discussed below.) Performing the same analyses on the replication versus cell level leads to differing effect sizes, as the cell-level data removes some of the variability in the replication-level data, inflating the effect sizes. For example, two ANOVA models for raw bias of the contextual effect were estimated in a subset of 112 conditions. Computed at the replication level, the effect of level-1 sample size was $\eta^2 = .0004$, but computed at the cell level, the effect of level-1 sample size was $\eta^2 = .258$. Thus, a somewhat exclusive $\eta^2 \geq .10$ threshold was chosen to account for the effect size inflation that occurred when analyzing cell-level means.

Ideally, ANOVAs would be conducted at the individual-replication level in order to account for within-cell variability. However, with a total of 7,500,000 replications, PROC GLM did not have enough memory to compute a factorial ANOVA. So, ANOVAs were conducted at the cell level. The initial cell-level ANOVAs included all possible interactions up to five-way interactions, but these models would not converge. The SAS User's Guide suggested that specifying Type III sums of squares (SS) may create computational difficulties in a large model, and other SS types could be specified to decrease the computational load. However, using Type III SS was necessary when were are missing design cells (as was the case with this data where some simulation cells had zero converged replications), so this solution was untenable. The User's Guide also suggested using the ABSORB or REPEATED statements or instead performing the analysis with PROC ANOVA or PROC REG, which were also not appropriate for this study. The other suggestion was to eliminate terms, especially high-level interactions. Eliminating the five-way interaction terms allowed the models to converge. Still, the effects of the five-way interaction terms were still of interest. This was addressed by conducting replication-level ANOVAs on a subset of the simulation data, as described below.

Convergence

No convergence problems were encountered when estimating the unconditional models (i.e., means-only multilevel models for Y with no predictors). However for the conditional multilevel models, one-tenth (750) of the 7,500 simulation cells failed to converge for all replications in the cell. This occurred where $ICC_x = 0.6$ and $\gamma_{01} = -.59$ or $\gamma_{01} = .59$. Many of the error messages for these replications indicated a non-positive definite covariance matrix. Because of the high rates of non-convergence for the $ICC_x = 0.6$ condition, this factor was removed from subsequent analyses, leaving 5,625 simulation conditions available for further analysis

Across the other conditions, 72 conditions had a least one replication that did not converge, totaling 81 non-converged replications. This only occurred in conditions where $\gamma_{01} = -.59$ or $\gamma_{01} = .59$ and level-1 sample size was equal to 5, 10, or 20 and $ICC_x = .2$ or $ICC_x =.5$. In most cases, only one replication per cell did not converge, but seven conditions had two replications that did not converge, and one condition had three replications that did not converge. All replications were also checked for out-of-bounds estimates. All converged replications had non-negative variance estimates.

Parameter Bias

Relative bias was originally proposed as the primary measure of parameter accuracy. However, relative bias cannot be computed when the true value of the parameter is zero, excluding 5 of the 25 total contextual effect size conditions. Relative bias was evaluated for the applicable conditions and found that while relative bias was generally low for the slope variance (M = -0.014%, SD = 0.41%), the contextual effect (M < 0.001%, SD = 0.57%), and standardized contextual effect (M = 0.001%, SD = 0.83%), in a few cases, relative bias was quite high for the standardized contextual effect estimate. For example, where the true value of the standardized contextual effect was small but non-zero (i.e., $\gamma_{01std} – \gamma_{10std} = \pm 0.049$), the average relative bias was as high as 112%. However, raw bias for the standardized contextual effect did not differ greatly between conditions with a small true standardized contextual effect and a large true standardized contextual effect. Thus, the large relative bias values were not due to the absolute difference between estimated and true values, but rather dividing the raw bias by a small true value ($\gamma_{01std} – \gamma_{10std} = \pm 0.049$). Across slope variance, contextual effect, and standardized contextual effect, the pattern of raw and relative bias did not differ greatly, so only raw bias was reported here.

In order to deal with the issue of zero and near-zero true values, standardized bias was estimated as well, by taking the mean of within-cell parameter estimates minus the true value of the parameter, and dividing that difference by the within-cell parameter estimate standard deviation. Standardized bias was reported as a percentage of the within-cell standard deviation (SD) of the parameter, 100 x (average estimate – true value of parameter)/within-cell SD of parameter estimate; standardized bias values above an absolute value of 40% are seen as problematic (Collins, Schafer, & Kam, 2001). However, standardized bias can be difficult to compare across sample size conditions, as the within-cell SD of a parameter estimate generally decreases with larger sample sizes. In the case where sample size is very large and the within-cell SD is very small, standardized bias may appear quite large compared to smaller sample size conditions, even if the raw bias is similar across sample size conditions. For that reason, raw bias was also examined in this study.

Standardized bias for the slope variance ranged from -19.00% to 23.8% (M = -6.33%, SD = 5.52%) across all cells included in the analysis. Level-1 sample size explained the largest amount of variance in slope variance standardized bias ($\eta_t^2 = .355$), followed by the interaction of level-1 and level-2 sample size ($\eta_{t \text{ x } i}^2 = .216$), and level-2 sample size ($\eta_i^2 = .212$). Slope variance standardized bias was generally reduced by increasing both level-1 and level-2 sample size, although the interaction of level-1 and level-2 sample size meant that in some cases, slope variance standardized bias increased slightly with larger sample size, as seen in Figure 1. However, this may have been an artifact of standardization. The variability of the slope variance estimate decreased with larger level-1 sample sizes, inflating the standardized bias of the slope variance for large level-1 sample size conditions.

Standardized bias for the contextual effect ranged from -2.50% to 9.83% (M = 3.15%, SD = 3.06%). Level-1 sample size explained the largest amount of variance in contextual effect standardized bias ($\eta_t^2 = .687$), followed by the interaction of level-1 and level-2 sample size (η_i^2 = .249). Contextual effect standardized bias actually showed a general increase with larger level-1 sample size, although because of the interaction effect, that increase was not always monotonic. In many cases, contextual effect standardized bias was lowest for *t* = 5 or *t* = 10 and highest for *t* = 40, as seen in Figure 2.

Standardized bias of the standardized contextual effect ranged from -27.17% to 28.18% (M = 2.98%, SD = 5.19%). The interaction of level-1 sample size with level-2 fixed effect explained the largest amount of variance in standardized bias of the standardized contextual effect ($\eta_{t \text{ x } \gamma 01}^2$ = .271), followed by level-1 sample size (η_t^2 = .213) and level-2 fixed effect ($\eta_{\gamma 01}^2$ = .139), as well as the interaction of ICC_x and level-2 fixed effect size ($\eta_{ICCx \text{ x } \gamma 01}^2$ = .108), as seen in Figure 3. In general, standardized bias increased with larger level-2 effect size, decreased with ICC_x values, and increased with level-1 sample size, although the highest average standardized bias occurred with *t* = 40. The interaction occurred such that the difference in standardized bias among level-2 effect size conditions increased with larger ICC_x values.

Average raw bias for the slope variance ranged from -0.010 to 0.019 (M = -0.002, SD = 0.002). Slope variance raw bias was affected by level-2 sample size (η_i^2 = .239), slope variance effect size ($\eta_{\tau 211}^2$ = .136), and level-1 sample size (η_t^2 = .114). Level-2 sample size also interacted with slope variance effect size ($\eta_{i \text{ x } \tau 211}^2$ = .111) and level-1 sample size ($\eta_{i \text{ x } t}^2$ = .105), as seen in Figure 4. While raw bias decreased with larger level-2 sample size, the amount of decrease was smaller with larger level-1 sample size. Likewise, the effect of increasing level-2 sample size was generally stronger for slope variance effect size conditions τ^2_{11} = .15 than for τ^2_{11} = .05.

Average raw bias for the contextual effect ranged from -0.004 to 0.027 (M = 0.004, SD = 0.005). Contextual effect raw bias was affected by level-1 sample size (η_t^2 = .404) level-2 sample size (η_i^2 = .132), and the interaction of level-1 and level-2 sample size ($\eta_{t \text{ x } i}^2$ = .265), as seen in Figure 5. In general, raw bias decreased with larger level-2 sample size and increased with larger level-1 sample size, although the largest average raw bias occurred with *t* = 40. The interaction among sample size conditions did not display a clear pattern of effects on raw bias. Therefore, interaction of level-1 and level-2 sample size was investigated by plotting the raw bias against effective sample size (as shown in Figure 6) calculated by the following

$$N_{effective} = \frac{N_n}{1+(n-1)ICC_y} \tag{15}$$

where N_n represents the total sample size (in this case $N_n = t*i$), *n* represents the level-1 sample size (in this case *n* = *t*), and ICC_y is the ICC of the outcome variable, which was generated to be .20 in all conditions in this study (Snidjers & Bosker, 2012). While the general trend was decreasing raw bias with larger effective sample size, the trend was not monotonic. Relative bias showed a similar pattern. However, considering that average raw bias for the contextual effect only ranged from -0.004 to 0.027, some of the fluctuations in raw bias among sample sizes could have been due to sampling error.

Average raw bias for the standardized contextual effect ranged from -0.11 to 0.12 (M = 0.003, SD = 0.01). Standardized contextual effect raw bias was affected by the interaction of level-1 sample size and level-2 effect size ($\eta_{t \text{ x } \gamma 01}^2$ = .267), such that the effect of raw bias decreasing with larger level-1 sample sizes was generally stronger for larger level-2 effect sizes (γ_{01} = ± 0.59), than for level-2 effect sizes near zero, as seen in Figure 7. There was also a three-way interaction of level-1 sample size with level-2 effect size and level-2 sample size ($\eta_{t \text{ x } \gamma01 \text{ x } i}{}^2$ = .298), such that the effect of increasing level-2 sample size on decreasing raw bias was largest for smaller level-1 sample size (i = 5) and large level-2 effect sizes (γ_{01} = ± 0.59).

In this study, contextual effects had positive average raw bias across all contextual effect sizes as seen in Table 1. Whereas positive raw bias was expected for contextual effects less than zero, negative raw bias was expected for contextual effects greater than zero (Lüdtke et al., 2008). However, Lüdtke et al. (2008) based their expectations on analytical, rather than simulation, work.

Although ICC_x corresponded to decreasing raw bias of the contextual effect (F = 1.4 x 10^7, $p < .001$), this hypothesized effect was not supported as the effect was small ($\eta_{\text{ICCx}}{}^2$ = .054). Contrary to expectation, level-1 sample size (t) actually increased the raw bias of the contextual effect ($\eta_t{}^2$ = .404), which went from an average of 0.001 at t = 5 to an average 0.006 at t = 200, as seen in Table 2.

Difference between Unstandardized and Standardized Contextual Effect

The raw difference between unstandardized and standardized contextual effect was also calculated as an outcome. The average difference ranged from -0.57 to 0.57 (M = 0.00, SD = 0.15). The largest factor affecting variability in the difference between unstandardized and standardized contextual effects was the interaction of level-2 effect size and ICC_x ($\eta_{\gamma01 \text{ x ICCx}}{}^2$ = .832). For ICC_x = .20, the difference between unstandardized and standardized contextual effects was near zero, and the difference was much greater for the ICC_x values above and below .20; however, this effect decreased with smaller level-2 effect sizes until γ_{01} = 0 conditions had near-zero differences for all ICC_x conditions, as seen in Figure 8. The difference in standardized and unstandardized contextual effects was negligible in conditions where $ICC_x = .20$, because then $ICC_x = ICC_y$, and the standardized contextual effect was computed to take into account differing ICC values for X and Y (Hoffman, 2015).

Parameter Variability

Parameter variability was assessed by calculating the empirical standard deviation and root mean squared error (RMSE) for the slope variance estimate, contextual effect estimate, and standardized contextual effect estimate. Results for empirical standard deviation and RMSE were very similar for each outcome, in terms of proportion of variance explained by each effect. So only RMSE results were reported here.

RMSE for the slope variance estimate ranged in magnitude from 0.01 to 0.12 (M = 0.03, SD = 0.02). While slope variance RMSE decreased with larger level-1 ($\eta_t^2 = .392$) and level-2 ($\eta_i^2 = .302$) sample sizes, it increased with larger slope variance effect sizes ($\eta_{\tau 211}^2 = .146$), as seen in Figure 9.

RMSE for the contextual effect ranged from 0.03 to 0.37 (M = 0.14, SD = 0.07). Contextual effect RMSE decreased with larger level-2 sample sizes ($\eta_i^2 = .528$) and larger ICC_x values ($\eta_{ICCx}^2 = .333$), as seen in Figure 10.

RMSE for the standardized contextual effect ranged from 0.02 to 0.80 (M = 0.12, SD = 0.07). Standardized contextual effect RMSE decreased by both increasing level-1 ($\eta_t^2 = .583$) and level-2 sample sizes ($\eta_i^2 = .133$), as seen in Figure 11.

Standard Error Accuracy

Standard error accuracy was assessed by calculating 95% coverage and power of the slope variance and non-zero contextual effect estimates. Coverage rates exceeding 90% (Collins et al., 2001), and power rates exceeding 80% (Cohen, 1988) were considered satisfactory. Type I error rates were also assessed for zero contextual effect estimates, and any Type I error rates that fell within the liberal interval criterion of [2.5%, 7.5%] were considered acceptable (Bradley, 1978). Cell-level raw standard error bias was also calculated for the slope variance and contextual effect by subtracting the within-cell standard deviation of the parameter estimate from the average standard error of the parameter, across replications in a cell.

Coverage

Coverage for the slope variance estimate ranged from 83.6% to 99.1% (M = 91.6%, SD = 2.97%). Both sample size conditions, as well as the interaction between them, accounted for at least 10% of the variance in slope variance 95% coverage. Slope variance coverage generally decreased with larger level-1 sample size ($\eta_t^2 = .152$) and increased with larger level-2 sample sizes ($\eta_i^2 = .419$), as seen in Figure 12. There was also an interaction among sample sizes ($\eta_{t \text{ x } i}^2 = .135$) such that the differences in coverage among level-2 sample size conditions were greater for larger level-1 sample size conditions. Coverage for the contextual effect ranged from 89.8% to 95.8% (M = 93.3%, SD = 1.1%). Contextual effect coverage increased as level-2 sample size increased ($\eta_i^2 = .670$), as seen in Figure 13.

The hypothesized interaction between ICC_x and level-2 sample size on contextual effect coverage was not supported for slope variance coverage rates as seen in Table 3. Although this effect was significant (F = 184.58, $p < .001$), it only explained .4% of the variance in slope variance coverage ($\eta_{i \text{ x ICCx}}^2 = .004$).

The hypothesized interaction between ICC_x and level-2 sample size was not supported, as seen in Table 3 ($\eta_{i \text{ x ICCx}}^2 = .037$). Coverage did increase with ICC_x for $i =$

30 and 50, but decreased with ICC_x for *i* =100 and 150. With *i* = 200, coverage decreased from ICC_x = .10 to .20, but then increased from ICC_x = .20 to .50. Furthermore, the interaction between ICC_x and level-2 sample size only explained 3.7% of the variance in contextual effect coverage.

Power and Type I Error

Power to detect the slope variance ranged from 2.1% to 100% (M = 85.8%, SD = 27.1%). Cell-level power to detect the slope variance increased monotonically by increasing either level-1 (η_t^2 = .405) or level-2 (η_i^2 = .139) sample size. There was an interaction of level-1 and level-2 sample size ($\eta_{t \text{ x } i}^2$ = .103), such that the power difference among level-2 sample size conditions was much greater at smaller level-1 sample size conditions, and this power difference decreased with larger level-1 sample size conditions, as seen in Figure 14. Power to detect the contextual effect ranged from 10% to 100% (M = 77.1%, SD = 30.8%). Cell-level power to detect the contextual effect was influenced level-2 effect size ($\eta_{\gamma01}^2$ = .368), level-1 effect size ($\eta_{\gamma10}^2$ = .348), and the interaction between level-1 and level-2 effect sizes ($\eta_{\gamma01 \text{ x } \gamma10}{}^2$ = .183). Cell-level power to detect the contextual effect was highest for the largest parameter values of $\gamma_{01} – \gamma_{10}$, as seen in Figure 15.

Level-2 sample size (*i*) was the only study factor that explained at least 10% of the variance in contextual effect Type I error rates (η_i^2 = .665), and Type I error decreased as level-2 sample size increased, as seen in Figure 16.

As expected, power to detect both within- and between-person effects increased dramatically as the effect size at each respective level increased as seen in Table 4. For example, power to detect the between-person effect was at 40.20% for γ_{01} = .14, but there was 95.53% power for $\gamma_{01} = .59$ ($\eta_{\gamma 01}{}^2 = .729$). Likewise, power to detect the within-person effect was at 77.86% for $\gamma_{10} = .14$ and 99.98% for $\gamma_{10} = .59$ ($\eta_{\gamma 10}{}^2 = .299$).

Although smaller slope variance effect sizes corresponded to greater power to detect the within-person effect (F = 166504, *p* < .001), as seen in Table 5, this hypothesized effect was not supported as the effect was small ($\eta_{\tau 211}{}^2 = .017$). Likewise, the hypothesis of contextual effect power increasing with smaller slope variance was not supported. Even though this effect was significant (F = 93.65, *p* < .001), the effect explained less than .1% of the variance in contextual effect power ($\eta_{\tau 211}{}^2 < .001$).

As expected, level-2 sample size (*i*) had a larger impact than level-1 sample size (*t*) on increasing power to detect the within-person effect ($\eta_i{}^2 = .226$, $\eta_t{}^2 = .072$), between-person effect ($\eta_i{}^2 = .062$, $\eta_t{}^2 = .005$), and contextual effect ($\eta_i{}^2 = .084$, $\eta_t{}^2 = .009$), as seen in Table 6. Also, level-1 sample size (*t*) had a greater impact than level-2 sample size (*i*) on increasing power to detect the slope variance ($\eta_t{}^2 = .405$, $\eta_i{}^2 = .139$). Type I error rates by level-1 and level-2 sample size for the fixed and random effects can be seen in Table 7.

Although larger ICC_x values corresponded to increased power to detect the contextual effect (F = 45936, *p* < .001) and decreased contextual effect Type I error (F = 445.28, *p* < .001), as seen in Table 8, these hypothesized effects were not supported as the effects were small (contextual effect power $\eta_{ICCx}{}^2 = .053$, contextual effect Type I error $\eta_{ICCx}{}^2 = .019$).

The hypothesized interaction of level-2 sample size and ICC_x was also not supported for contextual effect Type I error rates, also seen in Table 9 (F = 225.33, *p* < .001; $\eta_{i \text{ x } ICCx}{}^2 = .038$). Type I error did decrease with larger ICC_x for *i* = 30, 50, and 200, but it increased with ICC_x for *i* =100. With *i* = 150, Type I error increased from ICC_x = .10 to .20, but then increased from ICC_x = .20 to .50. Furthermore, the interaction between ICC_x and level-2 sample size explained 3.8% of the variance in contextual effect Type I error rates.

Raw Standard Error Bias

Average raw bias for the slope variance SE ranged from -0.006 to 1.06 (M = 0.002, SD = 0.033). None of the simulation factors or interactions accounted for more than 10% of the variability in slope variance SE raw bias. Average raw bias for the contextual effect SE ranged from -0.035 to 0.357 (M = -0.006, SD = 0.011). Larger level-2 sample sizes corresponded to decreasing contextual effect SE raw bias ($\eta_i{}^2$ = .222), as seen in Figure 17.

Replication-Level Analyses

In order to investigate the results more closely, a subset of the total simulation data were selected and the effects of the study conditions were assessed for outcomes that were calculated at the individual level. This included coverage, power, and Type I error for the slope variance and contextual effect. Logistic regression with effect coding for the simulation factors was used, except for the contextual effect, which was treated as continuous. All possible interactions were included in the analysis.

To choose the subset of conditions for the logistic regressions, conditions were considered that corresponded to variability in power and coverage for the slope variance and contextual effect. Among level-2 sample size conditions, the two smallest conditions *i* = 30 and *i* = 50 typically had the most variability in outcomes. Among the level-1 sample size conditions, *t* = 5 and *t* = 40 typically had the most variability in outcomes.

While increasing level-1 sample size generally improved outcomes, for some parameter recovery (for example, see Figure 2), t = 40 had worse outcomes. Thus, these four sample sizes were selected for further investigation. Also, there were 25 total effect size conditions in the full study: 5 conditions where $\gamma_{01} – \gamma_{10} = 0$, 10 conditions where $\gamma_{01} – \gamma_{10} > 0$, and 10 conditions where $\gamma_{01} – \gamma_{10} < 0$, as seen in Table 10. Outcomes were similar for contextual effect sizes of the same absolute value; i.e., $\gamma_{01} – \gamma_{10} = -.59$ showed similar outcome patterns to $\gamma_{01} – \gamma_{10} =.59$. Therefore, 7 unique effect size conditions where $\gamma_{01} – \gamma_{10} \geq 0$ were selected to further investigate. This included all $\gamma_{01} = .59$ conditions, as well as ($\gamma_{01} = .14$, $\gamma_{10} = -.14$) and ($\gamma_{01} =.14$, $\gamma_{10} = 0$).

Analysis of the full simulated data set showed that slope variance effect size did not have a large effect on many of the outcomes, nevertheless small and large slope variance effect sizes $\tau^2_{11} = .05$ and $\tau^2_{11} = .15$ were selected to study further as this was a primary factor of interest in this study. Finally, two ICC_x conditions $\text{ICC}_x = .10$ and $\text{ICC}_x = .50$ were chosen to investigate further. Table 9 shows the disparity between $\text{ICC}_x = .10$ and $\text{ICC}_x = .50$ in terms of power to detect the slope variance and the contextual effect. Whereas power to detect the contextual effect is greatest for $\text{ICC}_x = .50$, power to detect the slope variance is greatest for $\text{ICC}_x = .10$. Thus the subset of conditions chosen for analysis at the replication level included 112 conditions (2 x 2 x 7 x 2 x 2) with 1000 replications per condition. Only four replications in this subset did not converge.

As with the ANOVAs described earlier, there was high power to find even small effects with such a large sample size, so only effects with an odds ratio greater than 4 or less than .25 (indicating a large effect) are described (Rosenthal, 1987). An OR = 4 can be converted into Cohen's d = .765, or medium-large effect size (Chinn, 2000).

Replication-Level Coverage

Table 11 shows replication-level 95% coverage rates for the slope variance and the contextual effect. Slope variance coverage was above 90% for most conditions except where *t* = 40 and *i* = 30. In this case, it ranged from 83.7% to 85.8%. Slope variance coverage also ranged from 87.9% to 89.2% for conditions where (*t* = 5, *i* = 30, ICC_x = .20, τ^2_{11} = .15). There were several other *i* = 50 conditions where slope variance coverage was 89.8% to 89.9%. Slope variance coverage values were very similar, and in some cases identical, across some contextual effect sizes. This likely occurred because the absolute value of the between-person effect was equal for some contextual effect sizes, and the analysis on the full data set indicated that the between-person fixed effect had almost no effect on slope variance coverage ($\eta_{\gamma 01}{}^2$ = .00005).

No simulation factors or interactions had OR > 4 or OR < .25 for slope variance or contextual effect coverage. Replication-level slope variance coverage was higher in τ^2_{11} = .05 conditions than τ^2_{11} = .15, but only in conditions where the level-1 sample size was *t* = 5. For *t* = 40 conditions, there was negligible difference in coverage among slope variance effect sizes. Replication-level contextual effect coverage was above 90% for all conditions. Contextual effect coverage values were very similar, and in some cases identical, across some contextual effect sizes. This likely occurred because the absolute value of the between-person effect was equal for some contextual effect sizes, and the analysis on the full data set indicated that the between-person fixed effect had a very small effect on contextual effect coverage ($\eta_{\gamma 01}{}^2$ = .007). So in this smaller data set, there was not enough variability in between-person effect sizes to show any relationship between contextual effect size and contextual effect coverage.

Replication-Level Power

Table 12 shows replication-level power rates for the slope variance and contextual effect. Power to detect the replication-level slope variance was above 80% for all t = 40 conditions except where (i = 30, ICC_x = .50, τ^2_{11} = .05), where power ranged from 60.1% to 75.5%. For t = 5 conditions, power was only above 80% in two cases: when (i = 50, ICC_x = .10, τ^2_{11} = .15, $\gamma_{01} - \gamma_{10}$ = 0) or (i = 50, ICC_x = .10, τ^2_{11} = .15, $\gamma_{01} - \gamma_{10}$ = 1.18). In the rest of the t = 5 conditions, power to detect the slope variance ranged from 2.3% to 70.9%. In every case, power to detect the slope variance was much higher for $\gamma_{01} - \gamma_{10}$ = 0 or $\gamma_{01} - \gamma_{10}$ = 1.18 than for the other contextual effect sizes.

Increasing level-1 sample size (OR_t = 0.005) and increasing slope variance ($OR_{\tau 211}$ = 0.026) increased the power to detect slope variance. However, increasing the ICC of the predictor variable actually decreased power to detect the slope variance (OR_{ICCx} = 4.464). There was also an interaction between level-1 sample size and slope variance effect size ($OR_{t \text{ x } \tau 211}$ = 8.834) such that the difference in power between slope variance effect sizes was much more pronounced for t = 5 conditions than for t = 40 conditions. As was the case with coverage, slope variance power was identical for some contextual effect size conditions, likely because the analysis of the full data set showed the between-person fixed effect had almost no effect on slope variance power ($\eta_{\gamma 01}{}^2$ = .00007).

The only factor that consistently had more than 80% power to detect the contextual effect in all cases was γ_{01} - γ_{10} = 1.18. Across the other contextual effect sizes, power to detect the contextual effect ranged from 10.8% to 100%. Power to detect the contextual effect increased with larger contextual effect sizes ($OR_{\gamma 01 - \gamma 10}$ = 20819.32).

This effect was moderated by level-1 sample size (OR $_{\text{t x } \gamma 01 - \gamma 10}$ = 0.236) such that the influence of increasing contextual effect size was more pronounced for *t* = 5 conditions than for *t* = 40 conditions. The effect of contextual effect size was also moderated by the ICC of the predictor variable (OR $_{\text{ICCx x } \gamma 01 - \gamma 10}$ = 0.023) such that with ICC$_{x}$ = .10, power grew more slowly with increasing contextual effects compared to ICC$_{x}$ = .50, for which power increased more rapidly for increasing contextual effects.

Replication-Level Type I Error

Table 13 shows replication-level Type I error rates for the contextual effect. No simulation factors or interactions had OR > 4 or OR < .25 for contextual effect Type I error. For ICC$_{x}$ = .10, Type I error exceeded the liberal 7.5% criterion (Bradley, 1978) in all but one case—where (*t* = 5, *i* = 50, τ^{2}_{11} = .05), Type I error was 7.3%. For ICCx = .50, Type I error was below the 7.5% criterion in all but one case—where (*t* = 40, *i* = 50, τ^{2}_{11} = .05), Type I error was 7.7%.

Standardized Contextual Effect

Tables 14, 15, and 16 display the within-cell standard deviation of the standardized within-person effect, standardized between-person effect, and standardized contextual effect, respectively. Although these standardized effects take into account the ICC of both predictor and outcome variables, they do not have standard deviations of one. Still, these effects are useful to calculate, especially when the ICC of predictor and outcome variables differs greatly. Perhaps it is more helpful to view the standardized contextual effect as a descriptive effect size, rather than a truly standardized coefficient.

Autoregression

Although the simulation studies just described generated the data under the assumption of independence among the within-person residuals, it is likely that a longitudinal study would include some dependencies among the residuals. For example, the amount of stress experienced on day 1 of a daily diary study is likely to affect the amount of stress experienced the following days, although the longer time passes, the weaker this relation becomes.

First, the relation of autocorrelation to ICC in the simulated data was considered by estimating lag-1 autoregression of both X and Y in a subset of conditions, although autocorrelation was not specified in the data generation. Autoregression was estimated for the set of conditions where level-1 sample size $t = 5$ and $t = 10$, level-2 sample size $i = 30$ and $i = 50$, between-person effect $\gamma_{01} = 0$, within-person effect $\gamma_{10} = 0$, $\text{ICC}_x = .10$, .20, .50, .60, and slope variance $\tau_{11}^2 = .05, .10, .15$. For each replication, PROC MIXED TYPE = AR(1) with Y as the outcome variable was used in a model with no predictors to estimate the lag-1 autocorrelation of Y. Uncentered X was treated likewise to determine the lag-1 autocorrelation of X.

In general, the lag-1 autocorrelation for both X and Y was slightly less than the respective ICCs that had been generated ($\text{ICC}_x = .10, .20, .50$, $\text{ICC}_y = .20$). Table 17 shows the means and standard deviations of lag-1 autocorrelations across ICC_x conditions. Note that while the ICC of X was varied as a simulation condition, the ICC of Y was generated to be .20 in all simulation conditions.

In order to investigate how well the contextual multilevel model performed under the assumption of autocorrelation, data from a multilevel contextual model was generated with 75 persons and 5 observations per person, a level-2 fixed effect of $\gamma_{01} = .45$, a level-1 fixed effect of $\gamma_{10} = .14$, and a slope variance of $\tau_{11}^2 = .15$. The autocorrelation of the outcome variable was generated at $\rho = .7$, which meant that the correlation among residuals for time 1 and time 2 was $\rho = .7$, the correlation among residuals for time 1 and time 3 was $\rho^2 = (.7)^2 = .49$, …, and the correlation among residuals for time 1 and time 5 was $\rho^4 = (.7)^4 = .240$. Similar autoregression parameters have been used in other simulations of daily diary data structures (Seidman, Shrout, & Bolger, 2006).

To facilitate specification of the autoregression, the data were generated in a latent growth curve, rather than multilevel, framework. The data for the other simulation studies in this paper were generated as univariate, with each observation representing person *i*'s score on X and Y at time *t*. However, the autoregression data had to be generated as multivariate with one observation per person and variables $Y_1 \ldots Y_t$, $X_1 \ldots X_t$ representing person *i*'s scores for X and Y at each time point. Following Hoffman (2015) and Curran, Lee, Howard, Lane, and MacCallum (2012), individual Y variables were regressed on the individual X variables to estimate a random within-person effect. A latent intercept factor was estimated for the Y variables and for the X variables. In many applications, a slope factor for the outcome and/or predictor would be included as well, but the current study was focused on variables that fluctuate, rather than change steadily, over time. So, the data generating model was limited to latent intercepts for predictor and outcome. Regressing the latent Y intercept on the latent X intercept yielded the contextual effect, or the effect of average X on average Y above and beyond the within-person effect. Thus, the level-2 fixed effect in the data-generating model ($\gamma_{01} = .45$) represented the contextual effect, and the sum of the fixed effects ($\gamma_{01} + \gamma_{10} = .14 + .45 =$

.59) represented the between-person effect (Curran et al., 2012). For this single condition, 1000 replications were generated and a correctly specified model was fit to the data.

A second model was fit to the data where the autoregressive structure was not accounted for and the residuals of the Y variables were uncorrelated. To further compare the effect of autocorrelation on parameter recovery for the contextual multilevel model, the generated data were transformed to “long” univariate format and the contextual multilevel model was fit (as seen in Equation 5). Results of the three models can be seen in Table 18.

Across all models, standardized bias of the contextual effect and slope variance was below the 40% threshold, except for the contextual effect estimated in the contextual multilevel model, where it exceeded -110%. Variability of the contextual effect, assessed by the Mean Square Error (MSE) was lowest for the contextual multilevel model and highest for the latent growth model with uncorrelated residuals. MSE for the slope variance was lowest for the two growth models and highest for the contextual multilevel model. Coverage for the contextual effect was above 90% for all models except the contextual multilevel model, where it was 78.5%. Overall, power to detect the contextual effect was very low with across the four models, ranging from 14% to 32%. Coverage for the slope variance was above 89.3% for all models, and power to detect the slope variance was above 80% for all models.

Empirical Illustration

The example data were taken from Wave 2 of the National Survey of Midlife in the United States (MIDUS II): Daily Stress Project, 2004-2009 (Ryff & Almeida, 2004-2009). The main study sample was composed of (*N* = 1079) participants, 56.35% (*N* =

608) of which were women. The mean age was 57.6 years (SD = 12.50). Participants in the study answered questions via telephone interview each night over the course of eight days. Questions targeted the experience of day-to-day life stressors, as well as physical and emotional well-being. The effect of severity of physical symptoms on stressor negative affect was considered

To assess severity of physical symptoms, for every symptom experienced, participants rated the daily severity (1-10; 1 = Very Mild; 10 = Very Severe) of the physical symptoms. Severity of physical symptoms was calculated as an average of the individual severity items. Higher scores reflected higher severity. Severity of physical symptoms had *M* = 2.62, *SD* = 1.89. Figure 18 shows how severity of physical symptoms fluctuated over the course of the study for a subset (*N* = 242) of participants. Black lines represent an ordinary least squares (OLS) regression line for individuals' physical symptom severity, and the thick red line is an overall spline curve. The flatness of the overall curve and the fluctuation of the individual scores over time (also assessed using a spline function, not shown) indicated that using a person-mean centered contextual multilevel model was appropriate.

To assess stressor negative affect, the respondents indicated how much of the time they experienced each emotion over the past 24 hours: How angry were you feeling? How nervous or anxious were you feeling? How sad were you feeling? How shameful were you feeling? Each item was rated on a 0-3 scale (0 = Not at all; 3 = Very); and then the average across items was computed. Stressor negative affect had *M* = .81, *SD* = .57. Figure 19 shows how stressor negative affect fluctuated over the course of the study for a subset of participants. Black lines represent a smoothed spline curve for individuals' stressor negative affect, and the thick red line is an overall spline curve. The flatness of the overall curve and the fluctuation of the individual scores over time (also assessed using a spline function, not shown) indicated that using a person-mean centered contextual multilevel model was appropriate.

First, severity of physical symptoms was person-mean centered to reflect daily deviations in one's mean amount of physical symptom severity. As in Equation 5, person-mean centered physical symptom severity and mean physical symptom severity were used to predict the amount of stressor negative affect, allowing for variability in the slope parameter. Figure 20 shows the linear relation between physical symptom severity and stressor negative affect for a subset of the data, where each black line indicates one participant. The thick red line indicates a small positive relation between physical symptom severity and stressor negative affect, although some of the individual-level slopes had a strong positive or strong negative relation. Including a random slope in the model should account for this variability in the relation of physical symptom severity and stressor negative affect. Note that the model did not account for any weekly cycling in the predictor or outcome variables, as the data collection occurred over the course of eight days, making any weekly cycle difficult to detect. A similar study with a longer data collection period should consider a modeling strategy that accounts for weekly trends in the data (Liu & West, 2015). Likewise, a similar study with data collected monthly or yearly should account for systematic growth that might be seen in the variables (Wang & Maxwell, 2015; Curran & Bauer, 2011).

In order to aid in convergence, observations with missing values for severity of physical symptoms were excluded from the analysis, leaving N = 1024 participants in the final analysis. Analysis in Mplus showed that there was a significant average within-person effect of physical symptom severity on stressor negative affect (γ_{10} = 0.025, p = .019), indicating that on average, for every person's 1-unit increase in daily physical symptom severity, daily stressor negative affect increased on average by 0.025. There was also a significant between-person effect of physical outcome severity on stressor negative affect (γ_{01} = 0.072, $p < .001$), indicating that if an person increased their average stressor severity by 1 unit, their stressor negative affect would increase, on average, by 0.072. There was also a significant contextual effect ($\gamma_{01} - \gamma_{10}$ = 0.047, p = .004), indicating that the between-person effect was significantly larger than the average within-person effect. The slope variance was significant (τ^2_{11} = 0.007, p = 0.034), indicating that people significantly differed in how their daily physical symptom severity related to their daily stressor negative affect. The standardized contextual effect was $\gamma_{01std} - \gamma_{10std}$ = 0.297, indicating that the difference of between-person effect and within-person effect was not simply due to the discrepant ICCs of predictor and outcome variables.

CHAPTER 4

DISCUSSION

Daily diary studies can be used in social science research to assess within-person relations. However, data from daily diary studies also contain information about between-person relations. It is important that within-person and between-person relations are both estimated, as otherwise, the two effects are conflated. Contextual multilevel models can simultaneously estimate both between-person and within-person effects. The purpose of this study was to examine the performance of contextual multilevel models with random slopes for daily diary-type data. Sample size, effect size, and ICC were varied in a Monte Carlo simulation to investigate how these factors influenced parameter recovery.

Summary and Discussion of Results

Across all simulated data, coverage, power, and Type I error were assessed for the slope variance and contextual effect estimates. Parameter bias and variability were also assessed for the full set of simulation conditions. For the full simulation conditions, parameter estimates were generally unbiased, and in most cases, increasing level-1 or level-2 sample size had the largest effects on decreasing parameter bias.

Although larger ICC_x values corresponded to decreasing raw bias of the contextual effect as predicted (Lüdtke et al., 2008), this effect fell below the $\eta^2 \geq .10$ threshold set for this study ($\eta_{\text{ICCx}}^2 = .054$). This discrepancy may have been because Lüdtke et al. (2008) considered far fewer total conditions in their simulation study. Lüdtke et al. (2008) also showed analytically how contextual effect raw bias would be negative when the contextual effect was above zero ($\gamma_{01} > \gamma_{10}$) and positive when the contextual effect was below zero ($\gamma_{01} < \gamma_{10}$), although bias would decrease towards zero with larger level-1 sample size and larger ICC_x. However, the contextual effect in the current study was consistently overestimated for all fixed effect size conditions. This was likely because of the larger ICC_x conditions and larger level-1 sample size conditions considered in this study versus the analytical work of Lüdtke et al. (2008). Lüdtke et al. (2008) showed that asymptotically, as the level-1 sample size increased and as ICC_x increased, expected contextual effect bias decreased towards zero.

Raw bias of the contextual effect actually increased with larger level-1 sample sizes, contrary to another hypothesis that larger level-1 sample sizes would decrease contextual effect raw bias, although the increase was minimal and likely due to sampling error, as raw bias in general was near zero. Increasing level-1 or level-2 sample size had the largest effects on decreasing parameter variability. Slope variance RMSE also increased with larger slope variance effect size and contextual effect RMSE decreased with larger ICC_x.

Coverage for the contextual effect and slope variance was generally between 90% and 95% for most conditions—the exception being $i = 30$ conditions, where slope variance coverage averaged 88%. Ideally, coverage should be close to 95% to ensure that parameter estimates are accurate. However, based on a more liberal criterion of 90%, most of the conditions showed acceptable coverage for both contextual effect and slope variance estimates.

Average power was 89.6% for the within-person effect, 64.7% for the between-person effect, 77.1% for the contextual effect, and 84.8% for the slope variance across all the data conditions included in this study. However, power to detect random and fixed effects estimates varied widely across sample size and effect size conditions. Among the

25 total sample size conditions, there were no conditions where the average power to detect the between-person effect was above 80%, 20 conditions where the within-person effect power was at least 80%, 13 conditions where the contextual effect power was at least 80%, and 18 conditions where the slope variance power was at least 80%. This indicates that while the within-person effect and the slope variance were generally detectable, the contextual effect and between-person effect remains difficult to detect. Although some have claimed that it is often difficult to detect a non-zero slope variance (Snijders & Bosker, 2002; Hertzog et al., 2006), the current study found that there was adequate power to detect slope variance even with a small number of individuals, provided that there were enough repeated measurements (Rast & Hofer, 2014). This finding corresponded to the suggestion of Snijders and Bosker (2012) that at least 30 units at either level were needed to precisely estimate variance parameters. Future research should consider the accuracy of slope variance hypothesis tests for contextual multilevel models, and whether a bootstrap test might be appropriate based on the skewed distribution of slope variance parameters.

Average Type I error was 5.7% for the within-person effect, 6.8% for the between-person effect, and 6.7% for the contextual effect. Ideally, Type I error should be close to 5%, although based on a liberal criterion of 2.5%-7.5%, most conditions with at least i = 50 level-2 units showed acceptable Type I error rates for the contextual effect. However, the high rates of Type I error for i = 30 conditions indicate that estimating a contextual multilevel model with this sample size may lead to incorrect conclusions about the presence of a contextual effect.

When examined at the cell level, increasing level-1 or level-2 sample size had the largest effects on increasing coverage accuracy and power and decreasing Type I error, except for power to detect the contextual effect estimate. As predicted, level-2 sample size was more influential than level-1 sample size in increasing power to detect the within-person effect, between-person effect, and contextual effect (Bolger & Laurenceau, 2013; Raudenbush & Liu, 2000), but level-1 sample size was more influential in increasing power to detect the slope variance (Raudenbush & Liu, 2000; La Huis & Ferguson, 2009). Power to detect the contextual effect estimate was strongly influenced by increasing within-person and between-person effect sizes.

Power to detect a fixed effect is affected by the variability of the estimator. When the within-person effect estimate varies among people, as is the case when there is non-zero slope variance, then the effect becomes more difficult to detect. Although larger slope variance estimates corresponded to lower power to detect the within-person effect, consistent with previous findings (Bolger & Laurenceau, 2013), this effect ($\eta_{\tau211}^2 = .017$) was below the threshold set for this study. As the contextual effect is calculated as the difference of between-person and within-person effects, the standard error (and thus the power) of the contextual effect is affected by the standard error of the within-person effect. So, the power to detect the contextual effect should also be affected by slope variance. However, the hypothesized effect of the amount of slope variance on power to detect the contextual effect power was not supported in this study. Although the effect of slope variance on both the within-person effect and contextual effect power was not found in this study, this may have been due to the limited range of slope variance effect sizes chosen (τ^2_{11} = .05, .10, .15). Larger effects may have been found by increasing the range of slope variance effect sizes.

The hypothesized effect of ICC_x on contextual effect power was also not supported. Larger ICC_x values corresponded to increased contextual effect power as expected, but the effect was small ($\eta_{ICCx}{}^2$ = .053). The predicted interaction of level-2 sample size and ICC_x also did not have effects on slope variance coverage ($\eta_{i\ x\ ICCx}{}^2$ = .004), contextual effect coverage ($\eta_{i\ x\ ICCx}{}^2$ = .037), or Type I error ($\eta_{i\ x\ ICCx}{}^2$ = .038). This hypothesis had been developed based on findings from a pilot study. However, the data generation was specified slightly differently in the pilot study, where ICC_x was generated not as the unconditional ICC, but the model-conditional ICC of X. Furthermore, simulation conditions in the pilot study differed slightly from those chosen for the current study. In particular, the pilot study included conditions where ICC_x = .6, which were ultimately excluded from the current study, as those conditions where ICC_x = .6 and γ_{01} = .59 or -.59 failed to converge. Thus, any results in the pilot study that were due to the ICC_x = .6 condition may not have been present in the larger study where these conditions were excluded.

This failure to converge have been because it was impossible to generate multilevel data with a positive definite matrix that also fulfilled all of these conditions simultaneously. The data that were generated for these cases should be analyzed further to determine why the multilevel models were unable to converge using this data, and the limits of data generation procedures in Mplus should be explored as well.

When examined at the replication level for a subset of the simulation data, the effect size factors (slope variance effect size and contextual effect size) often interacted with sample size on standard error accuracy outcomes. For example, larger slope variance effect sizes corresponded to higher power to detect the slope variance estimate, but this effect became less pronounced with increasing level-1 sample size. Increasing ICC_x also had effects on increasing coverage of the contextual effect, decreasing power to detect the slope variance, and decreasing contextual effect Type I error.

I also investigated the effect of autocorrelation on estimation of the contextual effect in a multilevel modeling framework. Many daily diary data sets include serially dependent variables, and it was of interest to see how much autocorrelation would affect estimation of the contextual effect. For the data generated to have autocorrelated residuals (ρ = .70), most parameter recovery was generally acceptable across three models fit to the data: two models were contextual latent growth models (one that accounted for autocorrelation, one that did not), and the third model was the contextual multilevel model. However, power to detect the contextual effect remained under 32% for all models. Also, the contextual multilevel model had the poorest outcomes, with over 110% standardized bias of the contextual effect and 78.5% coverage of the contextual effect. Note that the latent growth models may have performed better with this simulated data, as it was generated within a latent growth, rather than multilevel, framework.

The contextual multilevel model was applied to a daily diary data set from 1,079 persons measured each day over 8 days. The effect of average severity of physical symptoms on average stressor negative affect, which was evaluated for person mean centered data with the means included in the model. Both between-person and within-person effects, as well as the contextual effect were significant. This indicated that the between-person relation of stressor negative affect and severity of physical symptoms was significantly larger than the within-person relation of stressor negative affect and severity of physical symptoms. The slope variance estimate was not significant, indicating that there were not significant differences in the within-person relation of negative affect and severity of physical symptoms.

Recommendations

The results of this study provide several recommendations to applied researchers. Note that these recommendations are offered based on the conditions included in this study. It is always good practice for a researcher planning a study to conduct a power analysis to determine the sample size needed to detect any effects. Such power analyses can easily be performed in the M*plus* Monte Carlo utility. Bolger, Stadler, and Laurenceau (2012) demonstrated how a Monte Carlo simulation could be used to perform a power analysis for a multilevel model with diary data.

1) If detecting a contextual effect is of interest, then studies should include at least i = 100 participants with at least t = 20 time points, or i = 150 participants with at least t = 5 time points in order to have at least 80% power to detect the contextual effect. Note that contextual effect power may be low with contextual effect sizes less than 0.45 (assuming the variables have a standard normal distribution).
2) For detecting slope variance, studies should include at least i = 150 participants with at least t = 5 time points, or i = 50 participants with at least t = 10 time points, or i = 30 participants with at least t = 20 time points.
3) Researchers who fit a person-mean centered multilevel model may find a non-significant between-person effect, but it is still important to estimate the contextual effect, which may be significant. All 25 sample size conditions in this

study had average power of less than 80% to detect the level-2 fixed effect, but there was at least 80% power to detect the contextual effect in over half of the sample size conditions. This indicated that there were several conditions where there was not adequate power to detect the level-2 fixed effect, but there was adequate power to detect the contextual effect.

4) The contextual multilevel model will likely provide biased estimates of the contextual effect if there is unaccounted autocorrelation. If the data contain autocorrelation, a better alternative would be either estimating the model in a latent growth framework that accounts for the autocorrelation, or using the method of Curran and Bauer (2011) where the autoregression is modeled, and then a contextual multilevel model is fit to the residuals.
5) It is useful to estimate the standardized contextual effect and compare it to the unstandardized contextual effect, especially if the ICC of predictor and outcome variables differ and the level-2 effect is non-zero. In this study, the difference between standardized and unstandardized contextual effects approached 0.4 in conditions with a large (positive or negative) level-2 effect size and $ICC_x = 0.5$. Although the standardized contextual effect is not a truly standardized parameter (i.e., it does not have a standard deviation of one), it can still be useful as an effect size that describes the difference among between- and within-person effects, accounting for differential ICCs of the predictor and outcome variables.

Limitations and Future Directions

The results of this simulation study produce many questions for future research. In particular, this study only examined three values for the ICC of the predictor variable:

ICC_x = .10, .20, and .50. Conditions with ICC_x = .60 were also generated, but there was a 100% rate of non-convergence in these conditions where the between-person effect was large (γ_{01} = -.59 or .59). This may indicate that such data combinations are problematic for estimating contextual multilevel models, although further research would be needed to determine exactly why these conditions failed to converge. Still, it is of interest to see how contextual multilevel models behave when the ICC of the predictor variable exceeds .50, that is, when the predictor variable contains more variability at level-2 than at level-1. Preliminary analyses that included the ICC_x = .60 simulation conditions indicated that ICC_x = .50 may have been a hinge point. For example, Type I error decreased with larger ICC_x values until ICC_x = .50, after which Type I error increased again. However, this finding is likely confounded with the high rate of missingness among ICC_x = .60 cells in this simulation. Further research including ICC_x values above and below .50 is needed to determine if this relationship holds.

Related to this, the present study did not vary the ICC of the outcome variable—it was set to be ICC_y = .2 across all conditions. Future studies should simultaneously vary the ICC of the predictor and outcome variables. In daily diary studies, both predictor and outcome variables are likely to have high ICC values (Bolger & Laurenceau, 2013). There may be high rates of non-convergence with data sets where both predictor and outcome variables have more variability at level-2 than at level-1.

This study included a random slope component that allowed for variation in the within-person relation of *X* and *Y*. In some cases, researchers may be interested in simply discovering and quantifying this variability, especially if they have no hypotheses about the sources of such variability. However in many cases, researchers are interested in uncovering the determinants of such variability, or even testing hypotheses about what might predict slope variability (Raudenbush & Liu, 2000; Bolger et al., 2003). Thus, future simulations studies should examine models where some variable, such as the between-person effect, predicts slope variability.

Another aspect of the current study that should be examined in future research is the intercept-slope covariance. In the present study, this parameter was generated to be zero in the population and constrained to zero when the models were fitted, in order to aid in convergence. However, it is likely that a non-zero intercept slope covariance may exist and have a substantively meaningful interpretation.

Although this study touched briefly on the effect of autocorrelation in estimating contextual multilevel models, it is important to also examine other trends and cycles that may occur in longitudinal data, such as daily or weekly cycling (Curran & Bauer, 2011; Liu & West, 2015). Ignoring trends and cycling in longitudinal data may lead to falsely detecting a significant *X*-*Y* relation (Liu & West, 2015). Future research should extend the present study, as well as the work of Liu and West (2015) to examine the effect of ignoring trends and cycling in daily diary data where there is a contextual effect and random slope variance.

Although a random slope estimate can capture some intraindividual variation in the *X*-*Y* relation over time, the model considered in this study assumed that every individual in the study had the same linear functional form of *X*-*Y* relation. Such an assumption may not be realistic. For example, while some people may have a linear relation between their daily stress and sugar consumption, others may have a quadratic or an exponential relation. Future statistical models that allow these intraindividual differences should be developed and tested.

Other potential research topics include evaluating the contextual multilevel models when there is missing data (often a problem with longitudinal studies), as well as imbalanced measurement designs and lagged relations, where the effect of X on Y may occur after some time lag.

Although this study was mainly concerned with using a multilevel manifest (MMC) approach to modeling the contextual effect, as seen in Equation 5, the multilevel latent covariate (MLC) model also has advantages as well. Whereas the MMC approach uses observed group means as a level-2 predictor, the MLC approach assumes there is some unreliability in the observed group means, and a latent group mean is instead used as a level-2 predictor.

Lüdtke et al. (2008) found that the MLC model had less biased estimates of the contextual effect than the MMC model examined in the present study. However, the MLC model produced contextual effect estimates with more variability than the MMC model. While the present study in many ways replicated and expanded this earlier simulation work, a key difference is that Lüdtke et al. (2008) generated data based on an MLC model, and the present study generated data based on an MMC model. Thus, it was of interest to see how well the MLC model performs under conditions examined here where the true population parameters are based on an MMC model.

In order to examine the relative performance of the two modeling strategies, a small subset of the total simulation conditions were selected and an MLC model was fitted to the generated data. The effect of ICC_x and contextual effect size on parameter outcomes for both MMC and MLC models on a total of six conditions. In general, the outcomes for the MLC model were worse than the corresponding MMC conditions—coverage and power for the contextual effect were lower with the MLC than the MMC model, and Type I error for the contextual effect was much higher with the MLC model. (Only power to detect the slope variance was comparable across models.) However, this discrepancy in accuracy of contextual effect estimates could be due to some property of the data generating mechanism in Mplus. Future work is needed to clearly explain the differences in the data generating mechanisms and why the MLC model might perform poorly when data is generated using the MMC model.

Conclusions

Many researchers need longitudinal data in order to test and develop psychological theories. Longitudinal data include information about both within-person relations and between-person relations, although models that separate these two relations have not been properly utilized in psychological research (Curran & Bauer, 2011). Any statistical model that fails to decompose effects into between-person and within-person components risks an ecological fallacy where the two effects are incorrectly assumed to be equal. As shown in the empirical example, a within-person effect between two variables (stressor negative affect and severity of physical symptoms) may be significantly different than the corresponding between-person effect among the same two variables.

The current study discussed how and why between-person and within-person relations often differ. A simulation study was used to assess how well contextual multilevel models could account for these differences, especially when the within-person relation differs among people. While estimates of the contextual effect and slope variance were, overall, unbiased, there was often inadequate power to detect these effects with sample sizes in typical daily diary studies. Still, the current study demonstrated how contextual multilevel models could be applied to longitudinal data and used to compare between- and within-person relations.

REFERENCES

Alker, H. R. (1969). A typology of ecological fallacies. In M. Dogan, & S. Rokkan (Eds.), *Quantitative ecological analysis in the social sciences* (pp. 69-86). Cambridge, MA: MIT Press.

Blakely, T. A., & Woodward, A. J. (2000). Ecological effects in multi-level studies. *Journal of Epidemiology and Community Health*, *54*(5), 367-374. doi:10.1136/jech.54.5.367

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*(1), 579-616. doi: 10.1146/annurev.psych.54.101601.145030

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford Press.

Bolger, N., & Schilling, E. A. (1991). Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. *Journal of Personality*, *59*(3), 355-386. doi: 10.1111/j.1467-6494.1991.tb00253.x

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*(22), 3127-3131.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330. doi: 10.1037/1082-989X.6.4.330

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*, 583. doi: 10.1146/annurev.psych.093008.100356

Curran, P. J., Lee, T. H., Howard, A. L., Lane, S. T., & MacCallum, R. C. (2012). Disaggregating within-person and between-person effects in multilevel and structural equation growth models. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 217-253).

Diez-Roux, A. V. (1998). Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health, 88*(2), 216-222. doi: 10.2105/AJPH.88.2.216

Diez-Roux, A. V. (2000). Multilevel analysis in public health research. *Annual Review of Public Health*, *21*(1), 171-192. doi: 10.1146/annurev.publhealth.21.1.171

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138. doi: 10.1037/1082-989X.12.2.121

Firebaugh, G. (2001). Ecological fallacy, statistics of. In N. J. Smelser & P. B Bates (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 14, pp. 4023-4026). Oxford, UK: Elsevier Science Ltd.

Firebaugh, G. (2009). Commentary: 'Is the social world flat? WS Robinson and the ecologic fallacy'. *International Journal of Epidemiology, 38*(2), 368-370. doi: doi: 10.1093/ije/dyn355

Gu, F., Preacher, K. J., & Ferrer, E. (2014). A state space modeling approach to mediation analysis. *Journal of Educational and Behavioral Statistics*, *39*(2), 117-143. doi: 10.3102/1076998614524823

Hertzog, C., Lindenberger, U., Ghisletta, P., & Oertzen, T. V. (2006). On the power of multivariate latent growth curve models to detect correlated change. P*sychological Methods*, *11*(3), 244. doi: 10.1037/1082-989X.11.3.244

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. New York, NY: Routledge.

Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development*, *6*(2-3), 97-120. doi: 10.1080/15427600902911189

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Idrovo, A. J. (2011). Three criteria for ecological fallacy. *Environmental Health Perspectives*, *119*(8), a332. doi: 10.1289/ehp.1103768

Iida, M., Shrout, P. E., Laurenceau, J.-P., & Bolger, N. (2012). Using diary methods in psychological research. In H. Cooper (Ed.) *APA Handbook of Research Methods in Psychology: Vol 1. Foundations, planning, measures, and psychometrics* (pp. 277-305). Washington, DC: American Psychological Association. doi: 10.1037/13619-016

Kreft, I. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*(1), 1-21. doi: 10.1207/s15327906mbr3001_1

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review*, *23*(4), 418-444. doi: 10.1177/0193841X9902300404

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods*, *12*(3), 418-435. doi: 10.1177/1094428107308984

Liu, Y., & West, S. G. (2015). Weekly cycles in daily report data: An overlooked problem. *Journal of Personality.* doi: 10.1111/jopy.12182

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*(4), 444-467. doi: 10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203-229. doi: 10.1037/a0012869

MacKinnon, D. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201-218. doi: 10.1207/s15366359mea0204_1

Molenaar, P. (2008). On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology*, *50*(1), 60-69. doi: 10.1002/dev.20262

Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*(2), 112-117. doi: 10.1111/j.1467-8721.2009.01619.x

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*(4), 371. doi: 10.1037/1082-989X.2.4.371

Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research. *Journal of Personnel Psychology*, *9*(2), 79-93. doi: http://dx.doi.org/10.1027/1866-5888/a000009

Pearce, N. (2000). The ecological fallacy strikes back. *Journal of Epidemiology and Community Health*, *54*(5), 326-327. doi:10.1136/jech.54.5.326

Rast, P., & Hofer, S. M. (2014). Longitudinal design considerations to optimize power to detect variances and covariances among rates of change: Simulation results based on actual longitudinal studies. *Psychological Methods*, *19*(1), 133. doi: 10.1037/a0034524

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

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199. doi: 10.1037/1082-989X.5.2.199

Riley, M. W. (1963). *Sociological research: A case approach.* San Diego, CA: Harcourt, Brace and World.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *International Journal of Epidemiology*, *38*(2), 337-341. doi: 10.1093/ije/dyn357

Ryff, C. D., & Almeida, D. M. (2004-2009). National Survey of Midlife in the United States (MIDUS II): Daily Stress Project. ICPSR26841-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-02-26. http://doi.org/10.3886/ICPSR26841.v1

Schwartz, S. (1994). The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences. *American Journal of Public Health*, *84*(5), 819-824. doi: 10.2105/AJPH.84.5.819

Seidman, G., Shrout, P. E., & Bolger, N. (2006). Why is enacted social support associated with increased distress? Using simulation to test two possible sources of spuriousness. *Personality and Social Psychology Bulletin*, *32*(1), 52-65. doi: 10.1177/0146167205279582

Selvin, H. C. (1958). Durkheim's Suicide and problems of empirical research. *American Journal of Sociology*, 607-619. Retrieved from: http://www.jstor.org/stable/2772991

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Subramanian, S. V., Jones, K., Kaddour, A., & Krieger, N. (2009). Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Journal of Epidemiology*, *38*(2), 342-360. doi: 10.1093/ije/dyn359

Tennen, H., Affleck, G., Armeli, S., & Carney, M. A. (2000). A daily process approach to coping: Linking theory, research, and practice. *American Psychologist*, *55*(6), 626. doi: 10.1037/0003-066X.55.6.626

Velicer, W. (2010, July). Applying idiographic research methods: Two examples. In *Proceedings of the 8th International Conference on Teaching Statistics*. Retrieved from: http://icots.info/icots/8/cd/pdfs/invited/ICOTS8_4F3_VELICER.pdf

Wang, L. P., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods*, *20*(1), 63. doi: 10.1037/met0000030

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality*, *59*(3), 609-662. doi: 10.1111/j.1467-6494.1991.tb00261.x

Yip, T., & Fuligni, A. J. (2002). Daily variation in ethnic identity, ethnic behaviors, and psychological well–being among American adolescents of Chinese descent. *Child Development*, *73*(5), 1557-1572. doi: 10.1111/1467-8624.00490

Table 1

Raw Bias of the Contextual Effect Estimate by Fixed Effect Sizes

Between-persons effect size γ_{01}	Within-persons effect size γ_{10}				
	γ_{10} = -.59	γ_{10} = -.14	γ_{10} = 0	γ_{10} = .14	γ_{10} = .59
γ_{01} = -.59	0.003	0.004	0.004	0.004	0.003
γ_{01} = -.14	0.004	0.004	0.004	0.004	0.004
γ_{01} = 0	0.004	0.004	0.004	0.004	0.004
γ_{01} = .14	0.004	0.004	0.004	0.004	0.004
γ_{01} = .59	0.003	0.004	0.004	0.004	0.003

Table 2

Raw Bias of the Contextual Effect Estimate by Level-1 Sample Size

No. (t) of level-1 units within each level-2 unit	Contextual effect raw bias
t = 5	0.001
t = 10	0.001
t = 20	0.003
t = 40	0.009
t = 80	0.006

Table 3

Coverage of Slope Variance and Contextual Effect Estimates by ICC$_x$ and Level-2 Sample Size

No. (i) of level-2 units	ICC$_x$		
	ICC$_x$ = .10	ICC$_x$ = .20	ICC$_x$ = .50
		Slope variance coverage	
i = 30	87.70%	87.82%	88.96%
i = 50	90.36%	90.55%	91.38%
i = 100	92.43%	92.44%	92.85%
i = 150	93.25%	93.24%	93.32%
i = 200	92.99%	92.97%	93.30%
		Contextual effect coverage	
i = 30	91.12%	91.49%	92.35%
i = 50	92.81%	92.89%	93.35%
i = 100	94.05%	93.99%	93.90%
i = 150	94.02%	93.93%	93.89%
i = 200	93.99%	93.92%	94.14%

Table 4

Power **[and Type I Error]** Rates for Between- and Within-Person Effects by Fixed Effect Size

Fixed effect size	Power **[and Type I error]**
Between-person effect	
γ_{01} = -.59	95.21%
γ_{01} = -.14	32.35%
γ_{01} = 0	**[6.83%]**
γ_{01} = .14	33.80%
γ_{01} = .59	95.53%
Within-person effect	
γ_{10} = -.59	99.99%
γ_{10} = -.14	79.54%
γ_{10} = 0	**[5.66%]**
γ_{10} = .14	79.05%
γ_{10} = .59	99.99%

Note. Power values are given in regular face; Type I error rates are given in bold in brackets.

Table 5

Power and Type I Error Rates for Within-Person and Contextual Effects by Slope Variance Effect Size

	Within-persons effect estimate		Contextual effect estimate	
Slope variance effect size τ^2_{11}	Power	Type I error	Power	Type I error
$\tau^2_{11} = .05$	92.73%	5.83%	77.51%	6.67%
$\tau^2_{11} = .10$	89.55%	5.62%	77.11%	6.67%
$\tau^2_{11} = .15$	86.66%	5.51%	76.72%	6.68%

Table 6

Power Rates of Fixed and Random Effects by Level-1 and Level-2 Sample Size

No. (i) of level-2 units	No. (t) of level-1 units within each level-2 unit				
	t = 5	t = 10	t = 20	t = 40	t = 80
			Between-persons effect estimate		
i = 30	46.58%	50.11%	52.48%	53.89%	53.81%
i = 50	52.95%	56.19%	58.46%	59.82%	60.03%
i = 100	61.60%	64.90%	66.81%	67.85%	68.05%
i = 150	66.89%	70.04%	71.96%	73.36%	73.53%
i = 200	70.57%	74.07%	76.31%	77.46%	77.80%
			Within-persons effect estimate		
i = 30	63.84%	70.29%	76.59%	81.13%	82.92%
i = 50	70.08%	78.78%	85.93%	89.96%	92.06%
i = 100	82.84%	92.36%	96.52%	98.17%	98.74%
i = 150	90.99%	97.30%	99.21%	99.64%	99.80%
i = 200	95.23%	99.01%	99.84%	99.94%	99.96%
			Contextual effect estimate		
i = 30	55.37%	61.09%	64.42%	66.17%	67.42%
i = 50	65.04%	69.57%	72.31%	73.95%	74.87%
i = 100	75.73%	79.21%	81.22%	82.10%	82.79%
i = 150	80.50%	83.43%	85.12%	86.09%	86.57%
i = 200	83.31%	86.07%	87.81%	88.58%	89.12%
			Slope variance estimate		
i = 30	21.76%	47.81%	77.72%	94.24%	99.10%
i = 50	36.35%	66.82%	90.98%	99.05%	99.97%
i = 100	59.50%	86.66%	98.44%	99.99%	100.00%
i = 150	72.35%	93.43%	99.63%	100.00%	100.00%
i = 200	79.61%	96.24%	99.93%	100.00%	100.00%

Note . Grey cells denote power for conditions with Type I error rates that exceed the 7.5% threshold.

Table 7

Type I Error Rates of Fixed Effects

No. (i) of level-2 units	No. (t) of level-1 units within each level-2 unit				
	t = 5	t = 10	t = 20	t = 40	t = 80
			Between-persons effect estimate		
i = 30	8.72%	10.12%	10.53%	9.77%	9.74%
i = 50	7.29%	6.56%	8.42%	7.94%	6.91%
i = 100	4.52%	5.10%	6.26%	6.42%	6.15%
i = 150	4.65%	5.32%	5.61%	6.30%	5.59%
i = 200	5.04%	5.67%	6.70%	6.26%	5.07%
			Within-persons effect estimate		
i = 30	7.56%	6.67%	5.48%	7.28%	6.26%
i = 50	4.49%	5.73%	5.05%	6.30%	6.02%
i = 100	4.85%	5.04%	4.67%	5.30%	5.39%
i = 150	5.38%	5.55%	5.27%	5.71%	6.19%
i = 200	5.63%	6.01%	5.40%	4.82%	5.36%
			Contextual effect estimate		
i = 30	7.46%	8.59%	8.51%	8.98%	8.15%
i = 50	6.95%	6.20%	6.82%	7.78%	7.22%
i = 100	5.20%	6.16%	6.18%	6.61%	5.98%
i = 150	6.08%	6.23%	5.61%	6.62%	5.73%
i = 200	5.63%	6.38%	5.74%	6.28%	5.77%

Table 8

Power and Type I Error Rates of the Contextual Effect by ICC_x

ICC_x	Power	Type I error
$ICC_x = .10$	68.18%	6.81%
$ICC_x = .20$	77.71%	6.75%
$ICC_x = .50$	85.46%	6.46%

Table 9

Power **[and Type I Error Rates]** for Slope Variance and Contextual Effect by ICC$_x$ and Level-2 Sample Size

No. (i) of level-2 units	ICC$_x$		
	ICC$_x$ = .10	ICC$_x$ = .20	ICC$_x$ = .50
	Slope variance power		
i = 30	75.20%	71.71%	57.46%
i = 50	84.84%	81.94%	69.13%
i = 100	93.39%	91.50%	81.86%
i = 150	96.33%	95.05%	87.86%
i = 200	97.70%	96.76%	91.01%
	Contextual effect power		
i = 30	49.11%	63.81%	75.76%
i = 50	59.79%	72.61%	81.03%
i = 100	72.73%	80.84%	87.06%
i = 150	78.12%	84.38%	90.53%
i = 200	81.12%	86.89%	92.91%
	Contextual effect Type I error		
i = 30	**8.88%**	**8.52%**	**7.62%**
i = 50	**7.23%**	**7.12%**	**6.63%**
i = 100	**5.95%**	**6.01%**	**6.11%**
i = 150	**5.99%**	**6.09%**	**6.08%**
i = 200	**6.02%**	**6.03%**	**5.85%**

Note. Power values are given in regular face; Type I error rates are given in bold in brackets. Grey cells denote power for conditions with unacceptable Type I error rates.

Table 10

Population Contextual Effect Sizes

Within-persons fixed effect	Between-persons fixed effect				
	γ_{01} = -.59	γ_{01} = -.14	γ_{01} = 0	γ_{01} = .14	γ_{01} = .59
γ_{10} = -.59	0	0.45	0.59	0.73	1.18
γ_{10} = -.14	-0.45	0	0.14	0.28	0.73
γ_{10} = 0	-0.59	-0.14	0	0.14	0.59
γ_{10} = .14	-0.73	-0.28	-0.14	0	0.45
γ_{10} = .59	-1.18	-0.73	-0.59	-0.45	0

Table 11

Coverage of the Slope Variance and Contextual Effect by Slope Variance Effect Size, ICCx, Level-2 Sample Size, Level-1 Sample Size, and Contextual Effect Size

		No. (i) of level-2 units							
No. (t) of level-1 units		i = 30				i = 50			
		ICC$_x$ = .10		ICC$_x$ = .50		ICC$_x$ = .10		ICC$_x$ = .50	
	γ_{01} - γ_{10}	τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15
		95% Coverage of the slope variance							
t = 5									
	0	94.6%	87.9%	97.1%	91.3%	94.0%	90.8%	98.9%	91.1%
	0.14	97.8%	89.2%	96.6%	93.5%	98.6%	90.5%	98.8%	93.9%
	0.28	97.7%	89.0%	96.6%	93.4%	98.6%	90.3%	98.8%	93.9%
	0.45	98.0%	88.9%	96.6%	93.9%	98.7%	89.8%	99.1%	93.7%
	0.59	98.0%	89.0%	96.6%	93.8%	98.7%	89.9%	99.1%	93.9%
	0.73	98.0%	88.9%	96.6%	93.9%	98.7%	89.8%	99.1%	93.7%
	1.18	94.6%	87.9%	97.1%	91.3%	94.0%	90.8%	98.9%	91.1%
t = 40									
	0	85.4%	85.6%	83.8%	85.8%	90.5%	90.5%	90.2%	90.4%
	0.14	84.5%	85.4%	83.9%	85.4%	90.1%	90.5%	90.3%	90.5%
	0.28	84.5%	85.4%	83.7%	85.3%	90.1%	90.5%	90.4%	90.4%
	0.45	84.5%	85.4%	84.0%	85.3%	90.1%	90.5%	89.9%	90.5%
	0.59	84.6%	85.4%	84.0%	85.3%	90.1%	90.4%	90.1%	90.5%
	0.73	84.5%	85.4%	84.0%	85.3%	90.1%	90.5%	89.9%	90.5%
	1.18	85.4%	85.6%	83.8%	85.8%	90.5%	90.5%	90.2%	90.4%
		95% Coverage of the contextual effect							
t = 5									
	0	91.6%	91.7%	92.8%	92.4%	92.7%	92.4%	94.7%	94.6%
	0.14	92.7%	91.8%	93.2%	93.1%	92.3%	92.7%	93.1%	93.4%
	0.28	92.6%	91.8%	93.2%	93.0%	92.5%	92.8%	93.1%	93.4%
	0.45	93.0%	91.9%	93.1%	92.5%	92.1%	92.9%	94.5%	94.4%
	0.59	92.9%	91.9%	93.1%	92.6%	92.1%	93.0%	94.5%	94.4%
	0.73	93.0%	91.9%	93.1%	92.5%	92.1%	92.9%	94.5%	94.4%
	1.18	91.6%	91.7%	92.8%	92.4%	92.7%	92.4%	94.7%	94.6%
t = 40									
	0	90.6%	90.5%	92.9%	93.0%	92.1%	91.5%	92.3%	93.3%
	0.14	90.9%	90.7%	91.1%	91.5%	91.9%	92.4%	92.3%	92.0%
	0.28	90.9%	90.7%	91.1%	91.5%	91.9%	92.4%	92.3%	92.0%
	0.45	91.0%	90.5%	92.6%	93.1%	92.1%	92.6%	92.7%	92.8%
	0.59	91.0%	90.5%	92.6%	93.2%	92.1%	92.6%	92.7%	92.8%
	0.73	91.0%	90.5%	92.6%	93.1%	92.1%	92.6%	92.7%	92.8%
	1.18	90.6%	90.5%	92.9%	93.0%	92.1%	91.5%	92.3%	93.3%

Table 12
Power to Detect the Slope Variance and Contextual Effect by Slope Variance Effect Size, ICCx, Level-2 Sample Size, Level-1 Sample Size, and Contextual Effect Size

No. (t) of level-1 units		No. (i) of level-2 units							
		i = 30				i = 50			
		ICC_x = .10		ICC_x = .50		ICC_x = .10		ICC_x = .50	
	γ_{01} - γ_{10}	τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15
		Power to detect the slope variance							
t = 5									
	0	13.8%	70.9%	3.0%	25.4%	25.4%	92.8%	7.1%	47.7%
	0.14	5.4%	40.0%	2.2%	15.0%	10.3%	64.8%	4.1%	27.9%
	0.28	5.4%	41.0%	2.2%	15.4%	10.6%	66.7%	4.2%	28.3%
	0.45	5.5%	41.6%	2.3%	16.2%	10.7%	67.1%	4.7%	32.1%
	0.59	5.5%	40.6%	2.3%	15.8%	10.4%	66.1%	4.5%	31.7%
	0.73	5.5%	41.6%	2.3%	16.2%	10.7%	67.1%	4.7%	32.1%
	1.18	13.8%	70.9%	3.0%	25.4%	25.4%	92.8%	7.1%	47.7%
t = 40									
	0	99.1%	100.0%	75.5%	99.9%	100.0%	100.0%	96.5%	100.0%
	0.14	91.0%	100.0%	60.1%	99.2%	99.6%	100.0%	88.5%	100.0%
	0.28	91.8%	100.0%	61.0%	99.2%	99.6%	100.0%	89.5%	100.0%
	0.45	91.8%	100.0%	62.5%	99.3%	99.6%	100.0%	89.4%	100.0%
	0.59	90.9%	100.0%	61.8%	99.3%	99.6%	100.0%	88.7%	100.0%
	0.73	91.8%	100.0%	62.5%	99.3%	99.6%	100.0%	89.4%	100.0%
	1.18	99.1%	100.0%	75.5%	99.9%	100.0%	100.0%	96.5%	100.0%
		Power to detect the contextual effect							
t = 5									
	0.14	10.8%	11.1%	13.5%	14.1%	11.4%	11.1%	18.0%	17.2%
	0.28	16.4%	16.9%	34.7%	32.2%	22.2%	21.7%	47.5%	44.5%
	0.45	29.8%	31.2%	77.2%	74.5%	43.5%	43.8%	93.9%	91.0%
	0.59	44.9%	45.6%	91.5%	91.4%	61.6%	61.3%	99.5%	99.4%
	0.73	60.0%	60.4%	97.6%	97.6%	78.8%	79.1%	99.9%	100.0%
	1.18	95.9%	96.1%	99.9%	100.0%	99.7%	100.0%	100.0%	100.0%
t = 40									
	0.14	14.4%	14.9%	28.1%	25.0%	15.6%	15.3%	33.0%	29.2%
	0.28	26.6%	26.1%	64.0%	57.5%	31.5%	30.9%	81.0%	74.8%
	0.45	50.7%	50.6%	99.9%	99.2%	65.7%	64.3%	100.0%	100.0%
	0.59	70.0%	68.8%	100.0%	100.0%	86.0%	84.8%	100.0%	100.0%
	0.73	82.5%	81.7%	100.0%	100.0%	95.0%	95.1%	100.0%	100.0%
	1.18	99.0%	98.8%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Note. Conditions where Type I error exceeded the 7.5% threshold are filled in grey.

Table 13

Contextual Effect Type I Error Rates by Slope Variance Effect Size, ICC$_x$, Level-2 Sample Size, and Level-1 Sample Size

No. (t) of level 1 units	No. (i) of level 2 units	ICC$_x$ = .10		ICC$_x$ = .50	
		τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15
t = 5					
	i = 30	8.4%	8.3%	7.2%	7.6%
	i = 50	7.3%	7.6%	5.3%	5.4%
t = 40					
	i = 30	9.4%	9.5%	7.1%	7.0%
	i = 50	8.0%	8.5%	7.7%	6.7%

Table 14

Standard Deviation of the Standardized Within-Person Effect by Slope Variance Effect Size, ICCx, Level-2 Sample Size, Level-1 Sample Size, and Contextual Effect Size

No. (t) of level-1 units		No. (i) of level-2 units							
		i = 30				i = 50			
		$ICC_x = .10$		$ICC_x = .50$		$ICC_x = .10$		$ICC_x = .50$	
	$\gamma_{01} - \gamma_{10}$	$\tau^2_{11} = .05$	$\tau^2_{11} = .15$	$\tau^2_{11} = .05$	$\tau^2_{11} = .15$	$\tau^2_{11} = .05$	$\tau^2_{11} = .15$	$\tau^2_{11} = .05$	$\tau^2_{11} = .15$
t = 5									
	0	0.299	0.308	0.335	0.468	0.181	0.182	0.187	0.178
	0.14	0.317	0.329	0.319	0.379	0.213	0.216	0.209	0.211
	0.28	0.313	0.320	0.300	0.308	0.213	0.216	0.208	0.211
	0.45	0.299	0.347	0.456	0.311	0.196	0.199	0.144	0.155
	0.59	0.296	0.336	0.294	0.322	0.196	0.200	0.141	0.160
	0.73	0.328	0.316	0.341	0.393	0.195	0.199	0.141	0.165
	1.18	0.306	0.358	0.492	0.457	0.175	0.179	0.144	0.154
t = 40									
	0	0.169	0.174	0.058	0.070	0.129	0.133	0.044	0.053
	0.14	0.204	0.214	0.195	0.201	0.158	0.165	0.151	0.155
	0.28	0.203	0.213	0.194	0.201	0.158	0.164	0.151	0.155
	0.45	0.177	0.187	0.059	0.076	0.136	0.144	0.045	0.058
	0.59	0.177	0.188	0.060	0.077	0.136	0.144	0.045	0.059
	0.73	0.176	0.187	0.059	0.076	0.136	0.144	0.045	0.058
	1.18	0.169	0.174	0.057	0.070	0.130	0.133	0.044	0.054

Table 15

Standard Deviation of the Standardized Between-Person Effect by Slope Variance Effect Size, ICCx, Level-2 Sample Size, Level-1 Sample Size, and Contextual Effect Size

		No. (i) of level-2 units							
		i = 30				i = 50			
No. (t) of level-1 units		ICC_x = .10		ICC_x = .50		ICC_x = .10		ICC_x = .50	
	γ_{01} - γ_{10}	τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15	τ^2_{11} = .05	τ^2_{11} = .15
t = 5									
	0	0.063	0.074	0.075	0.083	0.046	0.054	0.055	0.060
	0.14	0.102	0.117	0.098	0.107	0.074	0.085	0.071	0.078
	0.28	0.099	0.114	0.096	0.106	0.072	0.083	0.070	0.077
	0.45	0.099	0.114	0.093	0.103	0.072	0.083	0.068	0.075
	0.59	0.101	0.117	0.094	0.104	0.074	0.085	0.069	0.076
	0.73	0.099	0.114	0.093	0.103	0.072	0.083	0.068	0.075
	1.18	0.062	0.072	0.074	0.081	0.047	0.055	0.055	0.061
t = 40									
	0	0.033	0.052	0.035	0.051	0.026	0.041	0.027	0.039
	0.14	0.053	0.080	0.045	0.064	0.041	0.063	0.034	0.049
	0.28	0.052	0.079	0.044	0.063	0.040	0.061	0.034	0.049
	0.45	0.052	0.079	0.044	0.063	0.040	0.062	0.034	0.049
	0.59	0.053	0.080	0.045	0.064	0.041	0.063	0.034	0.049
	0.73	0.052	0.079	0.044	0.063	0.040	0.061	0.034	0.049
	1.18	0.033	0.051	0.035	0.050	0.025	0.040	0.027	0.039

Table 16

Standard Deviation of the Standardized Contextual Effect by Slope Variance Effect Size, ICCx, Level-2 Sample Size, Level-1 Sample Size, and Contextual Effect Size

No. (t) of level-1 units		No. (i) of level-2 units							
		i = 30				i = 50			
		$ICC_x = .10$		$ICC_x = .50$		$ICC_x = .10$		$ICC_x = .50$	
	$\gamma_{01} - \gamma_{10}$	$\tau^2_{11} = .05$	$\tau^2_{11} = .15$	$\tau^2_{11} = .05$	$\tau^2_{11} = .15$	$\tau^2_{11} = .05$	$\tau^2_{11} = .15$	$\tau^2_{11} = .05$	$\tau^2_{11} = .15$
t = 5									
	0	0.313	0.320	0.300	0.308	0.213	0.216	0.208	0.211
	0.14	0.317	0.329	0.319	0.379	0.213	0.216	0.209	0.211
	0.28	0.306	0.358	0.492	0.457	0.175	0.179	0.144	0.154
	0.45	0.328	0.316	0.341	0.393	0.195	0.199	0.141	0.165
	0.59	0.296	0.336	0.294	0.322	0.196	0.200	0.141	0.160
	0.73	0.299	0.347	0.456	0.311	0.196	0.199	0.144	0.155
	1.18	0.299	0.308	0.335	0.468	0.181	0.182	0.187	0.178
t = 40									
	0	0.203	0.213	0.194	0.201	0.158	0.164	0.151	0.155
	0.14	0.204	0.214	0.195	0.201	0.158	0.165	0.151	0.155
	0.28	0.169	0.174	0.057	0.070	0.130	0.133	0.044	0.054
	0.45	0.176	0.187	0.059	0.076	0.136	0.144	0.045	0.058
	0.59	0.177	0.188	0.060	0.077	0.136	0.144	0.045	0.059
	0.73	0.177	0.187	0.059	0.076	0.136	0.144	0.045	0.058
	1.18	0.169	0.174	0.058	0.070	0.129	0.133	0.044	0.053

Table 17

Mean and SD of Autocorrelation of X and Y by ICC of X

	ICC_x		
Autocorrelation	$ICC_x = .10$	$ICC_x = .20$	$ICC_x = .50$

ρ_y	M = 0.189
	SD = .084
		M = 0.189
		SD = .083
			M = 0.190
			SD = .083

ρ_x	M = .097
	SD = .074
		M = .195
		SD = .077
			M = .489
			SD = .072

Table 18

Standardized Bias, Mean Square Error (MSE), 95% Coverage, and Power for the Contextual Effect and Slope Variance Estimates for AR(1) Simulated Data Across Three Models

Model	Outcome variable	Contextual effect estimate	Slope variance estimate
AR(1) Latent Growth Model	Standardized Bias	6.8%	-11.3%
	MSE	0.134	0.002
	95% Coverage	93.9%	90.7%
	Power	27.4%	98.6%
Latent Growth Model with Uncorrelated Residuals	Standardized Bias	7.7%	-5.7%
	MSE	0.187	0.002
	95% Coverage	91.0%	94.0%
	Power	32.0%	90.0%
Multilevel Manifest Covariate Model	Standardized Bias	-110.7%	-11.1%
	MSE	0.074	0.033
	95% Coverage	78.5%	89.8%
	Power	29.0%	86.7%

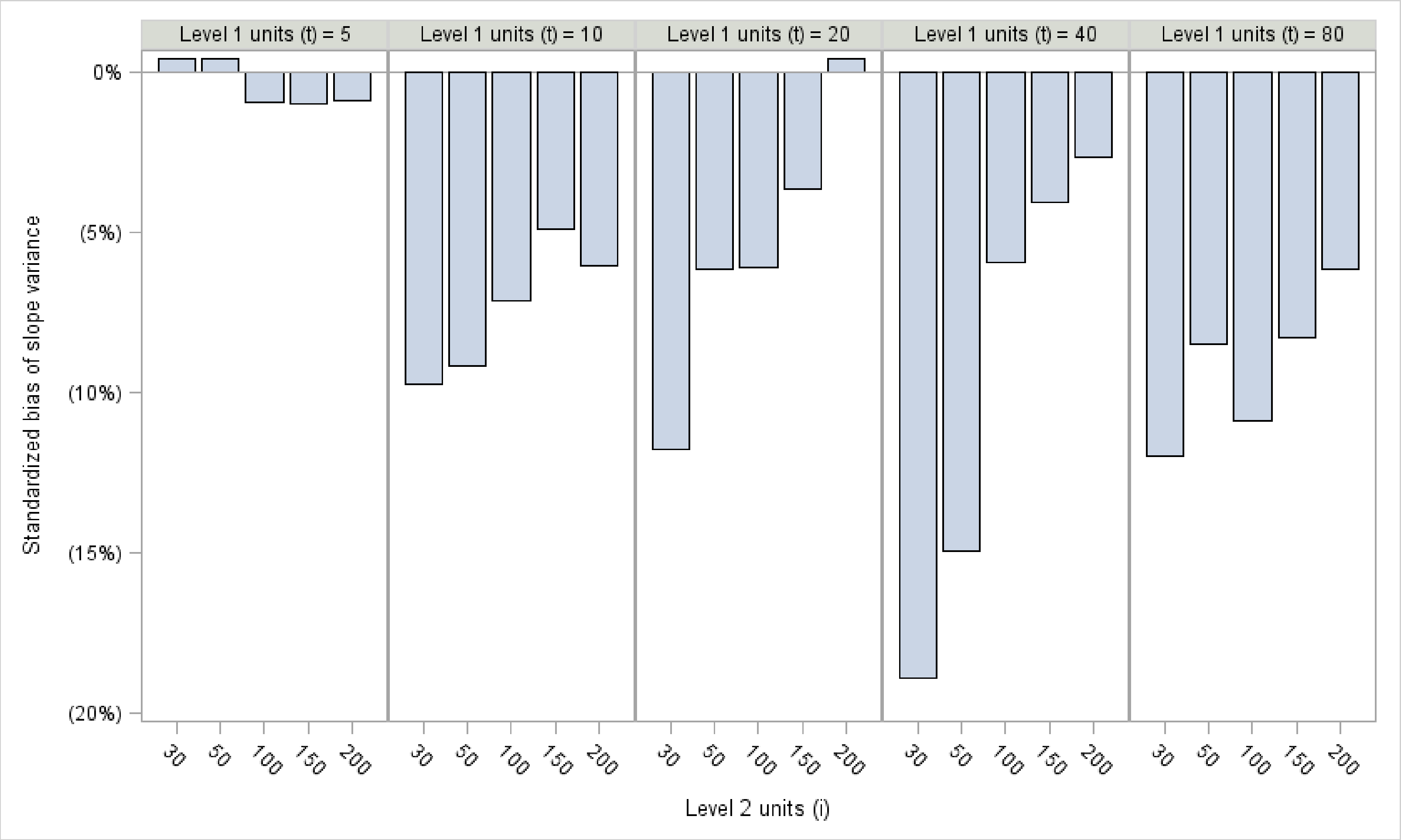

Figure 1. Average standardized bias of the slope variance estimate by level-2 sample size (*i*) and level-1 sample size (*t*).

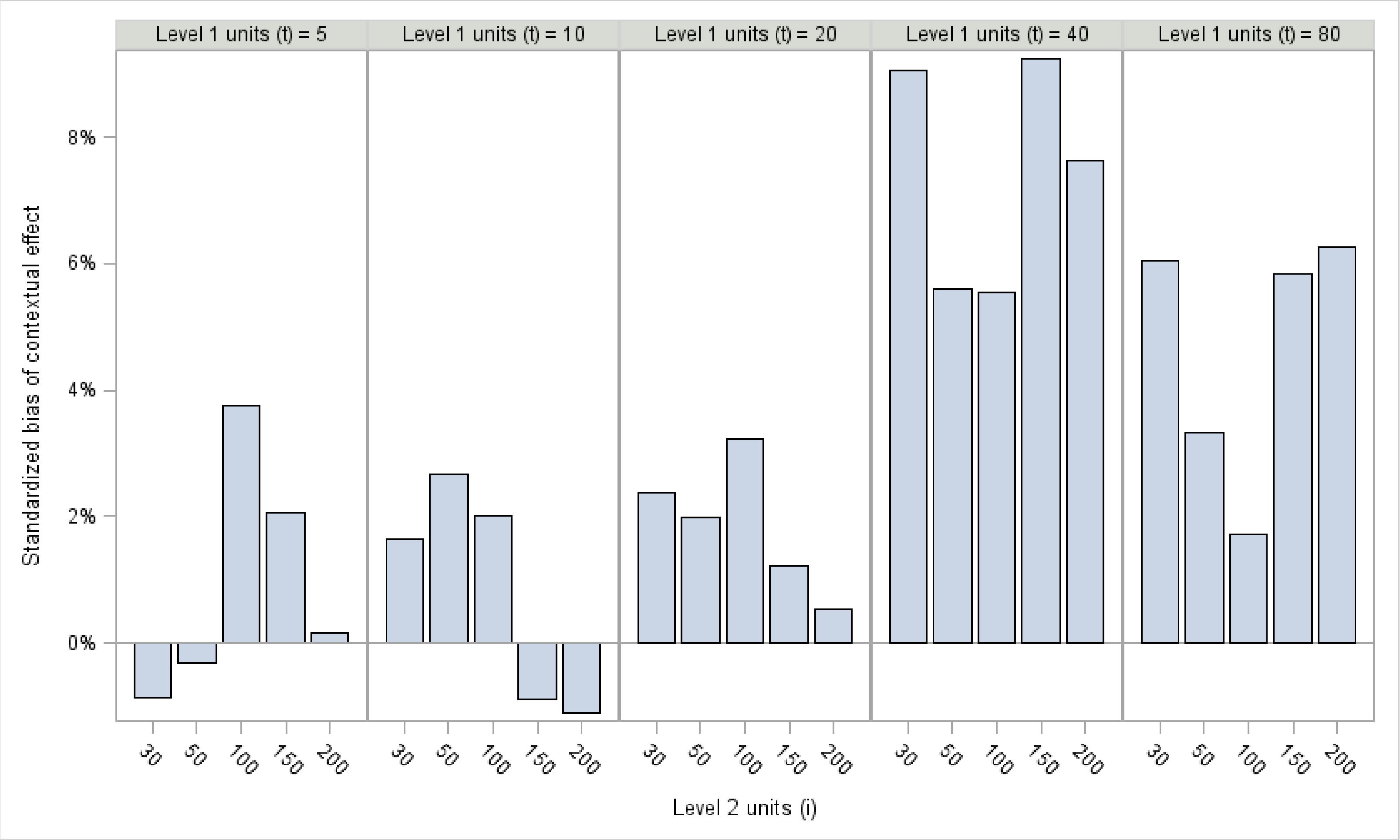

Figure 2. Average standardized bias of the contextual effect estimate by level-2 sample size (*i*) and level-1 sample size (*t*).

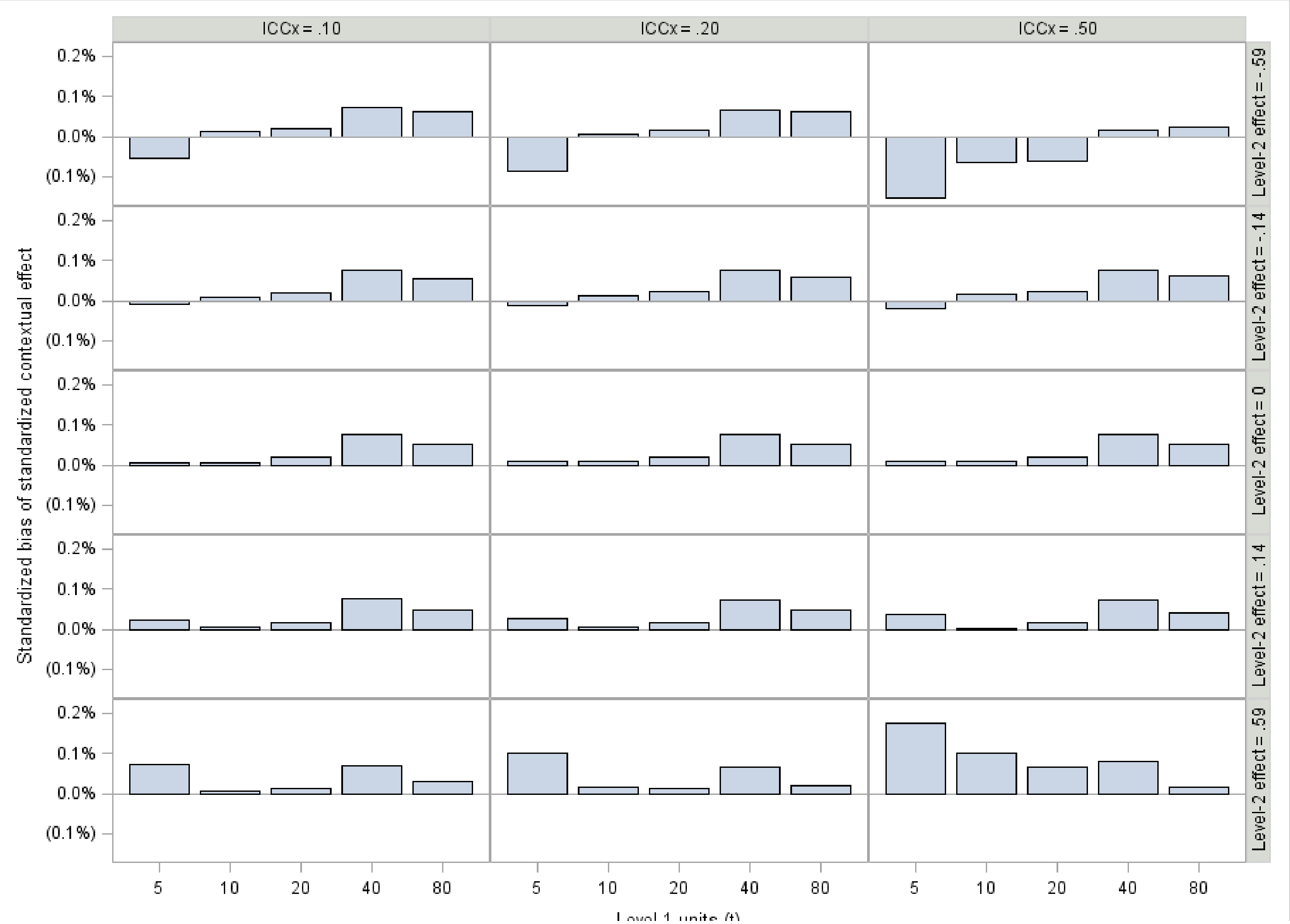

Figure 3. Average standardized bias of the standardized contextual effect estimate by level-2 effect size (γ_{01}) and level-1 sample size (t).

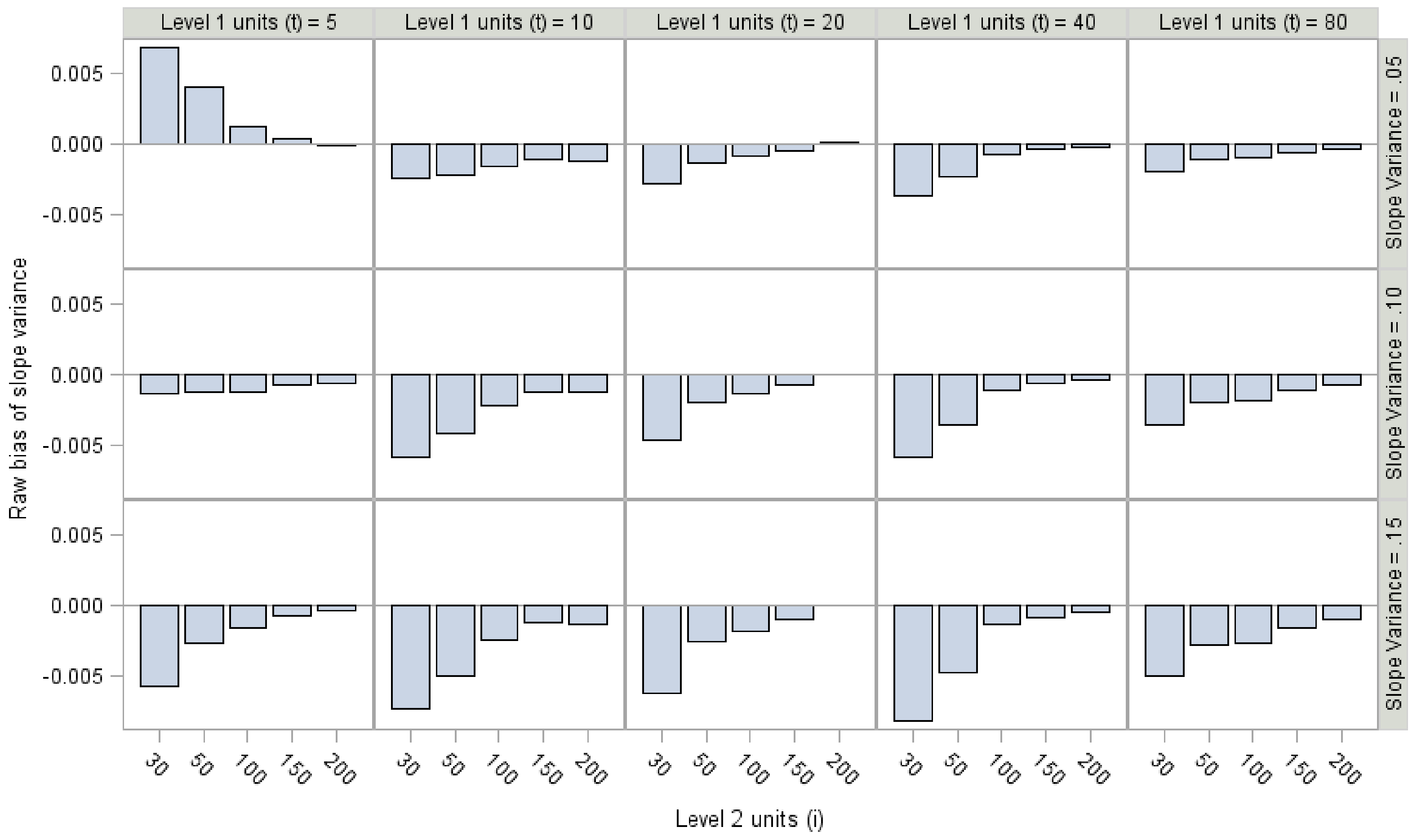

Figure 4. Average raw bias of the slope variance estimate by level-2 sample size (*i*), level-1 sample size (*t*), and slope variance effect size (τ^2_{11}).

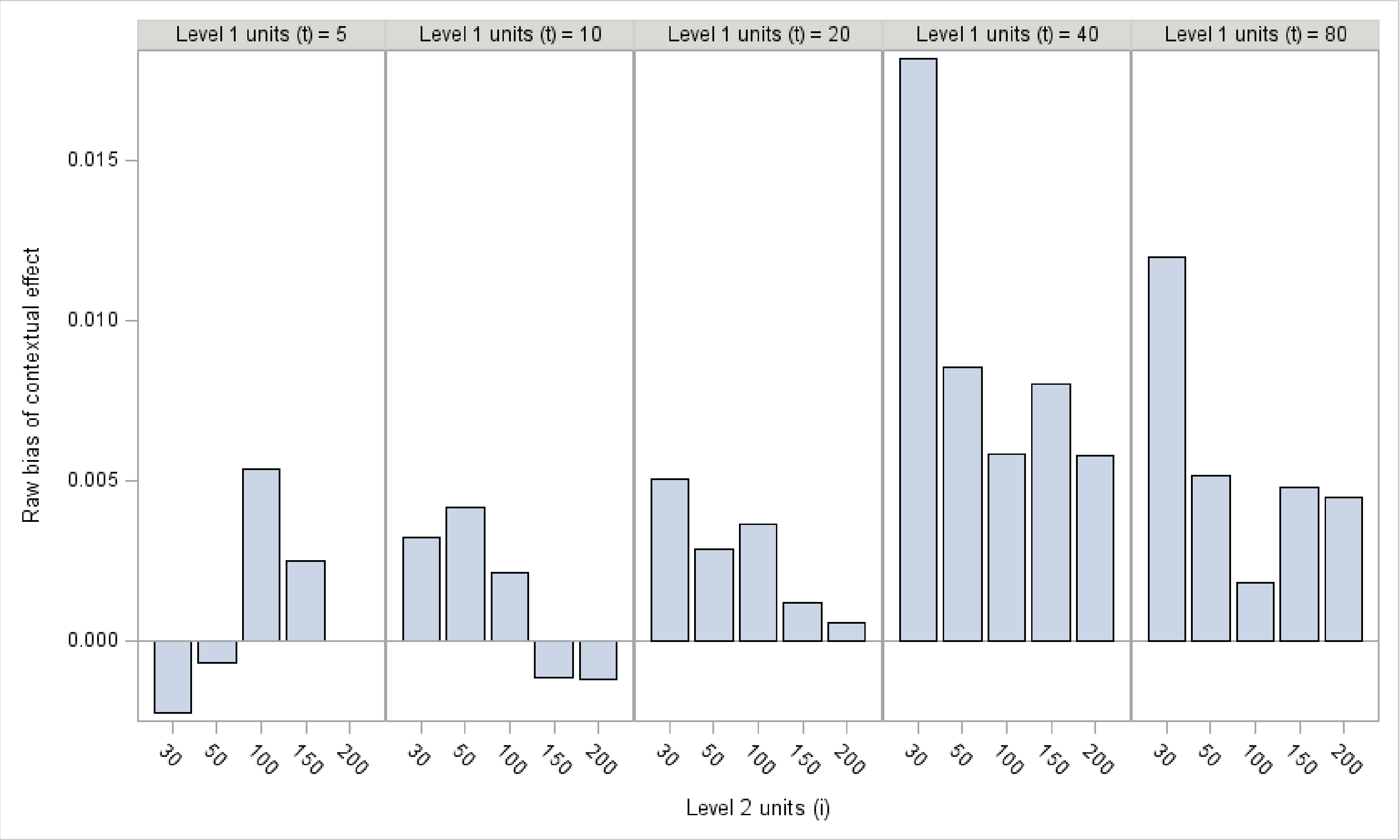

Figure 5. Average raw bias of the contextual effect estimate by level-2 sample size (*i*) and level-1 sample size (*t*).

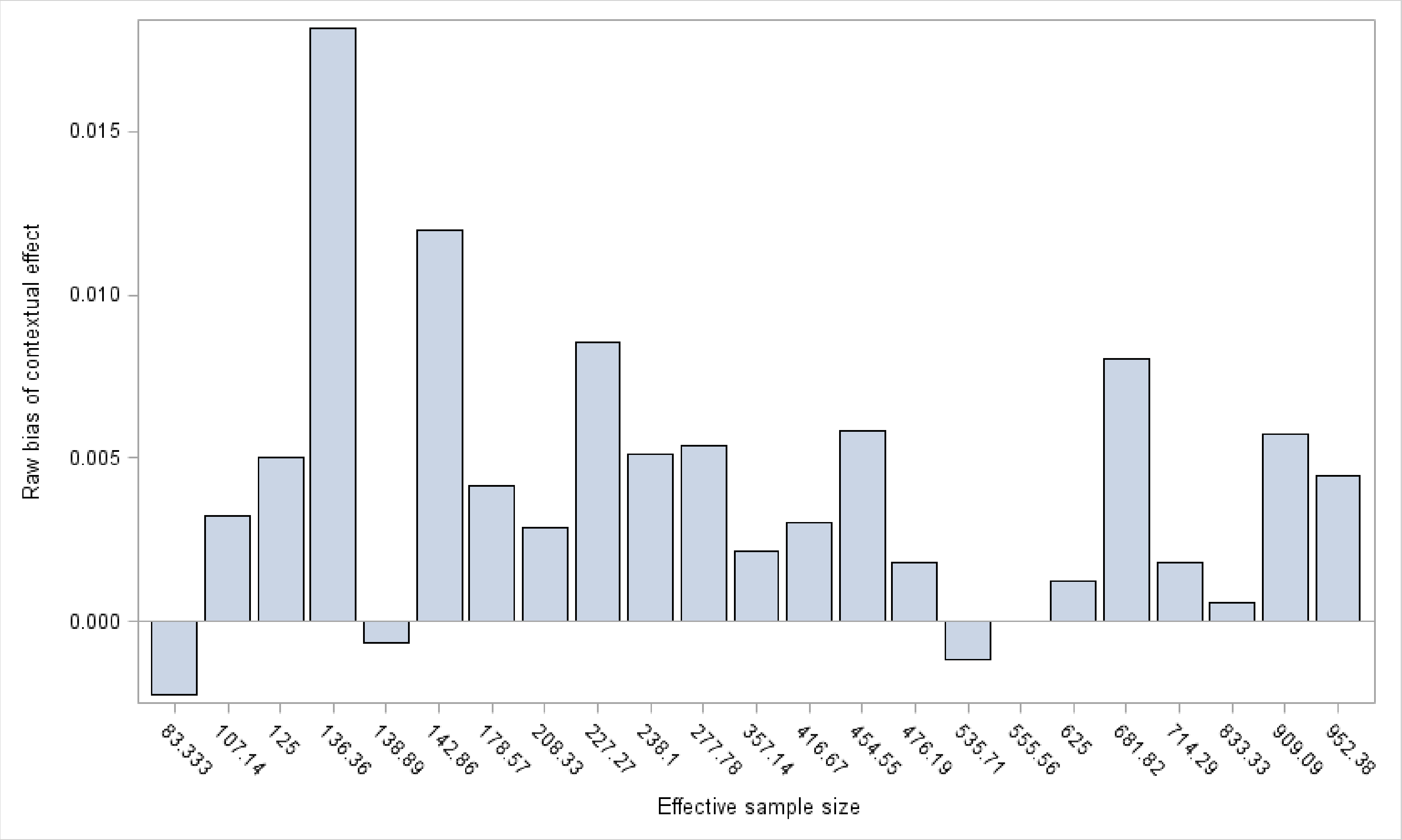

Figure 6. Average raw bias of the contextual effect by effective sample size.

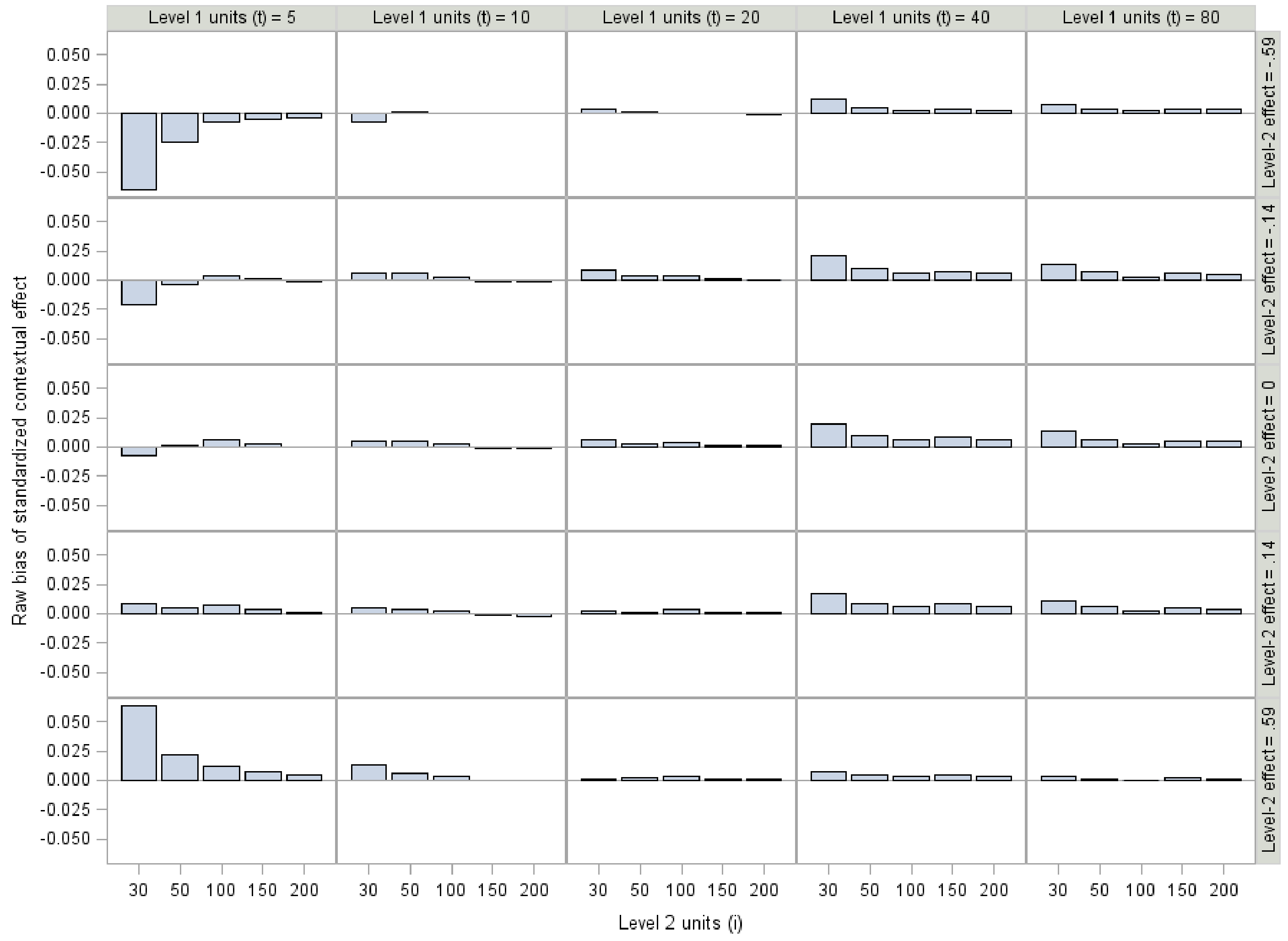

Figure 7. Average raw bias of the standardized contextual effect estimate by level-2 effect size (γ_{01}), level-2 sample size (*i*), and level-1 sample size (*t*).

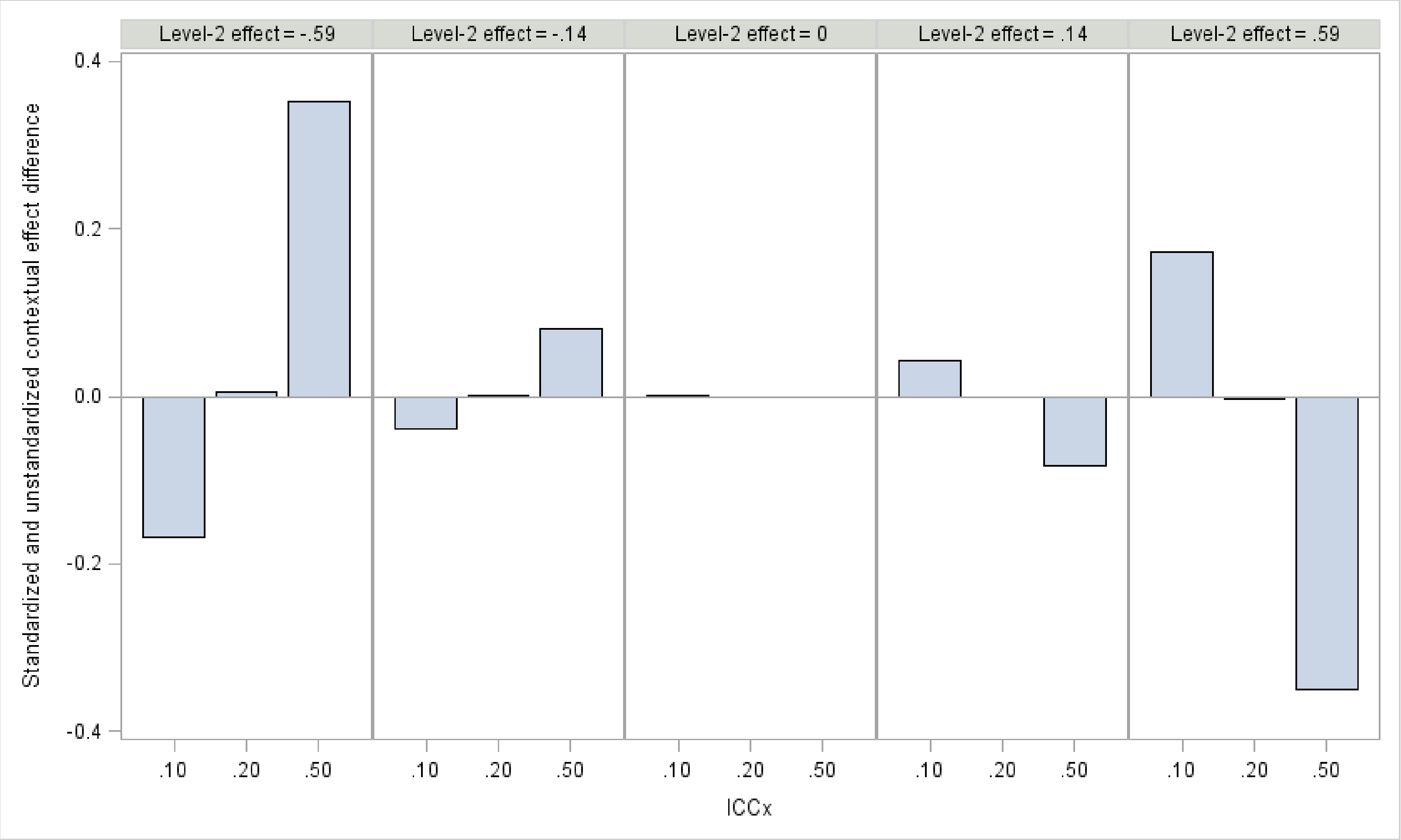

Figure 8. Average difference between unstandardized and standardized contextual effect by ICC of the predictor variable (ICC$_x$) and level-2 fixed effect size (γ_{01}).

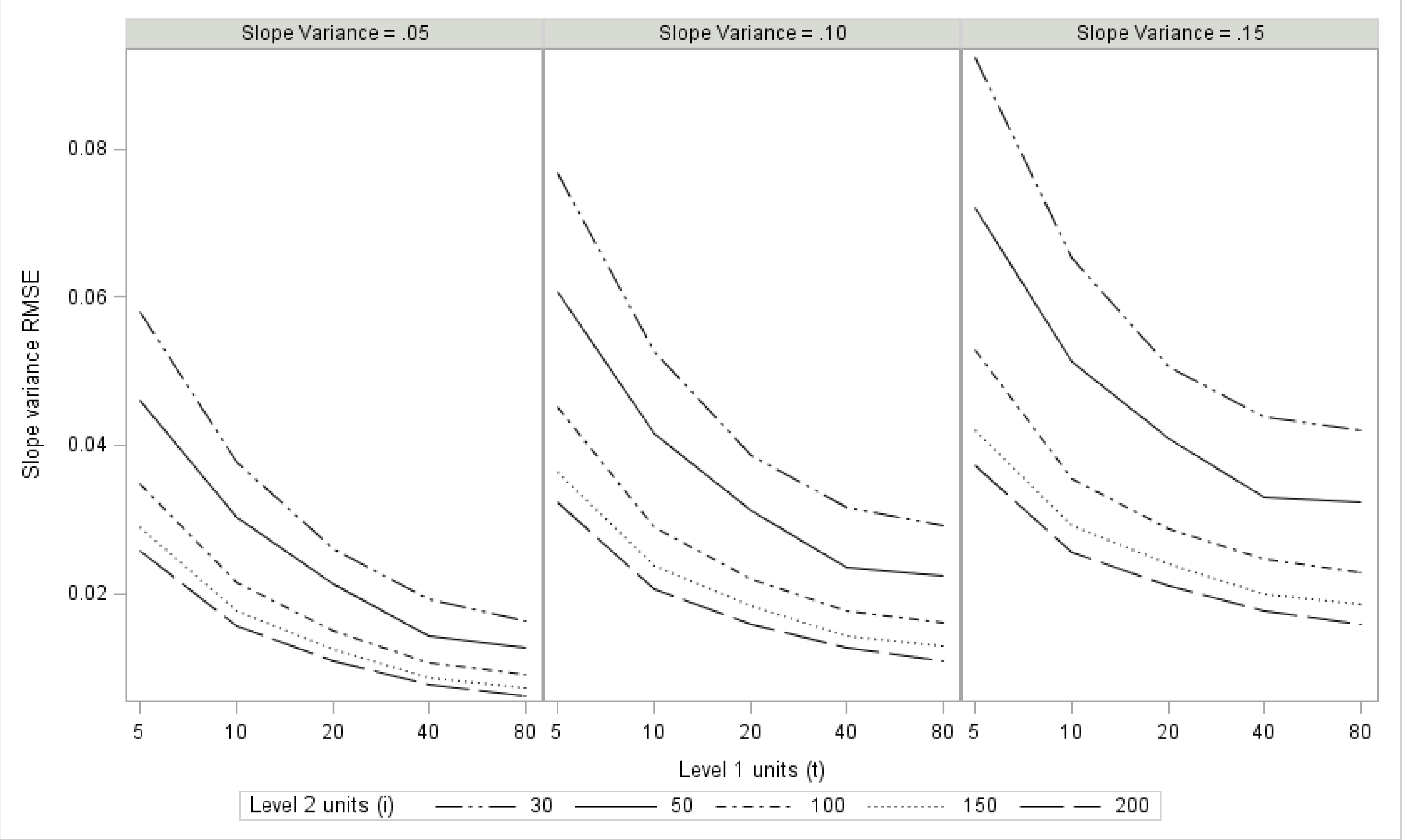

Figure 9. Average RMSE of the slope variance estimate by level-2 sample size (*i*), level-1 sample size (*t*), and slope variance effect size (τ^2_{11}).

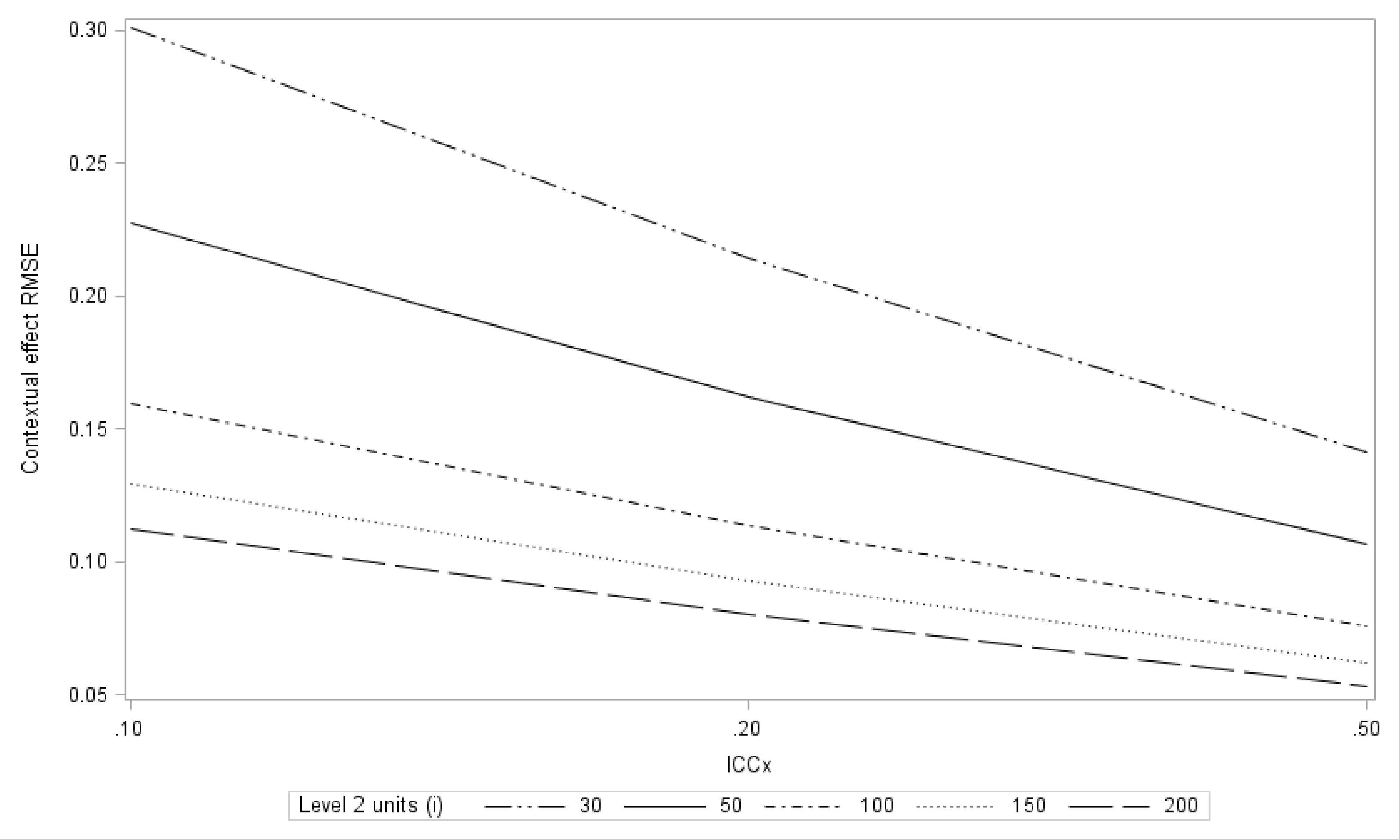

Figure 10. Average RMSE of the contextual effect estimate by level-2 sample size (*i*), and ICC of the predictor variable (ICC_x).

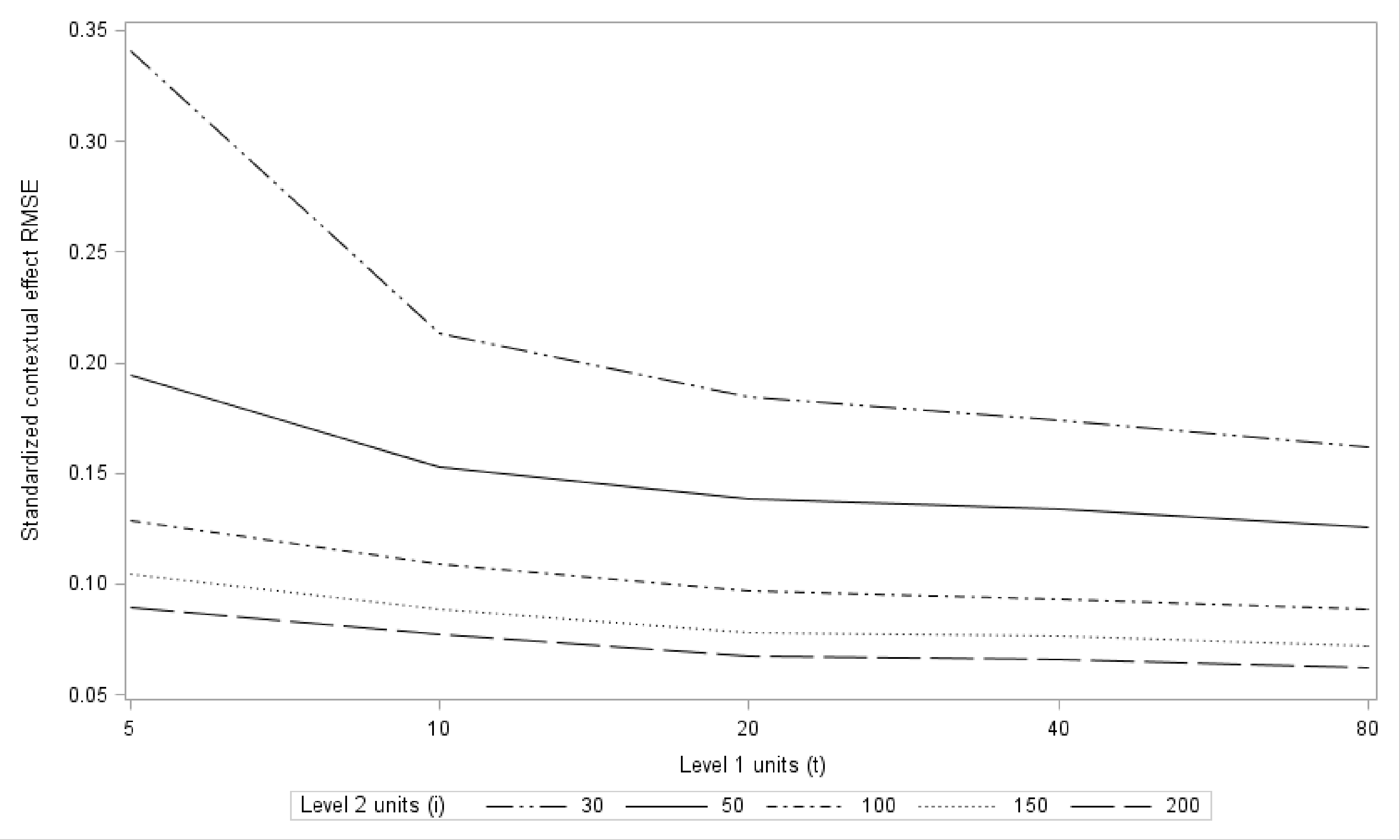

Figure 11. Average RMSE of the standardized contextual effect estimate by level-2 sample size (*i*) and level-1 sample size (*t*).

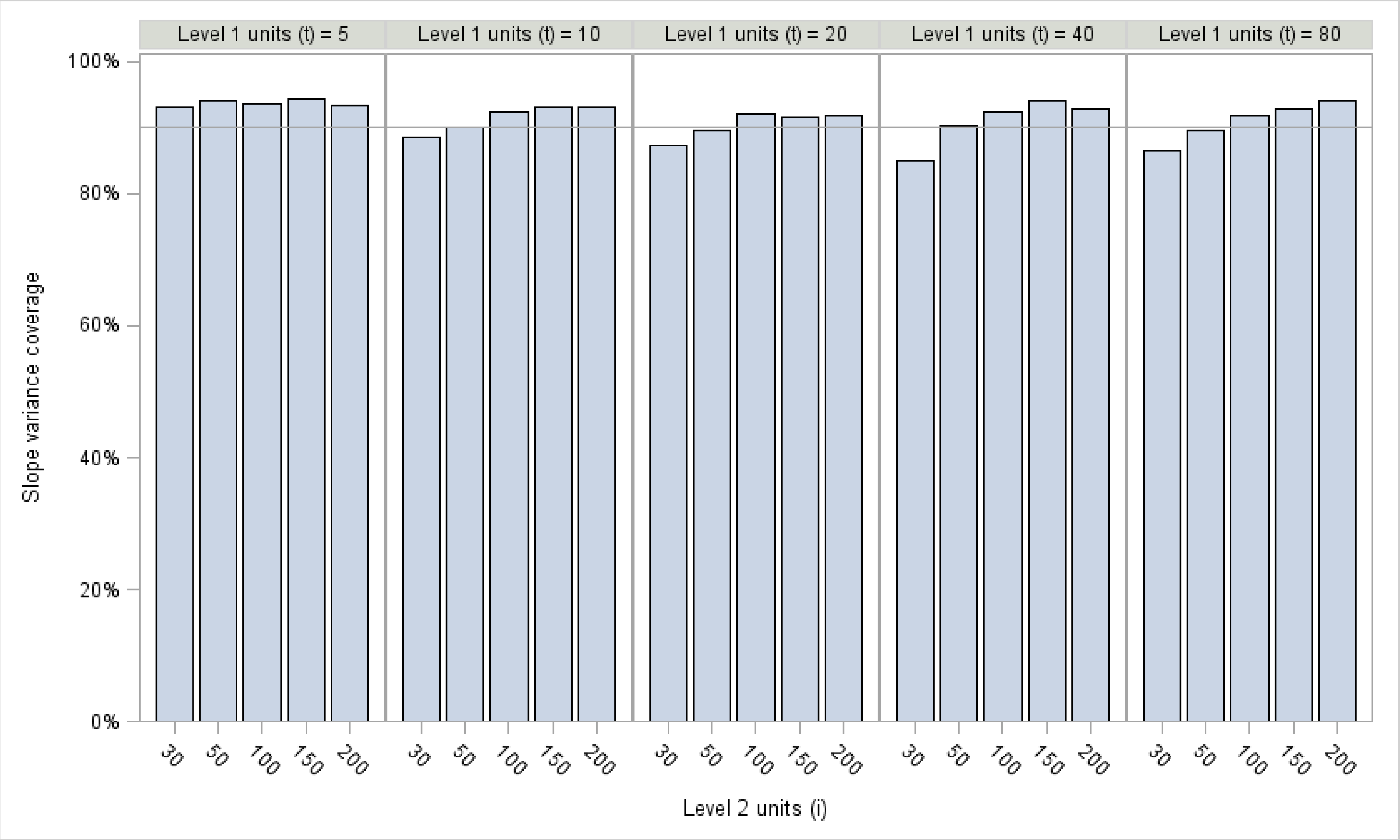

Figure 12. Average 95% coverage of the slope variance estimate by level-2 sample size (*i*) and level-1 sample size (*t*).

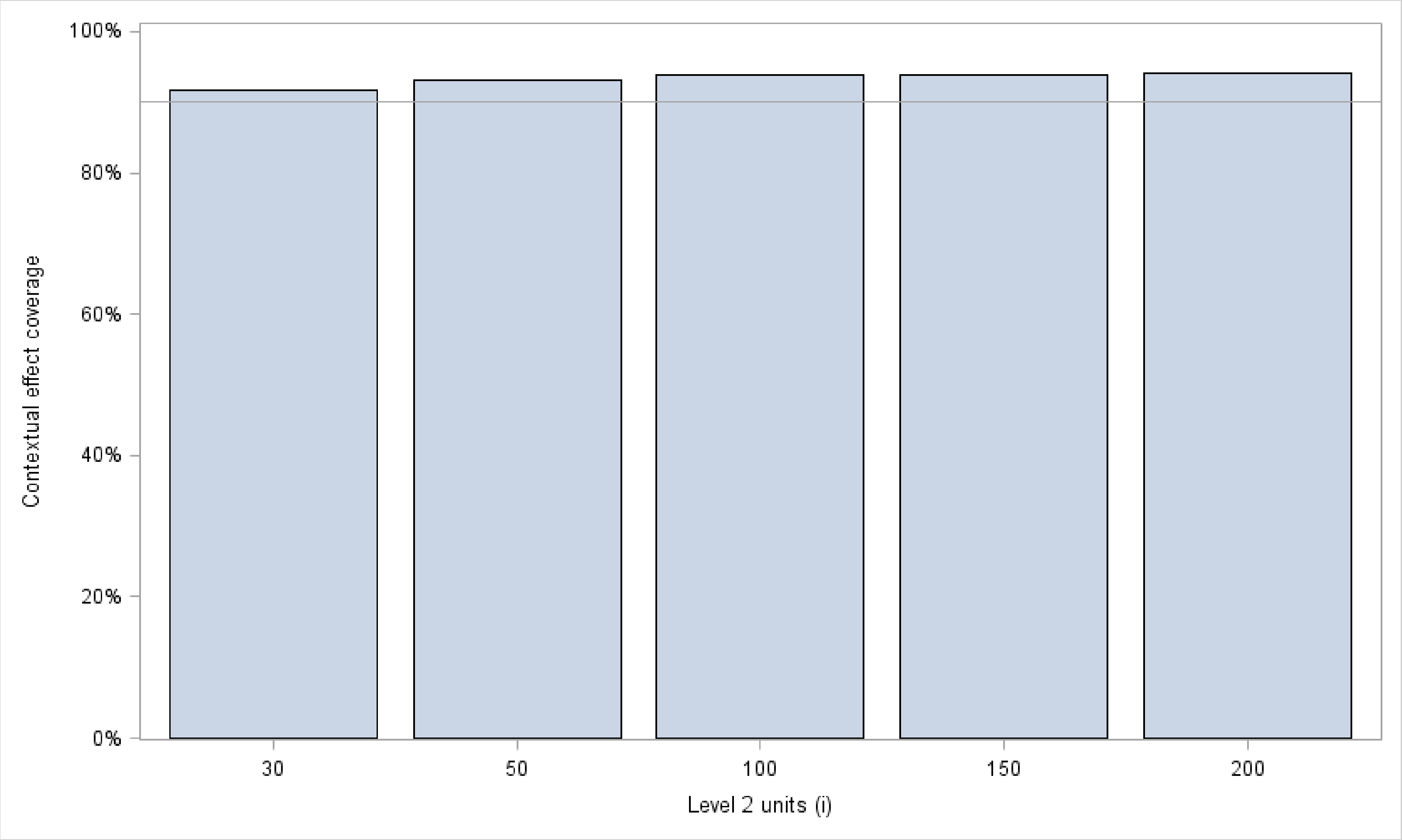

Figure 13. Average 95% coverage of the contextual effect estimate by level-2 sample size (*i*).

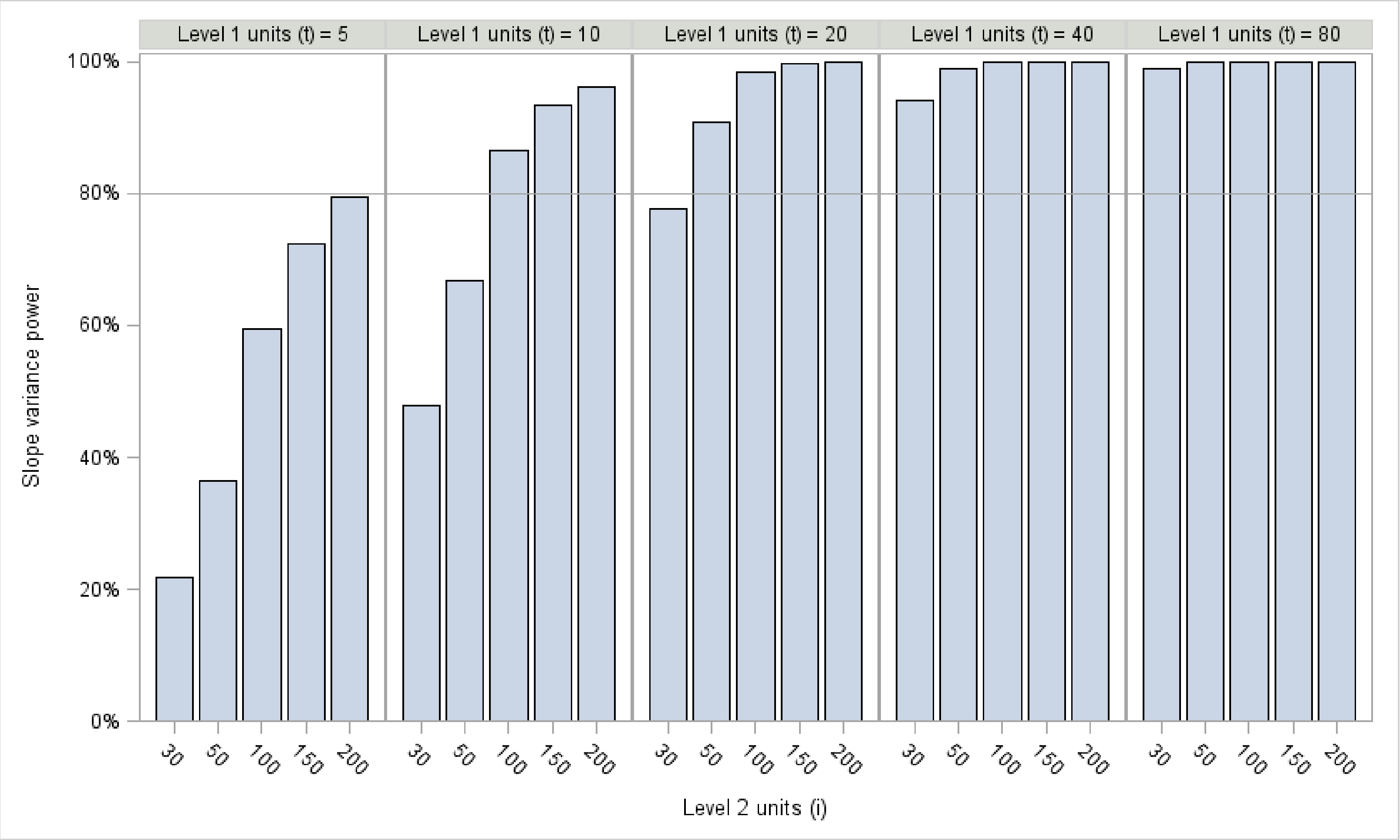

Figure 14. Average power to detect the slope variance estimate by level-2 sample size (*i*) and level-1 sample size (*t*).

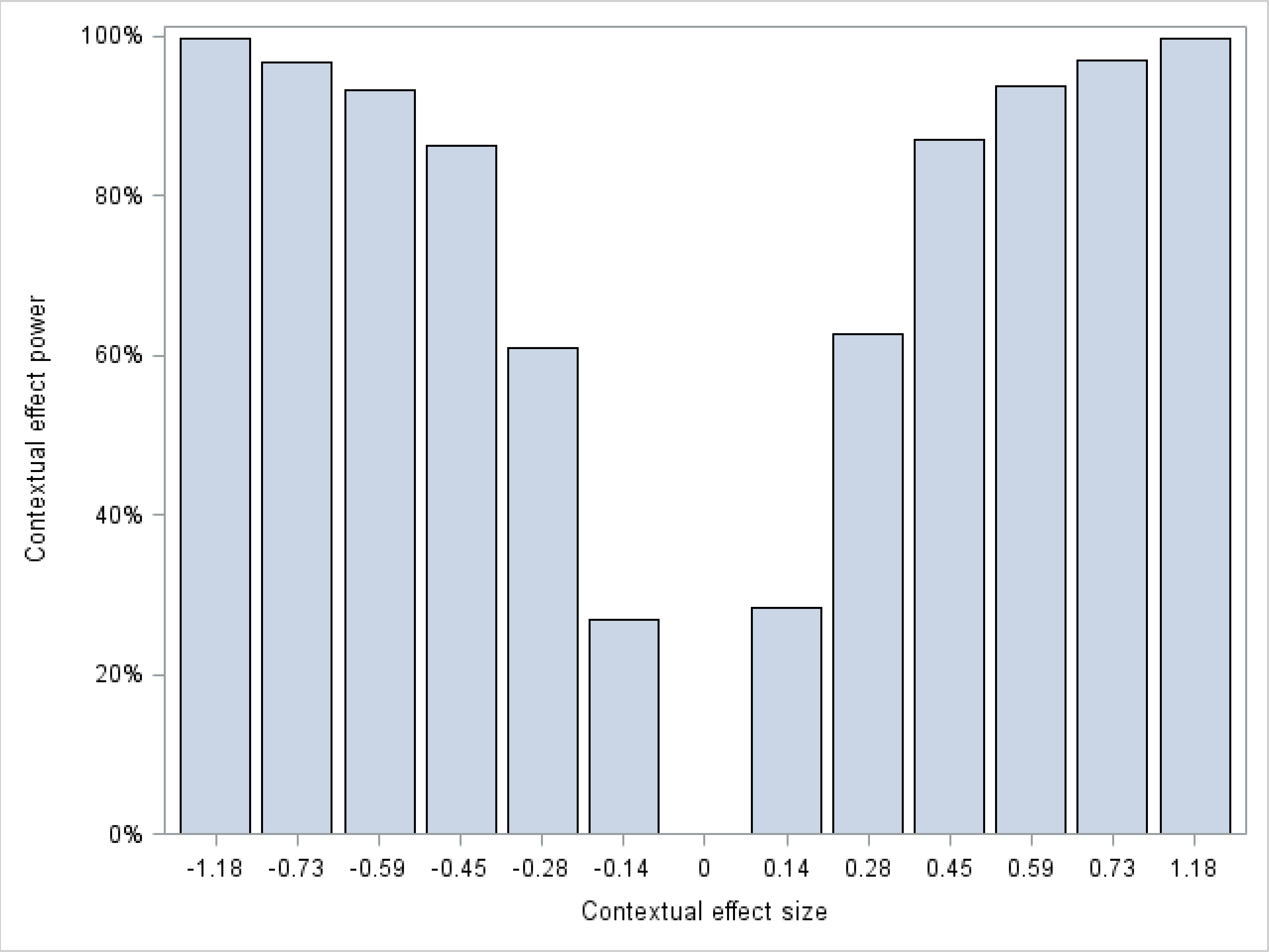

Figure 15. Average power to detect the contextual effect by contextual effect size (γ_{01}- γ_{10}).

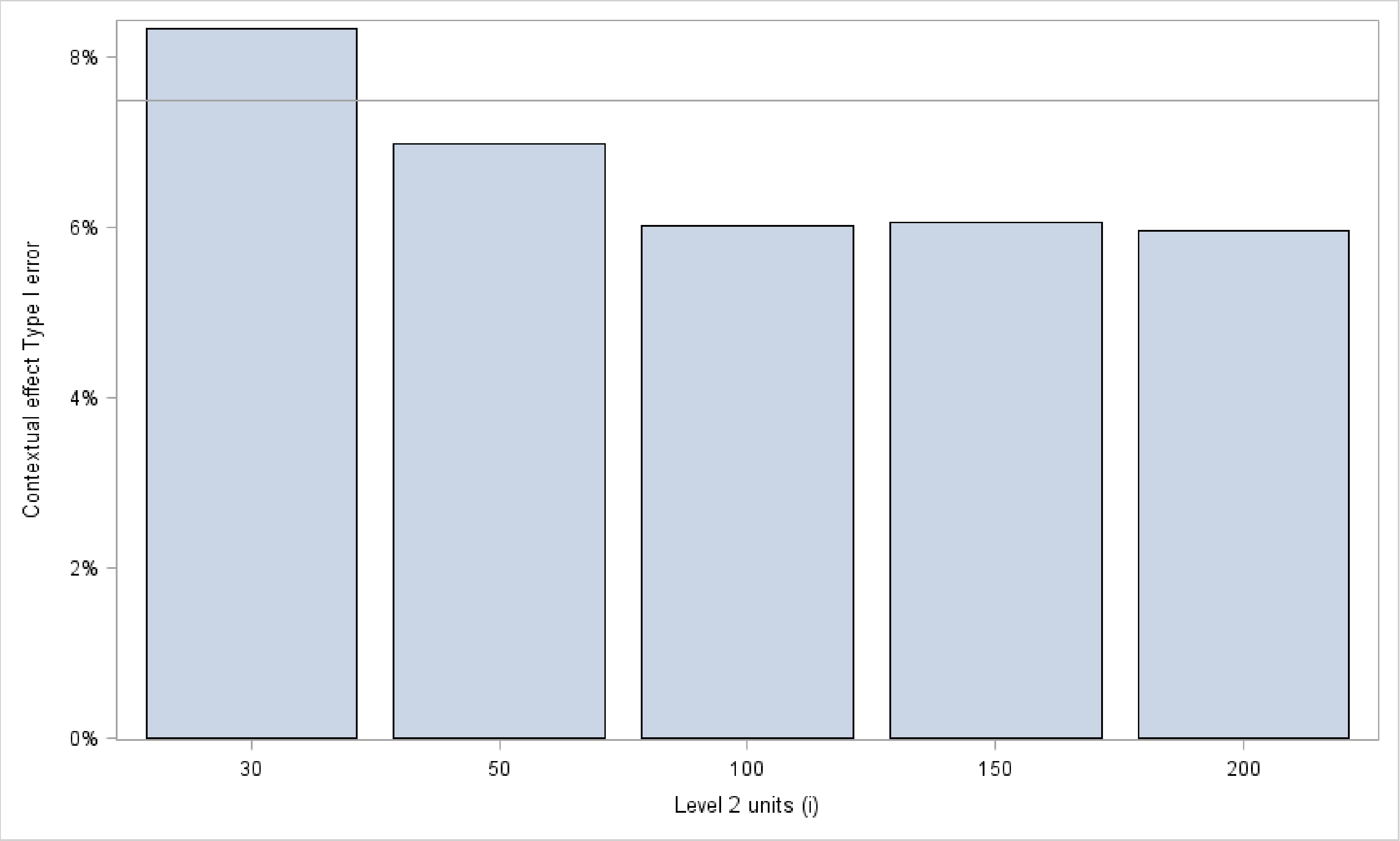

Figure 16. Average Type I error rate of the contextual effect by level-2 sample size (*i*).

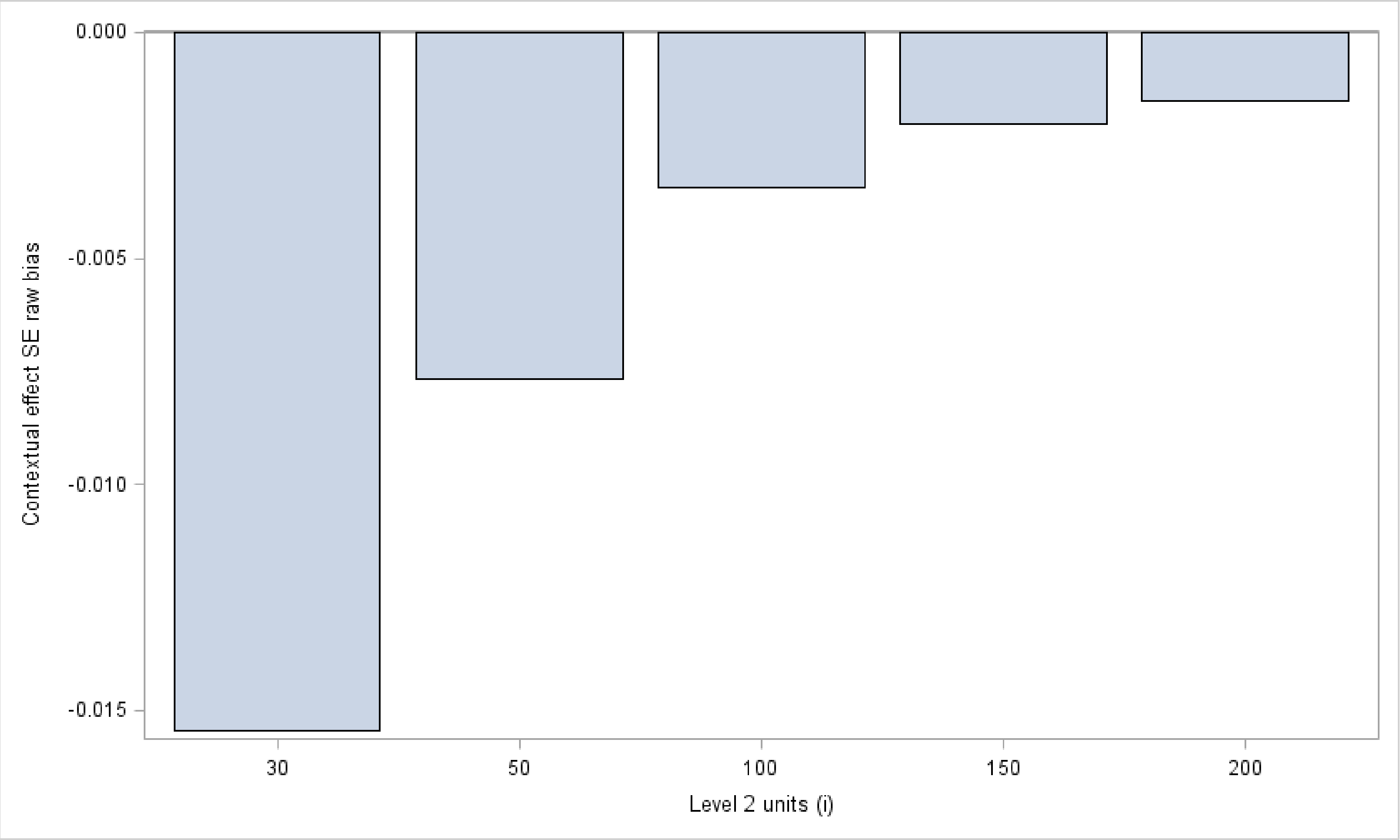

Figure 17. Average contextual effect standard error raw bias by level-2 sample size (*i*).

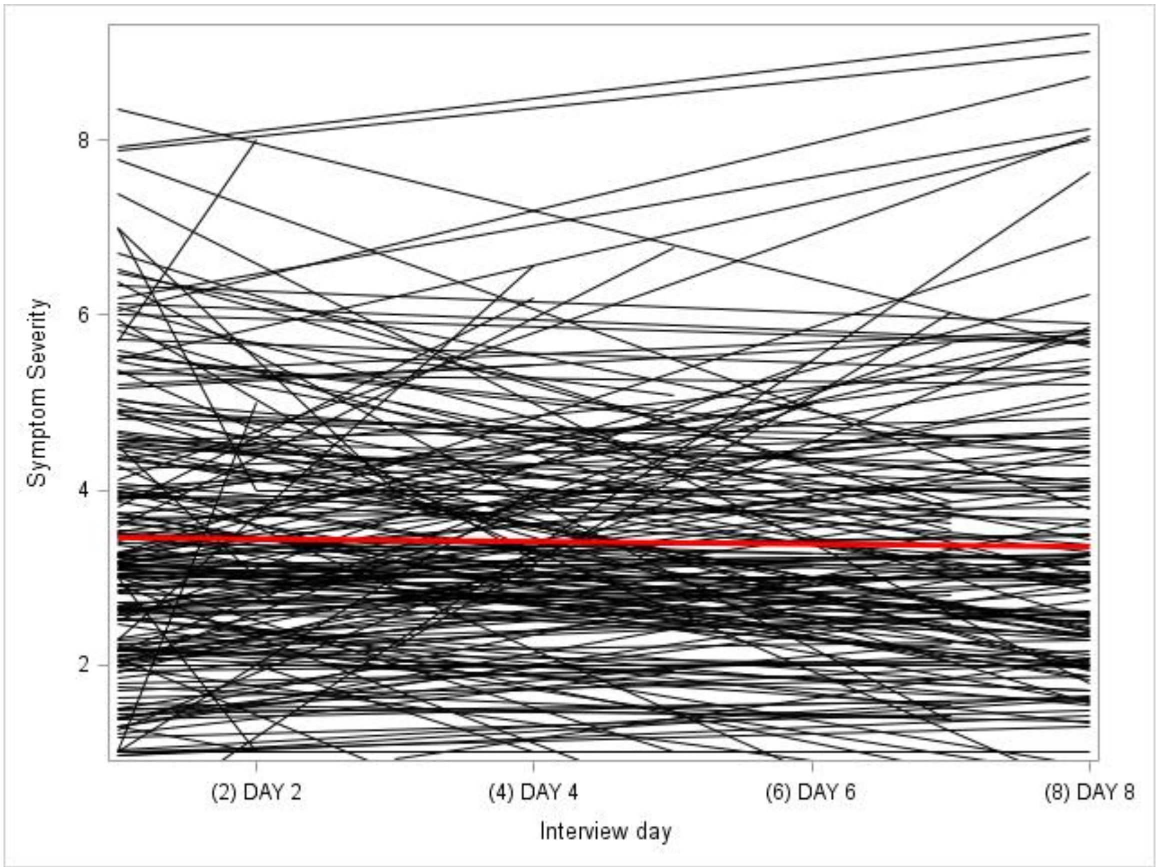

Figure 18. Spaghetti plot of physical symptom severity over time with overall spline curve.

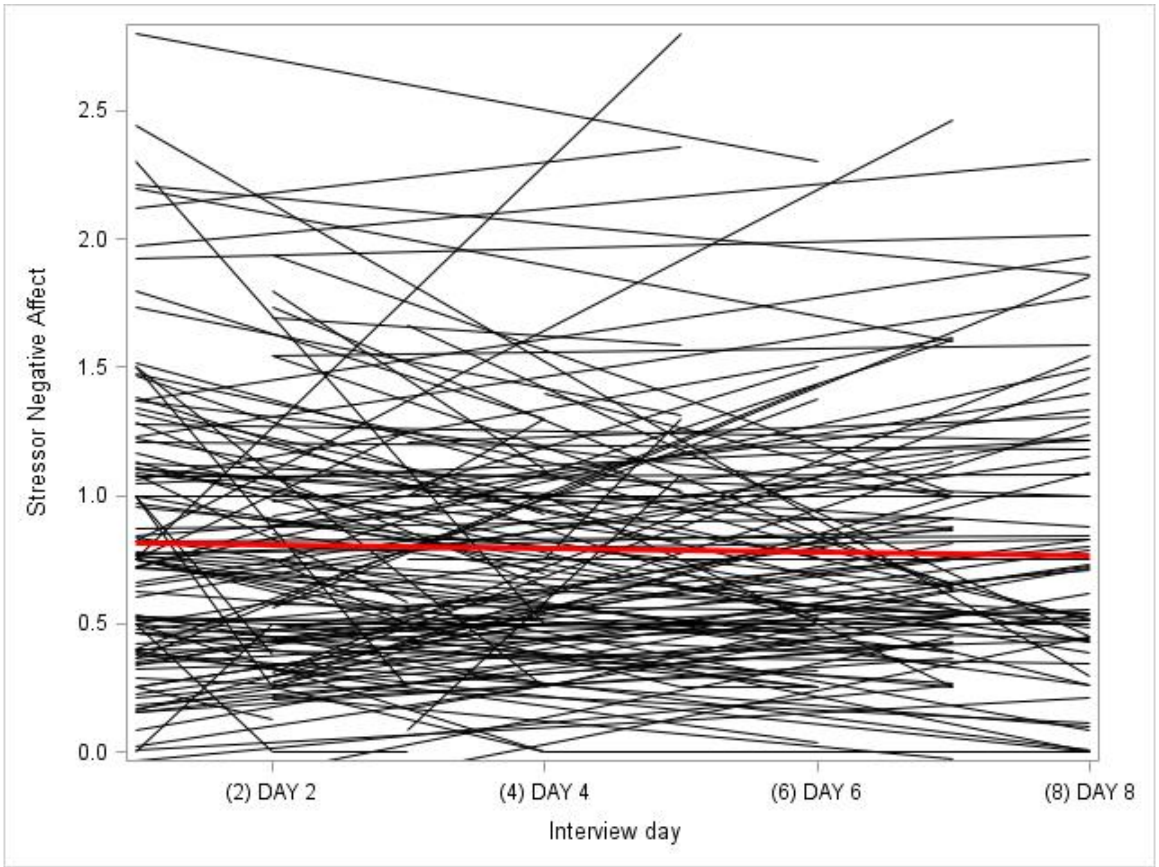

Figure 19. Spaghetti plot of stressor negative affect over time with overall spline curve.

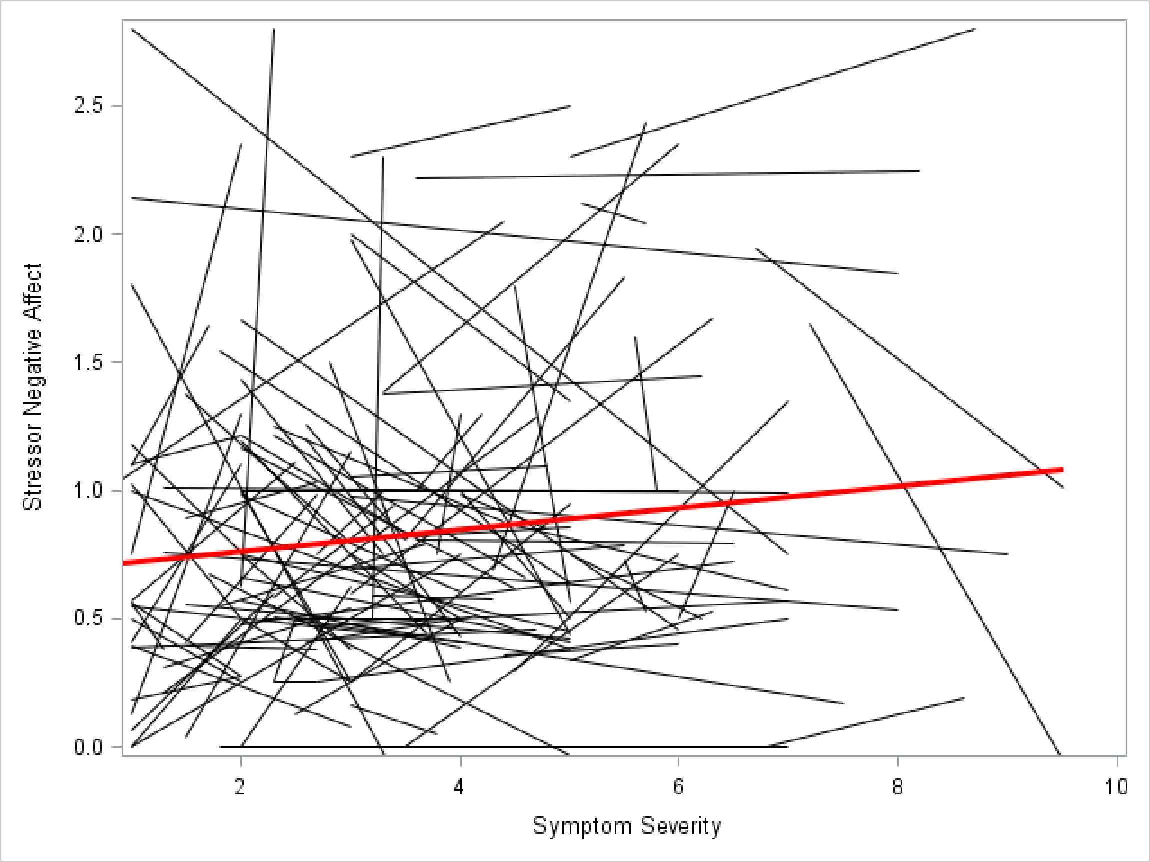

Figure 20. Individual and overall linear relation of physical symptom severity and stressor negative affect.

APPENDIX A

SYMBOL GLOSSARY

ICC$_x$	Intraclass correlation coefficient of X
ICC$_y$	Intraclass correlation coefficient of Y
r_{ti}	Within-persons (level-1) residual for person i at time t
$SD(X_B)$	Standard deviation of X at between-persons level (level-2)
$SD(X_W)$	Standard deviation of X at within-persons level (level-1)
$SD(Y_B)$	Standard deviation of Y at between-persons level (level-2)
$SD(Y_W)$	Standard deviation of Y at within-persons level (level-1)
T_{10}	Vector of intercept-slope covariances
T_{11}	Covariance matrix of random slopes
u_{0i}	Between-person (level-2) residual for person i
u_{1i}	Between-person (level-2) slope residual for person i
$VAR(X_B)$	Variance of X at between-persons level (level-2)
$VAR(X_W)$	Variance of X at within-persons level (level-1)
X_B	Portion of X that is purely between-persons (level-2)
X_{ti}	Value of predictor variable X for person i at time t

X_W	Portion of X that is purely within-persons (level-1)
$\bar{X}_i$	Mean of X for person *i*
$X_{ti} - \bar{X}_i$	Deviation of X at time t from the mean for person *i*
Y_B	Portion of Y that is purely between-persons (level-2)
Y_{ti}	Value of outcome variable Y for person *i* at time *t*
Y_W	Portion of Y that is purely within-persons (level-1)
β_{0i}	Intercept (predicted mean value of Y_{ti}) for person *i*
β_{1i}	Within-persons (level-1) regression slope coefficient of Y on a predictor variable for person *i*
γ	Vector of regression slope coefficients for all predictor variables
γ_{00}	Intercept (predicted mean value of Yti) for all persons
γ_{01}	Average within-persons (level-1) regression slope coefficient of Y on a predictor variable
$\gamma_{01} - \gamma_{10}$	Contextual effect
γ_{01std}	Standardized average within-persons (level-1) regression slope coefficient
$\gamma_{01std} - \gamma_{10std}$	Standardized contextual effect
γ_{10}	Between-persons (level-2) regression slope coefficient of Y on a predictor variable

γ_{10std} Standardized between-persons (level-2) regression slope coefficient of Y on a predictor variable

γ_B Vector of regression slope coefficients for between-persons (level-2) predictors

γ_W Vector of regression slope coefficients for within-persons (level-1) predictors

$\mu'_{X(q)}$ Vector of means of predictor variables

η^2 Proportion of explained variance in the outcome variable due to a specific factor in ANOVA; given by $\eta^2 = SS_{effect}/SS_{total}$

σ^2 Within-persons (level-1) residual variance in a multilevel model with predictors

σ_E^2 Within-persons (level-1) residual variance in a multilevel model without predictors

Σ_X^B Between-persons (level-2) covariance matrix of X

$\Sigma_{X(p)}$ Covariance matrix of predictor variables that have random slopes

$\Sigma_{X(q)}$ Covariance matrix of all predictor variables

Σ_X^W Between-persons (level-2) covariance matrix of X

τ_{00}^2 Between-persons (level-2) intercept residual variance in a multilevel model with predictors

τ_{01}^2 Between-persons (level-2) covariance of intercept residual and slope residual

τ_{11}^2 Between-persons (level-2) slope residual variance

τ_E^2 Between-persons (level-2) residual variance in a multilevel model without predictors

APPENDIX B

COMPONENTS OF VARIANCE

Snijders & Bosker (2012) provided a formula that shows the variance decomposition of *Y* in a multilevel model with q predictor variables, p of which have random slopes.

$$var(Y_{ij}) = \gamma'\Sigma_{X(q)}\gamma + \tau_{00}^2 + 2\mu'_{X(q)}T_{10} + \mu'_{X(q)}T_{11}\mu_{X(q)} + trace(T_{11}\Sigma_{X(p)}) + \sigma^2$$

So γ is the vector of regression coefficients of all predictor variables, $\Sigma_{X(q)}$is the covariance matrix of all predictor variables, $\Sigma_{X(p)}$ is the covariance matrix of predictor variables that have random slopes, $\mu'_{X(q)}$ is the mean vector of predictor variables, T_{10} is the vector of intercept-slope covariances, T_{11} is the covariance matrix of random slopes, and τ_{00}^2 and σ^2 are the between-group and within-group residuals, respectively. According to equation 5, the data-generating model has one predictor with a random slope at the within-level $(X_{ij} - \bar{X}_j)$ and one predictor at the between-level $(\bar{X}_j)$, so $\gamma'\Sigma_{X(q)}\gamma$ can be decomposed into within and between components

$$\gamma'\Sigma_{X(q)}\gamma = \gamma'_W \Sigma_X^W \gamma_W + \gamma'_B \Sigma_X^B \gamma_B$$

and

$$\Sigma_{X(p)} = \Sigma_X^W.$$

Also, the means of both within-group and between-group predictors are generated to be zero, so $\mu_{X(q)} = 0$. Then the residual variance of *Y* at each level can be found by:

$$.8 = \sigma_E^2 = \gamma'_W \Sigma_X^W \gamma_W + trace(T_{11}\Sigma_X^W) + \sigma^2$$

$$.2 = \tau_E^2 = \gamma'_B \Sigma_X^B \gamma_B + \tau_{00}^2$$

For example, let one simulation condition have ICC$_x$ = .4, γ_{10}= .59, γ_{01}= .59, and random slope variance τ^2_{11} = 0.01. Across all simulation conditions, *X* and *Y* have a variance of 1, *Y* has an unconditional ICC of .2($\tau^2_E =$ .2 and $\sigma^2_E =$ .8), and the mean of *X* is zero both between and within groups.

Then

$$\sigma^2 = .8 - (\,\gamma_W'\,\Sigma_X^W \gamma_W + trace(T_{11}\Sigma_X^W))$$

$$\sigma^2 = .8 - (\,.59(.6).59 + .01(.6))$$

$$\sigma^2 = .58514$$

and

$$\tau^2_{00} = .2 - (\gamma_B'\,\Sigma_X^B \gamma_B)$$

$$\tau^2_{00} = .2 - (.59(.4).59)$$

$$\tau^2_{00} = .06076$$

These values of σ^2 and τ^2_{00} are then given as the residual within-persons and between-persons variance of Y, respectively, in the Mplus data generation code.

APPENDIX C

EXAMPLE MPLUS CODE

Example Mplus code for generating multilevel data

```
TITLE: Monte Carlo Simulation within size = 20 between size
= 100 between effect = 0.59 within effect = 0.59 ICCx = 0.4
slopevar = 0.01

Montecarlo:

    NAMES ARE X Y xmean;
    NOBSERVATIONS = 2000;
    NREPS = 1000;
    SEED = 6712345;
    NCSIZES = 1;
    CSIZES = 100(20);
    WITHIN = X;
    BETWEEN = XMEAN;
    REPSAVE = ALL;
    SAVE = clustermeandata_20_100_0.59_0.59_0.4_0.01_*.dat;
    RESULTS =
clustermeanRESULTS_20_100_0.59_0.59_0.4_0.01.DAT;

MODEL POPULATION:
    %WITHIN%
    X*0.6;
    [X*0];
    SLOPE | Y ON X;
    Y*0.585;
    %BETWEEN%
    XMEAN *0.4;
    [XMEAN *0];
    [Y*0];
    Y ON XMEAN *0.59;
    [SLOPE*0.59];
    SLOPE*0.01;
    Y*0.06;
    Y WITH SLOPE@0;

ANALYSIS: TYPE = TWOLEVEL RANDOM;

MODEL:
    %WITHIN%
    X*0.6;
    SLOPE | Y ON X;
    Y*0.585;
    %BETWEEN%
    XMEAN *0.4;
    [Y*0];
    Y ON XMEAN *0.59 (b);
    [SLOPE*0.59] (w);
    SLOPE*0.01;
    Y*0.06;
    Y WITH SLOPE@0;

MODEL CONSTRAINT:
    NEW(CONTEXTUAL*0);
    CONTEXTUAL = b - w;
```

Example Mplus code for estimating means-only model

```
TITLE: Monte Carlo Simulation within size = 20 between size
= 100 between effect = 0.59 within effect = 0.59 ICCx = 0.4
slopevar = 0.01

DATA:

    FILE = data_20_100_0.59_0.59_0.4_0.01_LIST.dat;
    TYPE = MONTECARLO;

VARIABLE:

    NAMES = XMEAN Y X CLUSTER;
    USEVARIABLES = Y;
    CLUSTER = CLUSTER;

ANALYSIS: TYPE = TWOLEVEL RANDOM;

MODEL:
    %WITHIN%
    Y*.80;
    %BETWEEN%
    [Y*0];
    Y*.20;

SAVEDATA:

    RESULTS ARE
    UNRESULTS_20_100_0.59_0.59_0.4_0.01.dat;
```