December 2018

Determining Predictor Importance in Multilevel Models for Longitudinal Data: An Extension of Dominance Analysis

LUCIANA PACHECO CANCADO
University of Wisconsin-Milwaukee

DETERMINING PREDICTOR IMPORTANCE IN MULTILEVEL MODELS FOR LONGITUDINAL DATA: AN EXTENSION OF DOMINANCE ANALYSIS

by

Luciana Pacheco Cançado

A Dissertation Submitted in

Partial Fulfilment of the

Requirements for the Degree of

Doctor of Philosophy

in Educational Psychology

at

The University of Wisconsin-Milwaukee

December 2018

ABSTRACT

DETERMINING PREDICTOR IMPORTANCE IN MULTILEVEL MODELS FOR LONGITUDINAL DATA: AN EXTENSION OF DOMINANCE ANALYSIS

by

Luciana Pacheco Cançado

The University of Wisconsin-Milwaukee, 2018
Under the Supervision of Professor Razia Azen

Longitudinal models are used not only to analyze the change of an outcome over time but also to describe what person-level and time-varying factors might influence this trend. Whenever a researcher is interested in the factors or predictors impacting an outcome, a common follow-up question asked is that of the relative importance of such factors. Hence, this study aimed to extend and evaluate Dominance Analysis (DA), a method used to determine the relative importance of predictors in various linear models (Budescu, 1993; Azen & Budescu, 2003; Azen, 2013), for use with longitudinal multilevel models. A simulation study was conducted to investigate the effect of number of measurement occasions (level-1 units), number of subjects (level-2 units), different levels of model complexity (i.e., number of predictors at level-1 and level-2), size of predictor coefficients, predictor collinearity levels, misspecification of the covariance structure, and measures of model fit on DA results and provide recommendations to researchers who wish to determine the relative importance of predictors in longitudinal multilevel models. Results indicated that number of subjects was the most important factor influencing the accuracy of DA in rank-ordering the model predictors, and that more than 50 subjects are needed to obtain adequate power and confidence in the reproducibility of DA results. The McFadden pseudo R^2 is recommended as the standard measure of fit to use when

performing DA in multilevel longitudinal models. Finally, asymptotic standard error and percentile confidence intervals constructed through bootstrapping can be used to determine if one predictor significantly dominates another but might not provide sufficient power unless there are at least 200 subjects in the sample or the magnitude of the general dominance difference measure is greater than 0.01 using McFadden's R^2.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are many people who helped me through the long journey of completing this dissertation and Ph.D. whom I would like to acknowledge and thank. First of all, I would like to thank my advisor, Dr. Razia Azen, for her support, guidance and the many hours spent reading drafts of this work and providing invaluable feedback. Dr. Azen's generous and patient advice and mentoring not only helped make this dissertation better, but it also helped make me a better researcher. I would also like to extend my gratitude to my committee members, Bo Zhang, David Klingbeil, David Budescu and Wen Luo, for their time and help throughout this process. Additionally, I must thank Dr. Cindy Walker who, as program director when I applied to the Ph.D., was the first to trust my potential to be successful in this program. I am also endlessly thankful to my parents, Walter and Fátima, my sister Flávia, and my brothers for their unwavering support and belief in me. I am fortunate to have a wonderful family who was always by my side and never doubted that I could achieve my goals. Finally, I would like to thank the friends who lent me their ears, provided much needed distraction, and were patient and empathetic as I disappeared into seclusion in order to complete this dissertation. My deepest gratitude goes to Simone Conceição, who made me aware of, and encouraged me to pursue this Ph.D. program. I am also profoundly thankful to Andrea for being an endless source of emotional support and good laughs when I needed the most. I also want to say thanks to the friends I made in the Consulting Office for Research and Evaluation at UWM throughout the years, Kevin, Logan, Yao, Stanley, Dian, Shuang, Sam and Sonja. Being able to share laughs, lessons and frustrations throughout the Ph.D. was invaluable both professionally and emotionally. And to many other friends, close and far, who believed in me, thank you.

CHAPTER 1. INTRODUCTION

Multilevel models, also called hierarchical linear models (Raudenbush & Bryk, 2002), random coefficient models (de Leeuw & Kreft, 1986), or linear mixed models (Littell, Milliken, Stroup, & Wolfinger, 1996), were developed to analyze nested or hierarchical data where individuals are nested into groups. Longitudinal data, obtained when the same unit or person is measured at multiple points in time, are commonly found in applied fields such as educational and psychological research, among many others. The need to understand how an outcome changes over time, and what factors might influence these changes, is a typical research question in areas such as school effectiveness, human development, and program evaluation to name a few. The analysis of systematic change over time, commonly called growth curve modeling, is a straightforward application of multilevel modeling where the repeated measures are seen as nested within persons, which could be further nested within higher level units (Fox, 2010).

Traditional methods of analyzing longitudinal data, such as repeated measures analysis of variance and multivariate analysis of variance, can be highly restrictive, imposing assumptions such as equal spacing between observations (i.e., time points), equal number of observations for all individuals, and complete (i.e., no missing) data. The use of multilevel models for the analysis of longitudinal data has become increasingly popular because these models are very flexible in terms of the inclusion of complex features including partially missing data, unequally spaced time points, non-normally distributed or discretely-scaled repeated measures, complex nonlinear growth paths, time-varying predictors, and multivariate growth processes. Longitudinal data analysis using multilevel models has been the focus of, among others, the books by Verbeke and Molenberghs (2000), Singer and Willett (2003), and Hedeker and Gibbons (2006).

Growth curve models are so popular because they can be used not only to analyze the change of an outcome over time, but also to describe what person-level and time-varying factors might influence this trend. Whenever a researcher is interested in the factors or predictors impacting an outcome, a common follow-up question is that of the relative importance of such factors. In the context of multiple regression, such questions have been answered by utilizing what has been called relative importance analysis (Budescu & Azen, 2004; Tonidandel & LeBreton, 2011). Given the widespread use of multilevel longitudinal models and the need to understand the relative contributions of predictors in such models, this dissertation aims to answer the following question: once a given growth curve model has been identified, how can the relative importance of the predictors (or explanatory variables) contained within this model be determined? Answering this question might seem straightforward but, especially when the predictors are correlated, it is not. The difficulty lies in the very definition of importance and how it is supposed to be measured.

Relative importance is defined here as the additional contribution of a given explanatory variable, in comparison to others in the selected model, in predicting the outcome (Azen & Budescu, 2003; Budescu, 1993). In multiple linear regression, several measures and corresponding analytical methods have been proposed to determine relative importance, such as Dominance Analysis (DA; Azen & Budescu, 2003; Budescu, 1993), relative weight analysis (J.W. Johnson, 2000), and measures based on information (Retzer, Soofi, & Soyer, 2009). Dominance analysis is regarded by many researchers as a comprehensive approach for determining relative importance when predictors are correlated, and it is generally recommended when it is computationally feasible to do so (Gromping, 2015; LeBreton, Ployhart, & Ladd, 2004; Thomas, Zumbo, Kwan & Schweitzer, 2014). In fact, in a review of the research on predictor importance published in 2004, Johnson and LeBreton state that DA is “the first measure that was theoretically meaningful and

consistently provided sensible results (pp. 241).” Therefore, this dissertation focuses on the use of dominance analysis to quantify the concept of predictor importance.

There is a large body of literature on variable importance for traditional statistical methods (e.g., Budescu, 1993; Darlington, 1968; Green, Carroll, & Desarbo, 1978; J.W. Johnson, 2000; Kruskal, 1987; Lindeman, Merenda & Gold, 1980; Pratt, 1987). However, the issue of relative importance in multilevel models in general, and growth curve models in particular, has received much less attention (Liu, Zumbo & Wu, 2014; Luo & Azen, 2013). The still widespread reliance on *p*-values, standardized regression coefficients, and other less informative measures for evaluating the importance of predictors in these models suggest that there is a need for better ways to understand the relative contributions of explanatory variables in growth models. The purpose of this study is to help fill this gap by demonstrating how to assess and rank-order the relative importance of predictors in a multilevel model for longitudinal data using dominance analysis.

Dominance analysis examines all possible subset models formed from a set of predictors and compares the incremental fit obtained when each predictor is added to each subset model. One predictor is said to dominate another if it produces a larger incremental fit in each of the subset models or, more weakly, on average across models. The dominance relationship can be defined at three levels, providing a rich picture of the relative contributions of the predictors to explaining the outcome. Complete dominance is established when a predictor dominates (contributes more than) another in each and every subset model. Conditional dominance is established when a predictor’s average additional contribution within all subset models of a given size is greater than that of the other predictor. General dominance is achieved when the average conditional additional contribution of a predictor across all model sizes is greater than that of another predictor. General dominance is the weakest of the three levels but also the most straightforward to determine.

Additionally, it provides an intuitive quantitative representation of the predictor's relative contributions to overall model fit (Luo & Azen, 2013).

DA only requires a measure of model fit in order to assess the additional contribution of a predictor to a model. In linear (multiple) regression, the coefficient of determination, R^2, is well understood and can be easily decomposed in a variety of ways to help determine relative importance in terms of each predictor's relative contribution to the total variance explained in the model. In other linear models, such as generalized linear models (e.g., logistic regression), linear multilevel models, or generalized linear multilevel models (e.g., logistic multilevel models), there is no universal analogue to the coefficient of determination, and the estimation of predictor relative importance in terms of contributions to the model's explanatory power is less clear. Multilevel models for longitudinal data pose additional challenges since one must account for the time-varying structure of the data, including the modeling of errors that might demand special covariance structures. The concept of variance explained in multilevel models is an active area of research, and no single definition exists as to how to measure it. Recently, Nakagawa and Schielzeth (2013) and Jaeger, Edwards, Das, & Sen (2017) have each proposed new measures that claim to overcome some of the problems, such as negative values, that have plagued older pseudo-R^2 measures for multilevel models such as those proposed by Raudenbush and Bryk (2002) and Snijders and Bosker (2012). However, there are still very few comparative studies of these measures, most of them appearing in the original papers proposing the newer measures (Jaeger et al., 2017; LaHuis, Hartman, Hakoyama, & Clark, 2014; Nakagawa & Schielzeth, 2013).

Prior research (Luo & Azen, 2013) provided some indication that DA might be a suitable method for determining relative importance in multilevel models with a continuous outcome. Work by Azen and Cancado (2017, July) provided further evidence of the suitability of DA for

linear multilevel models. This dissertation aims to further extend and evaluate the use of dominance analysis for determining the relative importance of predictors in multilevel models for longitudinal data with continuous (normally distributed) outcomes. Monte-Carlo simulations are used to investigate the impact of model complexity, sample size, collinearity and covariance misspecification on the accuracy of dominance analysis results in terms of (1) rank-ordering of predictors by relative importance, (2) the performance of bootstrap-based inferential procedures for the quantitative general dominance measure, and (3) the reproducibility rates of the qualitative general dominance measure over many bootstrap samples. Of added interest is the performance of different measures of model fit, especially those proposed more recently, on the inferential procedures for the general dominance measure. This study aims to contribute to the literature on both relative importance and multilevel models by examining and demonstrating the use of dominance analysis to answer questions about how predictors compare in terms of their influence on outcomes that change over time.

CHAPTER 2. LITERATURE REVIEW

This literature review will cover six major topics. First, an overview of the general theory of multilevel models will be presented. Next, the specific concepts related to multilevel models for longitudinal data will be discussed, and the specific models used in the simulation study will be introduced and their rationale explained. Then, the concept of predictor importance will be introduced along with an exploration and critique of methods currently used to measure predictor importance in multilevel models. An in-depth review of dominance analysis follows. Subsequently, different measures of model fit that have been proposed for multilevel models are discussed. Lastly, the different bootstrap methods that could be used to carry out the inferential analyses are presented and critiqued, and the chosen method justified.

Multilevel Models

Many areas of research must deal with data that are nested or clustered, such as students within schools, patients within hospitals, or yearly screenings within individuals. This nesting introduces dependencies between individual observations, since observations from a given cluster or group (e.g., school or hospital) are often more similar to each other than to observations from a different cluster. If this dependence is ignored by, for example, using analysis of variance (ANOVA) or a linear regression model to analyze the data, standard error estimates are downward biased leading to erroneous rejection of the null hypothesis (Raudenbush & Bryk, 2002; Hox, 2010; Snijders & Bosker, 2012). Multilevel models (also commonly known as hierarchical models or mixed models) have been proposed to handle data structures where observations are not independent, a central assumption of linear regression models. Multilevel models are designed to

combine information about variables from different levels of a hierarchical structure in a single model, modeling the dependency inherent in the lower level units by including different variance terms for the various levels (Hox, 2010). One of the primary differences between multilevel models and other linear "single-level" models, such as multiple regression, is the ability to estimate one or more of the coefficients or "effects" in the model as either fixed or random. A fixed effect has only a single parameter value in the whole model and is applied to each observation in the analysis regardless of the cluster under which the observation is nested. A random effect, on the other hand, is allowed to vary between clusters.

Multilevel models can be formulated in different ways, resulting in different notations. Matrix representation, a more concise way of organizing the different models, is widely used in software documentation such as the SAS/STAT® User's Guide (SAS Institute Inc., 2017). Algebraic or scalar notation, on the other hand, allow multilevel models to be formulated by either presenting separate equations for each of the levels, combining them into a single equation, or, usually for didactic reasons, writing separate equations at multiple levels and then substituting in to arrive at a single equation (Singer, 1998).

In matrix form, the linear multilevel model can be expressed as:

$$\underset{(N\times 1)}{\boldsymbol{y}} = \underset{(N\times(p+1))}{\boldsymbol{X}}\ \underset{((p+1)\times 1)}{\boldsymbol{\beta}} + \underset{(N\times Q)}{\boldsymbol{Z}}\ \underset{(Q\times 1)}{\boldsymbol{u}} + \underset{(N\times 1)}{\boldsymbol{e}} \quad (1)$$

where $\boldsymbol{y}$ is an N x 1 stacked vector of observed outcome measures for all M subjects, N is the total number of observations where $N = \sum_{i=1}^{M} n_i$, n_i is the number of observations for subject i, $\boldsymbol{X}$ is the $N \times (p+1)$ design matrix corresponding to the $(p+1) \times 1$ fixed-effects parameter vector $\boldsymbol{\beta}$ that contains the intercept and all the p fixed-effects (e.g., main effects and interactions); $\boldsymbol{Z}$ is the $N \times Q$ blocked design matrix corresponding to the $Q \times 1$ vector $\boldsymbol{u}$ that contains all the

random effects Q, where $Q = M \cdot (q + 1)$ (q random slopes and 1 intercept) aggregates the subject specific random effects, and $\boldsymbol{e}$ is the N x 1 vector of level-1 residuals. We assume that $\boldsymbol{u}$ and $\boldsymbol{e}$ are uncorrelated and normally distributed with zero mean and covariance matrices **G** and **R**, respectively. Hence,

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{pmatrix}\right] \tag{2}$$

The expected value and variance (**V**) of the observation vector $\boldsymbol{y}$ are given by:

$$E[\boldsymbol{y}] = \boldsymbol{X\beta} \tag{3}$$

$$Var[\boldsymbol{y}] = \boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R} \tag{4}$$

The vector of observations $\boldsymbol{y}$ is assumed to be normally distributed, $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$. The variance **V** can be modeled by configuring the random-effects design matrix **Z** and specifying **G**, the covariance matrix for the random-effects parameters, and **R**, the covariance matrix of the level-1 errors. If we assume that level-1 errors are homoscedastic, then $\boldsymbol{R} = \sigma^2 \boldsymbol{I_N}$, where $\boldsymbol{I_N}$ corresponds to the $N \times N$ identity matrix. The general linear model can be defined as a special case where **Z** = **0** and $\boldsymbol{R} = \sigma^2 \boldsymbol{I_N}$ (SAS Institute Inc., 2017).

Multilevel Models for Longitudinal Data

Longitudinal data is collected as a set of repeated measurements on individuals across time. Longitudinal data can be considered a special case of multilevel data with the repeated measures (level 1) nested within individuals or subjects (level 2). These models can also be extended to include higher-level units (such as schools or clinics). Longitudinal data has specific characteristics that makes it an ideal candidate for multilevel analysis: (1) there are (at least) two sources of

variability: one within subjects (intra-individual) and one between subjects (inter-individual); (2) the within-subject observations are generally not independent of each other; (3) the between-subject observations may not be constant over time; (4) the data set is usually incomplete or unbalanced, for example due to participants dropping out of a study or missing one or more measurement occasions; and (5) the points in time at which different subjects are measured might not be the same. The analysis of longitudinal data is complex because models must simultaneously account for within-subject observation dependence, between-subject variation, non-constant variance, and unbalanced data (Hox, 2010; Singer & Willett, 2003).

Traditionally, repeated measures data were analyzed with univariate Analysis of Variance (ANOVA) or its multivariate extension (MANOVA), where the main focus is testing the null hypothesis that the means are equal across all occasions. The biggest advantages of these approaches are their simplicity and well-understood properties. However, ANOVA and MANOVA methods make several assumptions about the data that are likely to be violated in practice. ANOVA assumes sphericity, or equal variances for the differences between all possible pairs of time points, which is unlikely to occur if variances increase over time or the correlations between measurements decrease as a function of time. MANOVA, on the other hand, allows a general covariance structure for the repeated measures. However, both MANOVA and ANOVA have the disadvantage of requiring complete data for all subjects and identical measurement occasions. The use of multilevel models for longitudinal data overcomes these limitations and provides a flexible framework to study change over time (Hox, 2010; Singer & Willett, 2003). Specifically, incomplete data resulting from missed measurement occasions is handled seamlessly as long as it can be reasonably assumed that data is missing at random (MAR; Rubin, 1976), and

the procedure allows for a variable number of measurement occasions as well as different spacing between time points (Hox, 2010).

Longitudinal studies can have substantially different designs in terms of number, timing and the balanced nature of the repeated measurements. In this study, the focus is on growth curve models where analysis is aimed at estimating change in the outcome variable over time and on predictors of such change. In general, longitudinal models include two types of predictors: those whose values are constant throughout the duration of the study (e.g., birth year, race), here referred to as time-invariant, person-level, or level-2 predictors, and those whose values change depending on time (e.g., weight), which are referred to as time-varying, time-level, or level-1 predictors.

Additionally, in growth curve models an explanatory variable for time needs to be explicitly included in the models to represent the time points directly. Consider a study with four waves of measurement, where data is collected at baseline and then after 6, 12 and 24 months. An individual measured at all four occasions would have $t = 1, 2, 3, 4$, whereas an individual who missed the second wave would have $t = 1, 3, 4$, and would be missing the measurement obtained at month 6. However, t is not used to index time directly (e.g., $t = 2$ represents month = 6), so the explanatory variable for time, representing the number of months since baseline and denoted $Time_{ti}$, needs to be explicitly included in the model to indicate the actual differences between time points (e.g., $Time_{1i} = 0$, $Time_{2i} = 6$, $Time_{3i} = 12$, $Time_{4i} = 24$). That is, t represents the measurement wave (on an ordinal scale) whereas *Time* represents the actual number of months (or a continuous time scale). The coding of time in growth curve models is also important since it affects the meaning of the fixed and random intercept components in the model. The time variable should be scaled so that the time point with a value of zero corresponds to the time point when the researcher wants a snapshot of the between-subjects differences. Care should be taken in the coding of time

because the fixed intercept is the average value of the response variable at whatever occasion time is coded as zero and the intercept variance represents inter-individual differences at that particular time point (Hox, 2010).

Using an example to put the models studied here into context, suppose a researcher has data on reading comprehension from a random sample of students measured once a year from first to fourth grade. Through appropriate model selection procedures, a set of explanatory variables is selected to model the direction and rate of change in reading comprehension over the four measurement occasions. The data set contains both scale scores on the reading test and a pass/fail decision according to a cut-score. Some of the predictors change over time, such as number of books read in the past year, a measure of social skills, expressive vocabulary skill, and verbal memory; other predictors are measures that do not change over time, such as gender, SES at baseline, number of books at home at baseline, and mother's years of schooling at baseline. To better understand the impact of the predictors, the researcher wants to rank order the variables in terms of their relative importance in predicting the outcome across time.

The focus of this study was the two-level longitudinal model where measurement occasions (level-1 units) are nested within individuals (level-2 units). To introduce the notation, assume the data described above were collected for *M* students, each denoted by *i* and having a total number of measurement occasions n_i. The total sample size is defined as $N = \sum_{i=1}^{M} n_i$. Let the outcome (e.g., reading comprehension score) for student *i*, measured at measurement occasion *t*, be denoted by y_{ti}. Also assume the design is unbalanced; that is, students might be measured at different times and measurement spacing is unequal, such that the gap in time between two consecutive measurement occasions need not be exactly the same. Let *t* index the measurement occasion, where

t varies from 1 to n_i. Then, the general multilevel longitudinal model for student *i*, as introduced by Laird and Ware (1982), can be written as:

$$\underset{(n_i \times 1)}{\boldsymbol{y_i}} = \underset{(n_i \times (p+1))}{\boldsymbol{X_i}} \underset{((p+1) \times 1)}{\boldsymbol{\beta}} + \underset{(n_i \times (q+1))}{\boldsymbol{Z_i}} \underset{((q+1) \times 1)}{\boldsymbol{u_i}} + \underset{(n_i \times 1)}{\boldsymbol{e_i}} \tag{5}$$

$$\begin{pmatrix} \boldsymbol{u_i} \\ \boldsymbol{e_i} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R_i} \end{pmatrix}\right] \tag{6}$$

$$Var(\boldsymbol{y_i}) = \boldsymbol{V_i} = \boldsymbol{Z_i G Z_i'} + \boldsymbol{R_i} \tag{7}$$

where $\boldsymbol{y_i}$ is the n_i-dimensional vector of observed outcomes for student *i*, $1 \leq i \leq M$, *M* is the number of students (subjects), $\boldsymbol{X_i}$ is the known design matrix for the fixed effects, $\boldsymbol{\beta}$ is a (*p*+1)-dimensional vector (*p* predictor effects and 1 intercept) of unknown population (fixed) effects, $\boldsymbol{Z_i}$ is the known design matrix for the random effects, $\boldsymbol{u_i}$ is a (*q*+1)-dimensional vector of unobserved subject-specific (random) effects (*q* random slopes and 1 intercept), and $\boldsymbol{e_i}$ is a n_i-dimensional vector of residual components (i.e., level-1 random errors).

In the context of the reading comprehension example, the $\boldsymbol{X_i}$ matrix is composed of a column of 1's representing the fixed intercept, a column of the values of the $Time_{ti}$ variable, and columns with the values of all fixed predictors, both time-invariant (e.g., gender, SES, number of books at home at baseline, and mother's years of schooling at baseline) and time-varying (e.g., number of books read in the past year, social-skills score, expressive vocabulary skill score, and verbal memory score), as well as their interactions. The components of the **β** vector are the fixed effects or slopes for the variables in $\boldsymbol{X_i}$, which have the same value for all students in the sample. The random effects would be the random intercept (u_{0i}) and the random slope of *Time* (u_{1i}). Assuming the *Time* variable has a value of zero for the first measurement (i.e., for *t* = 0, $Time_{ti} = 0$), the random intercept indicates that the value of the outcome variable, in this example reading

comprehension score, at baseline and when all covariates are zero is allowed to vary among students. Additionally, in longitudinal multilevel models for change (i.e., growth models), the slope of the *Time* variable (e.g., the rate of change of reading comprehension scores over time), is also allowed to vary among students; therefore, a random component for the slope of *Time* (u_{1i}) is included in the $\boldsymbol{u_i}$ vector. The variances and covariance of the random intercept and random slopes are captured in the covariance matrix **G**. The $\boldsymbol{Z_i}$ design matrix in this example would be composed of two columns since we have two random components, one for the intercept, composed of 1's, and another composed of the values of the $Time_{ti}$ variable from $t = 1$ to $t = n_i$. The values in $\boldsymbol{e_i}$ are the time-specific residuals for student *i*, so they can differ per student. The variance of these residuals for student *i* are captured in the covariance matrix $\boldsymbol{R_i}$.

The random effects and level-1 residuals are assumed to be independent ($\boldsymbol{u_i} \perp \boldsymbol{e_i}$), where $\boldsymbol{G}$ is the $(q + 1) \times (q + 1)$ covariance matrix of the student-level (level-2) random effects and $\boldsymbol{R_i}$ is a $(n_i \times n_i)$ covariance matrix for student *i*, but which does not depend on *i* other than through its dimension n_i. Thus, it follows that $\boldsymbol{y_i}|\boldsymbol{u_i} \sim N(\boldsymbol{X_i\beta} + \boldsymbol{Z_i u_i}, \boldsymbol{R_i})$; that is, conditional on the random effects $\boldsymbol{u_i}$, $\boldsymbol{y_i}$ is normally distributed with mean $\boldsymbol{X_i\beta} + \boldsymbol{Z_i u_i}$ and covariance matrix $\boldsymbol{R_i}$. The variance of $\boldsymbol{y_i}$ is $\boldsymbol{V_i} = \boldsymbol{Z_i G Z_i'} + \boldsymbol{R_i}$ and the marginal distribution of $\boldsymbol{y_i}$ is assumed to be $\boldsymbol{y_i} \sim N(\boldsymbol{X_i\beta}, \boldsymbol{V_i})$.

Multilevel analyses usually start with fitting an unconditional, or intercept-only, model in order to calculate the intraclass correlation (ICC). The ICC can be defined as the proportion of total variability in the outcome that is due to the nested structure of the data, or, alternatively, the expected correlation in the outcome of two random level-1 units belonging to the same level-2 unit.

A two-level unconditional model can be represented as:

Level 1: $$y_{ti} = \beta_{0i} + e_{ti} \tag{8}$$

Level 2: $$\beta_{0i} = \gamma_{00} + u_{0i} \tag{9}$$

Combined model: $$y_{ti} = \gamma_{00} + u_{0i} + e_{ti} \tag{10}$$

where $e_{ti} \sim N(0, \sigma^2)$, $u_{0i} \sim N(0, \tau_0^2)$, and $Cov(e_{ti}, u_{0i}) = 0$.

The combined model can be divided into two parts: a fixed part containing the overall intercept, γ_{00}, and a random part containing two random effects: the random intercept coefficient u_{0i} and the level-1 residual e_{ti}. The model shows that the reading comprehension measurement at time *t* for student *i* is a function of three components: the overall mean reading score across all students and time points (γ_{00}), how much student *i*'s mean score deviates from this grand mean (u_{0i}), and how much the actual reading score at time *t* for student *i* differs from the student's model predicted score at that time point (e_{ti}).

The same model in matrix notation for a given student *i* with 4 measurement occasions (n_i = 4) would be:

$$\begin{array}{ccccccccc} \boldsymbol{y_i} & = & \boldsymbol{X_i} & \boldsymbol{\beta} & + & \boldsymbol{Z_i} & \boldsymbol{u_i} & + & \boldsymbol{e_i} \\ \begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \end{bmatrix} & = & \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} & [\gamma_{00}] & + & \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} & [u_{0i}] & + & \begin{bmatrix} e_{1i} \\ e_{2i} \\ e_{3i} \\ e_{4i} \end{bmatrix} \\ (n_i \times 1) & & (n_i \times (p+1)) & ((p+1) \times 1) & & (n_i \times (q+1)) & ((q+1) \times 1) & & (n_i \times 1) \end{array} \tag{11}$$

$$\begin{pmatrix} \boldsymbol{u_i} \\ \boldsymbol{e_i} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R_i} \end{pmatrix} \right] \tag{12}$$

$$with\ \boldsymbol{G} = [\tau_0^2]\ and\ \boldsymbol{R_i} = \sigma^2 \boldsymbol{I_{n_i}}$$

$$\boldsymbol{V_i} = \boldsymbol{Z_i G Z_i'} + \boldsymbol{R_i} = \begin{bmatrix}1\\1\\1\\1\end{bmatrix}[\tau_0^2][1 \quad 1 \quad 1 \quad 1] + \begin{bmatrix}\sigma^2\\\sigma^2\\\sigma^2\\\sigma^2\end{bmatrix} =$$

$$\begin{bmatrix} \tau_0^2 + \sigma^2 & \tau_0^2 & \tau_0^2 & \tau_0^2 \\ \tau_0^2 & \tau_0^2 + \sigma^2 & \tau_0^2 & \tau_0^2 \\ \tau_0^2 & \tau_0^2 & \tau_0^2 + \sigma^2 & \tau_0^2 \\ \tau_0^2 & \tau_0^2 & \tau_0^2 & \tau_0^2 + \sigma^2 \end{bmatrix} \tag{13}$$

The formulation of the unconditional model implies that $Var[\boldsymbol{y}]$ has a *compound symmetry* structure, where the variance for any y_{ti} is $\tau_0^2 + \sigma^2$, the covariance of any two measurements for the same student is τ_0^2, and the covariance between any two measurements from different students is zero. The structure of **V**, the variance of the full vector of responses $\boldsymbol{y}$, is a $M \times M$ block diagonal matrix with M blocks ($\mathbf{V_i}$) of dimension $(n_i \times n_i)$ for each student i (i = 1,…, M) represented by:

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V_1} & 0 & \cdots & 0 \\ 0 & \boldsymbol{V_2} & \cdots & 0 \\ 0 & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{V_M} \end{bmatrix} \tag{14}$$

Once variances at the time and student levels are known, the ICC can be calculated as:

$$ICC = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} \tag{15}$$

The magnitude of the ICC can be used as evidence for the need for multilevel modeling to account for the clustering in the data, since it represents how homogeneous the level-1 units are, or, equivalently, how different from each other the level-2 units are. For longitudinal data, the ICC measures the degree to which an outcome (e.g., reading literacy score, depression score, etc.) of the same individual is more similar to his/her own outcomes across time in comparison to outcomes from other individuals in the sample. Here, since the level-2 units are individuals and

the level-1 units are repeated measures within the same individual, a large ICC is usually expected as there is usually much more variation between individuals than between measurement occasions within individuals.

Next, the specific growth models used in this study are presented. These models are summarized in Table 1 and described in more detail below.

Table 1 Summary of longitudinal models.

Model	$y_{ti} =$	Equation
Time-invariant predictors of random intercept (Model 1)	$\gamma_{00} + \gamma_{10}Time_{ti} + \sum_{h=1}^{4} \gamma_{0h}w_{hi} + u_{1i}Time_{ti} + u_{0i} + e_{ti}$	18
Time-invariant predictors of Time effect (Model 2)	$\gamma_{00} + \gamma_{10}Time_{ti} + \sum_{h=1}^{4} \gamma_{0h}w_{hi} + \sum_{h=1}^{4} \gamma_{1h}w_{hi}Time_{ti} + u_{1i}Time_{ti} + u_{0i} + e_{ti}$	21
Time-varying predictors (Model 3)	$\gamma_{00} + \gamma_{10}Time_{ti} + \sum_{g=2}^{5} \gamma_{g0}x_{(g-1)ti} + \sum_{h=1}^{4} \gamma_{0h}w_{hi} + u_{1i}Time_{ti} + u_{0i} + e_{ti}$	24

Model 1: Growth model with time-invariant predictors of the random intercept. The first model represents a growth model in which both the intercept and the rate of change (over time) vary across individuals. In terms of the reading comprehension example, this model would be used to estimate the effects of student-level (i.e., level-2, time-invariant) predictors (gender, SES, number of books at home and mother's years of schooling) on reading comprehension after accounting for the effect of time, where both the starting point (intercept) and the rate of change

(slope) of the time variable can differ across students. This model can be represented by the following equations:

Level 1: $$y_{ti} = \beta_{0i} + \beta_{1i} Time_{ti} + e_{ti} \quad (16)$$

Level 2: $$\beta_{0i} = \gamma_{00} + \gamma_{01} w_{1i} + \gamma_{02} w_{2i} + \gamma_{03} w_{3i} + \gamma_{04} w_{4i} + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \quad (17)$$

Combined: $$y_{ti} = \gamma_{00} + \gamma_{10} Time_{ti} + \sum_{h=1}^{4} \gamma_{0h} w_{hi} + u_{1i} Time_{ti} + u_{0i} + e_{ti} \quad (18)$$

where y_{it} is the outcome for student *i* at time *t*; γ_{00} is the overall intercept (i.e., the value of the outcome when all predictors are zero); the time variable, $Time_{ti}$, is a continuous measure of time at level-1 scaled so that the first measurement occasion ($Time_{1i}$) has a value of zero and the subsequent values correspond to the distance in months to the first occasion, and γ_{10} is the fixed effect of the *Time* variable (i.e., the linear time trend, measuring the population effect of time on the reading score across all students). Note that this coding of *Time* implies that the overall intercept γ_{00} is the average outcome score at the first measurement occasion. The student-level (time-invariant) predictors are denoted by w_{hi}, where $h = 1,\ldots, p$ predictors, with corresponding fixed effects γ_{0h}. The model components described so far correspond to the *fixed* part of the model. The *random* model components are u_{0i}, representing the deviation of student *i* from the overall mean (γ_{00}), u_{1i}, the random coefficient of the *Time* variable representing the difference between the estimated change over time in the outcome of the *i*-th participant from the average growth in outcome (γ_{10}) across the *M* students, and e_{ti}, representing the residual of student *i* at time *t*, or the difference between the observed and predicted outcome at time *t* for student *i*.

The level-1 errors are modeled by the first-order autoregressive covariance structure, AR(1). Specifically, the level-1 variance is defined as $\sigma^2_{tt`} = \sigma^2(\phi^{|t-t`|})$, where σ^2 is the variance of the independent level-1 errors, *t, t'* are two different time points from the same individual, and ϕ is the first order autoregressive parameter, or the autocorrelation between observations measured at times *t* and *t*-1. For *m* time points, the level-1 residual covariance matrix $\boldsymbol{R_i}$ would be:

$$\boldsymbol{R_i} = \sigma^2 \begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{m-1} \\ \phi & 1 & \phi & \cdots & \phi^{m-2} \\ \phi^2 & \phi & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{m-1} & \phi^{m-2} & \cdots & \cdots & 1 \end{bmatrix}$$

The random effect for the intercept is distributed as $u_{0i} \sim N(0, \tau_0^2)$, the random effect of the time variable is distributed as $u_{1i} \sim N(0, \tau_1^2)$, and their covariance is $Cov(u_{0i}, u_{1i}) = \tau_{01}$. The variation of the student intercepts around the overall average intercept γ_{00} is represented by τ_0^2 and the variation in the individual students' growth rates (differences from the average growth γ_{10}) is represented by τ_1^2. A positive covariance τ_{01} (for example) would indicate that students who have a higher reading comprehension score (outcome) at the first time point measure are more likely to have larger predicted time change in reading scores than those who score lower on reading comprehension at the first time point. The random intercept and random slopes of time are modeled with an unstructured covariance matrix, that is, $\boldsymbol{u_i} \sim N(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{G} = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix}$.

Model 2: Growth model with time-invariant predictors of the time effect. This model adds cross-level interaction terms to the previous model to allow for predictors of the effect of time on the outcome (i.e., the time slope or growth rate). In the reading example, the researcher would use this model to investigate the relative importance of student-level predictors and the interactions between student-level predictors and time on reading comprehension after accounting

for the (fixed and random) effect of time. This model is represented by the following growth model equations:

Level 1: $$y_{ti} = \beta_{0i} + \beta_{1i} Time_{ti} + e_{ti} \quad (19)$$

Level 2: $$\beta_{0i} = \gamma_{00} + \gamma_{01} w_{1i} + \gamma_{02} w_{2i} + \gamma_{03} w_{3i} + \gamma_{04} w_{4i} + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} w_{1i} + \gamma_{12} w_{2i} + \gamma_{13} w_{3i} + \gamma_{14} w_{4i} + u_{1i} \quad (20)$$

Combined:

$$y_{ti} = \gamma_{00} + \gamma_{10} Time_{ti} + \sum_{h=1}^{4} \gamma_{0h} w_{hi} + \sum_{h=1}^{4} \gamma_{1h} w_{hi} Time_{ti} + u_{1i} Time_{ti} + u_{0i} + e_{ti} \quad (21)$$

This model includes level-2 (student-level) variables (w_{hi}) as predictors of both the random intercept and of the random slope of time, with the latter entering as cross-level interactions in the model. The first four terms in the model are fixed components. The last three terms are the random components. The covariance structure for the random components is again modeled as unstructured. Since a cross-level interaction effect was added, the regression coefficients of the individual predictors are conditional effects and must be interpreted along with the interaction term.

Model 3: Growth model with time-varying predictors. This model adds time-varying (level-1) predictors to model 1 to try and explain intra-individual variability in the outcome. In the hypothetical reading example, this model would be used to investigate the effect of the predictors that vary with time (e.g., social skills, number of books read, expressive vocabulary and verbal memory) on change in reading comprehension after accounting for the (fixed and random) effect of time and all other (i.e., student-level, time-invariant) predictors in the model. This model can be represented by the following equations:

Level 1: $$y_{ti} = \beta_{0i} + \beta_{1i}Time_{ti} + \beta_{2i}x_{1ti} + \beta_{3i}x_{2ti} + \beta_{4i}x_{3ti} + \beta_{5i}x_{4ti} + e_{ti} \quad (22)$$

Level 2: $$\beta_{0i} = \gamma_{00} + \gamma_{01}w_{1i} + \gamma_{02}w_{2i} + \gamma_{03}w_{3i} + \gamma_{04}w_{4i} + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}$$

$$\beta_{2i} = \gamma_{20};\ \beta_{3i} = \gamma_{30};\ \beta_{4i} = \gamma_{40};\ \beta_{5i} = \gamma_{50} \quad (23)$$

Combined:

$$\gamma_{00} + \gamma_{10}Time_{ti} + \sum_{g=2}^{5}\gamma_{g0}x_{(g-1)ti} + \sum_{h=1}^{4}\gamma_{0h}w_{hi} + u_{1i}Time_{ti} + u_{0i} + e_{ti} \quad (24)$$

where the $x_{1ti}, \dots, x_{4ti}$ variables represent the time-varying covariates and $\gamma_{20}, \dots \gamma_{50}$ their (fixed) effects on the outcome. Although the value of the time-varying predictors changes across time (i.e., the x_{ti} vary across time t and student i), the parameter value estimating the effect of these variables on the response variable (the γ_{g0}) is assumed to be constant across time.

In practice, when time-varying predictors are included in longitudinal models, some decisions regarding parametrization must be made. Here this model represents a parametrization using grand-mean centered predictors. Details on the issues of centering with time-varying predictors are discussed in the next section.

Time-varying (TV) predictors

Time-varying predictors can be modeled in different ways, resulting in different model parametrizations. The additional complexity in modeling TV predictors is due to the fact that these predictors are usually composed of two sources of variation, one within- and one between-subjects. Therefore, they are actually two variables, one representing the effect of the time-varying predictor on the outcome at a given time point for a given person, and another representing the average (over time) effect of that predictor on the outcome across all individuals (Hoffman & Stawski, 2009).

The different parametrizations allow for teasing out the different effects the TV predictor can have on the outcome. Let *x* be a time-varying predictor indexed by time *t* for student *i*. The simplest parametrization, and usually the incorrect one, is to just grand-mean center *x* and include it in the model at level 1 by itself; i.e., $GMCx_{ti} = x_{ti} - \bar{x}$. This is problematic because the estimate of the effect of this variable conflates the effects of the variable at a specific occasion for a specific person with its effect across all subjects in the sample. The second and third options involve including a new time-invariant (level-2) variable formed by calculating the person-mean (PM) of the time-varying predictor across time for person *i*. The new PM variable should be centered at the grand mean or another constant so that 0 is meaningful, just like any other predictor (i.e., $PMx_i = \bar{x}_i - C$, where *C* could be the grand mean of *x*). Alternatively, the third option transforms the time-varying variable by person-mean centering it; that is, by subtracting the average value of that variable for the given person from it ($PMCx_{ti} = x_{ti} - \bar{x}_i$). The second parametrization allows the investigation of a contextual effect; that is, to find out if after controlling for the absolute value of the time-varying predictor at each occasion, there is still an incremental contribution from having a higher person mean of the TV predictor. The third parametrization allows investigation of the between-person and within-person effects of the time-varying predictors on the outcome separately, by looking at the significance of the fixed effects of each of these predictors. The different parametrizations will impact model estimates and interpretation if the within-person and between-person effects of the TV predictor are different, which is often the case in practice. In the simulation study, for the sake of simplicity, it is assumed that the between-person effects of the time-varying covariates are not substantially different than the within-person effects (i.e., there is no contextual effect) and therefore no additional person-mean variables are added, corresponding to the first parametrization described above.

Estimation methods and inference

Estimation for multilevel models is usually performed using Maximum Likelihood (ML) methods. These methods select as estimates of model parameters the set of values that maximize the likelihood function and produce standard errors for the estimates that can be used for significance testing (Singer & Willett, 2003; Hox, 2010). Two types of ML estimation are commonly implemented in software packages for multilevel analyses: Full Information Maximum Likelihood (FIML) and Residual (or Restricted) Maximum Likelihood (REML). The difference between these methods is that in FIML both the fixed and the random effects are included in the likelihood function while in REML only the variance components (i.e., random effects) are included. Both ML procedures produce a deviance statistic which indicates how well the model fits the data.

The standard errors estimated by ML methods are used in Wald *z* and *t*-tests, where the test statistic is computed as the parameter estimate divided by its asymptotic standard error. The test statistic is compared to a standard normal distribution for the Wald *z* or a *t*-distribution for the *t*-tests to determine *p*-values for the null hypothesis that the given parameter is zero in the population. The degrees of freedom (*df*) for *t*-tests depend on the level of the variable and how many variables are being tested. The Wald *z* is valid only for large samples because it relies on asymptotic standard errors. In general, deviance (likelihood ratio) tests are preferred for testing hypotheses about variance components (Singer & Willett, 2003; Snijders & Bosker, 2012).

Deviance (or likelihood ratio) tests can also be used to test the difference in fit between nested models using the difference between the deviance statistics of two models as the test statistic and comparing this value to a χ^2 distribution with the appropriate degrees of freedom. For testing random components, since the null hypothesis is on the boundary of the parameter space, the

likelihood ratio test does not have the usual large sample χ^2 distribution. In this case, a mixture χ^2 distribution should be used (Self & Liang, 1987). For deviance tests of fixed effects, FIML estimation should be used to obtain the deviance statistic since REML does not include the fixed regression coefficients in the likelihood function. For tests of random effects where the fixed effects are the same, either REML or FIML can be used. The models in this study were estimated using the FIML method since dominance analysis is used here to compare the additional contribution of predictors with fixed effects only.

Missing data in longitudinal studies

In longitudinal studies each subject is measured at a series of time points. It is often the case that some subjects are not measured at every time point or do not have measurements on all variables for various reasons, such as missing one or more measurement occasions, dropping out of a study or failing to answer questions in a test or questionnaire. All of these scenarios produce a situation where the researcher must deal with missing data. The mechanism underlying the tendency of data to be missing, or their "missingness", has important implications to the analysis since modeling incomplete data appropriately depends on the assumptions about these mechanisms. Rubin (1976) introduced the terms that are commonly used to describe the missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

MCAR occurs when missingness is not related to the observed or unobserved outcomes or covariates; that is, the missing values are just a random subset of the complete data. MCAR is the most restrictive assumption regarding the missing data mechanism. A less restrictive case of MCAR is the covariate-dependent MCAR (Little, 1995), where missingness may depend on observed covariates but not on the observed outcome.

MAR occurs when missingness depends only on the observed outcomes and covariates and not on the unobserved (missing) data. For example, a subject might drop out of a study because the symptoms being measured got much better or much worse in the previous measurement. In this case missingness is random conditional on the observed characteristics of the sample data. This is a less restrictive and therefore more realistic assumption regarding the missing data mechanism than MCAR, but now adjustments must be made because observed responses are no longer a random sample.

Lastly, MNAR occurs when missingness depends on the unobserved data; that is, the failure to observe a value depends on the value of either the outcome or covariates that would have been observed. This is the least restrictive but also the most problematic missing data mechanism from the perspective of statistical analysis.

In terms of model estimation, the missing data mechanisms can be divided into ignorable (MCAR and MAR) and non-ignorable (MNAR) missingness (Laird, 1988). Multilevel models for longitudinal data with maximum likelihood estimation have been shown to provide valid inferences in the case of ignorable missingness (MCAR or MAR) without the need to explicitly model the missing data mechanism (Laird, 1988). However, if missingness is non-ignorable, valid inferences require specifying either the correct model for the missing data mechanism or the distributional assumptions for the response variable, or both, and estimators and tests are usually sensitive to these assumptions (Ibrahim & Molenberghs, 2009). For this dissertation, the assumption is that any missing data are at least MAR and therefore can be handled seamlessly through multilevel estimation procedures without the need to specify the missing data mechanism. This is a reasonable assumption commonly made in practice (Collins, Schafer & Kam, 2001; Wang, Fisher, & Xie, 2011).

Predictor Importance Methods

Multilevel models are tools to investigate the relationship between a set of predictors and an outcome variable when the observations are not independent due to clustering. In general, when selecting the appropriate model for the data, researchers have two main goals: explanation and prediction (Pedhazur, 1997). If focus is on prediction, the researcher is interested in finding a set of predictors that will account for the highest amount of variability in the outcome. When the goal is explanation, it is of interest to identify the correct model responsible for the outcome and theory or prior research should be used to determine the best model. In the prediction framework, it matters less which specific predictors are included and their coefficients are not of primary interest; so, while theory can and should guide model selection, the variables might be interchangeable if they produce the same model fit. In the explanation framework, the focus is on finding the specific model responsible for producing (i.e., explaining) the outcome, so the predictors and their coefficient values need to be specified correctly and are of primary interest. Regardless of the approach, but perhaps more applicable for prediction purposes, after a model is selected the focus shifts to evaluating the relationships between specific predictors and the outcome, and the issue of the relative importance of predictors becomes relevant. Researchers and analysts are usually interested in knowing which predictors among the set of explanatory variables in the chosen model have the highest impact on, or the strongest relationship with, the outcome. One might think this question has simple answers, but that is not the case when the predictors are correlated with each other, and the clustering of observations in multilevel data makes the determination of importance even more complicated. The complexity of devising a method (and measure) to clearly determine which predictor is more important than another in a linear model with several predictors comes

not only from the fact that, in most cases, predictors show some degree of correlation between them, but also from a lack of agreement on what is meant by "importance".

The problem with correlated predictors, or multicollinearity, is that the effect of a given predictor on the outcome will vary with values of the other predictors, so separating the effects of individual predictors is difficult. Furthermore, assessing the effect of a predictor that is added to a model depends on the set of predictors already in the model. If predictors are uncorrelated, relative importance is usually easy to determine using either standardized coefficients or simple bivariate correlations. However, because multilevel data violates the independence assumption, calculating simple bivariate correlations and standardized coefficients is not so straightforward. In multilevel data some level of correlation is usually always present due to the nested nature of the data. Observations in the same group, or, in the case of longitudinal data, measurements from the same person, will naturally be more similar than observations from different groups/persons. There is an extensive literature on predictor importance for linear regression models, but even for these simpler models there is no consensus on a generally accepted measure. In the multilevel modeling literature there is still little discussion on the issue of relative importance. Following is a review of some methods that have been proposed for measuring relative importance of variables in linear (ordinary least squares) regression models that might be applicable to multilevel models.

Zero-order correlation (*r*)

The zero-order correlation or bivariate correlation is one of the most basic measures of variable importance. For any given predictor, *x*, the zero-order (simple, bivariate) correlation with *y* is

$$r_{y,x} = \frac{cov(y,x)}{sd(y)sd(x)} \tag{25}$$

where $cov(y, x)$ is the covariance between *y* and *x*, $sd(x)$ is the standard deviation of *y* and $sd(x)$ is the standard deviation of *x*. This measure is appropriate only if one is interested in the isolated relationship between a predictor and the outcome, not accounting for any of the other predictors in the model. Comparing zero-order correlations will not provide a complete picture of the relative importance of a predictor variable in the presence of others if the predictors are correlated.

Standardized regression coefficients

Estimation of the relationship between the outcome and predictors included in a model produces (unstandardized) regression coefficients, which inform researchers as to the incremental or partial predictive power of each predictor in the model. Specifically, unstandardized regression coefficients represent the mean change in the outcome variable (*y*) for one single raw unit of change in the predictor variable (e.g., X_1) while holding the other *p*-1 predictors (e.g., $X_2,\ldots,X_p$) in the model constant. Standardized regression coefficients (β coefficients, beta weights) represent the change in standard deviation units (i.e., the mean change in standard deviation units of the outcome variable) for one standard deviation unit of change in the predictor variable, while holding the other predictors constant. The use of standardized coefficients ignores the predictors' (and outcome's) scale of units, making comparisons between coefficients more straightforward. When predictors are perfectly uncorrelated, each predictor's β weight equals each predictor's zero-order correlation with the criterion variable (*r*, discussed above). In multilevel models, however, standardizing the coefficients is not straightforward because there might be two or more levels and separate sets of variables that account for variance at each level. Thus, it is not clear whether a variable should be standardized with respect to the standard deviation of the outcome at the lowest or higher levels. The issue is even less clear in growth models where one must decide on a specific

occasion to use for calculating the variance since there might be different variance estimates at each occasion.

Pratt index (product measure)

A measure introduced by Hoffman (1960) and later justified by Pratt (1987), the Pratt index (called the "product measure" by Bring, 1996, or "Pratt's measure" by Thomas, Hughes, and Zumbo, 1998) assigns importance to an explanatory variable *j* in proportion to the product of its standardized regression coefficient (β_j) and its zero-order correlation (r_j) with the outcome variable. In multiple regression, the model explained variance (R^2) can be expressed as $R^2 = \Sigma_j \beta_j r_j$, which can be partitioned by computing the Pratt index d_j as:

$$d_j = \frac{\beta_j \times r_j}{R^2} \tag{26}$$

Recently, Liu, Zumbo, and Wu (2014) demonstrated the use of Pratt's measure to determine the relative importance of predictors in multilevel models with a random intercept, fit using a structural equation modeling (SEM) framework. The authors use the purported ability of SEM to partition the variance of a random-intercept-only multilevel model into orthogonal within and between covariance components to obtain the correlations and total explained variances needed to calculate the Pratt index. The main criticism of the Pratt measure is that it can produce negative values for variables in the model, rendering it an inappropriate metric of predictor relative importance (Bring, 1996; Johnson & LeBreton, 2004; Gromping, 2007, 2015). It is not clear if the same problem occurs in the extension of the Pratt measure to multilevel models proposed by Liu et al. (2014).

Akaike weights

A widely used information-theoretic approach for model selection in the biological sciences, Akaike weights (AW; Burnham & Anderson, 2002) can also be used to form a measure of the relative importance of a variable based on the model's Akaike information criterion (AIC; Akaike, 1973). In order to obtain the model's AW, the researcher must first determine a set of *M* candidate models that could be fit to the data, usually based on theoretical grounds. Then, for each model, a difference measure called delta AIC is calculated as:

$$\Delta AIC_i = AIC_i - min(AIC) \tag{27}$$

where AIC_i is the AIC value for model *i*, and *min*(AIC) is the AIC value of the "best" model out of the set (i.e., the model with the smallest AIC). The weights are then calculated as the proportion of a given model's Δ_{AIC} to the sum of delta AICs from all models in the set using an exponential scale transformation. The Akaike weight for model *i* in a set of *M* candidate models is given by:

$$Akaike\ w_i = \frac{\exp\left(-\frac{\Delta \mathrm{AIC}_i}{2}\right)}{\sum_{m=1}^{M} \exp\left(-\frac{\Delta \mathrm{AIC}_m}{2}\right)} \tag{28}$$

A measure of the relative importance of a predictor variable *x* can then be formed by summing the Akaike w_i of models including *x* and comparing it to the same sum for other predictors. It is important to note that the weight of each variable is determined by the number of models in which the variable appears, in addition to the weight of those models. Therefore, the set of candidate models containing each variable must be balanced; that is, all variables should appear the same number of times across the set of candidate models in order to make sensible predictor importance comparisons based on Akaike weights (Burnham & Anderson, 2002). One way to accomplish this would be to include all subset models of the "full" model.

Dominance Analysis (DA)

Dominance analysis, the method that is the focus of this study, determines the relative importance of explanatory variables in a statistical model based on the additional contribution of each predictor to an overall model fit statistic across all subset models. We discuss DA in more detail in the next section.

Other measures

There are many other measures suggested in the literature to quantify predictor relative importance based on different approaches for estimating linear models, but most of these have not been extended to multilevel models. In linear regression, the relative weights measure proposed by J.W. Johnson (2000) is usually compared to dominance analysis as a measure of relative importance of predictors in the context of multiple correlated predictors. J.W. Johnson (2000) and others (Johnson & LeBreton, 2004; Gromping, 2015) argue that the relative weights method produce similar results to the dominance analysis general dominance weights but is much less computationally intensive. However, Thomas, Zumbo, Kwan, and Schweitzer (2014) showed that the method used to derive Johnson's relative weights measure is theoretically flawed and recommend that it no longer be used. In any case, this measure has not yet been extended to multilevel models.

In the Bayesian framework, Bayesian Model Averaging (BMA) has been proposed as a measure on relative importance (Shou & Smithson, 2015). Soofi and colleagues (e.g., Soofi, 1994; Soofi, Retzer, & Yasai-Ardekani, 2000; Retzer, Soofi & Soyer, 2009) used information theory to define importance in terms of the information provided by a predictor for reducing the uncertainty in predicting the outcome, and provide a formal justification and generalization of the "averaging over all orderings" procedure based on the maximum entropy (ME) principle, where importance

measures are provided for categorical and continuous predictors in a unified manner. In the field of machine learning, methods such as random forests offer variable importance measures such as decrease in node impurity (i.e., Gini) for categorical outcomes and permutation-based mean square error reduction for continuous responses (Gromping, 2009, 2015). None of these measures, however, seem to have been extended to applications with multilevel data.

Summary of predictor importance measures

Critics of variable importance research usually say that these are atheoretical techniques that do not provide valuable information (Ehrenberg, 1990; Stufken, 1992; Christensen, 1992). Conversely, the view taken here is that we use statistical methods to gain insight into real world phenomena. The models we use, as the aphorism goes, are useful at best. Variable importance analysis is one tool in a researcher's toolkit that allows for a greater understanding of the process that might have generated the data and helps answer many questions related to the most relevant factors affecting an outcome (Kruskal, 1984).

The general purpose of relative importance analysis is to uncover the contributions of multiple predictors relative to each other (i.e., in relation with or compared to each other) within a selected model (Azen & Budescu, 2003). The (mis)use of significance testing for quantification of relative importance has been a widespread issue for decades (Kruskal & Majors, 1989). The use of raw or standardized regression coefficients for determining predictor importance is commonly found in the literature when one searches for "predictor importance" or "relative importance", despite the fact that the former is known to have an importance-irrelevant association with the scale of the predictor, and that both are misleading when predictors are correlated. When the concept of importance is understood in terms of a predictor's direct, total, and partial effects, dominance analysis is without match and has been generally recommended as the preferred method

for relative importance analysis (Gromping, 2015; LeBreton et al., 2004; Tonidandel & LeBreton, 2011). Next a more detailed description of dominance analysis is presented.

Dominance Analysis

Dominance Analysis was originally proposed by Budescu (1993) and refined by Azen and Budescu (2003) as a method to qualitatively and quantitatively determine the relative importance of predictors in a linear regression model. In this framework, relative importance is measured in terms of the predictor's additional contribution to R^2 in all subsets of a given model of interest. Predictors are compared in a pairwise manner based on a common subset reference model, and this is performed across all possible subset models. Hierarchical levels of dominance are established depending on the pattern of dominance: complete dominance, conditional dominance, or general dominance.

Due to the ambiguous definition of variable "importance", Budescu (1993) proposed three criteria for a method designed to measure relative importance: (1) the importance of a predictor should be related to its contribution to reducing the prediction error, or, equivalently, to the total explained variance, in the outcome (as this is the most intuitive interpretation of importance in a social sciences context); (2) the method should provide a clear way to directly compare the relative importance of predictors so that one can distinguish situations when there is a meaningful difference in the importance of two predictors from situations where this difference cannot be defined or is not meaningful; and (3) the measure of relative importance should provide information of a predictor's contribution at multiple levels: direct, total, and partial. Budescu (1993) then proposed Dominance Analysis as a methodology that satisfied all of these criteria.

The appeal of dominance analysis (DA) is that it provides an intuitive and meaningful definition of importance. Additionally, DA can be performed using any appropriate measure of model fit, which makes it easily extendable to other models. Indeed, so far DA has been successfully extended to a variety of statistical models such as multivariate regression (Azen & Budescu, 2006), logistic regression models (Azen & Traxel, 2009), canonical correlation analysis (Huo & Budescu, 2009), hierarchical linear models (Luo & Azen, 2013), models with multicategory dependent variables (Luchman, 2014), and beta general linear models (Shou & Smithson, 2015). This study proposes extending DA to multilevel models for longitudinal data.

Next, I follow the explanation provided by Azen (2013) with notation found in Shou and Smithson (2015) to describe how to carry out dominance analysis for a generic linear model with p predictors using a measure of model fit that denoted F (e.g., this would be R^2 in multiple regression). DA uses the "all subset models" approach in which each possible combination of the predictors from the (full) model is considered as a subset model, and the measure of fit (F) is recorded for each model. For p explanatory variables, the number of possible subset models is 2^p. Also, a subset model has size k if k predictors are included in the model. Therefore, there will be a number $\binom{p}{k} = \frac{p!}{k!(p-k)!}$ of models of size k. For example, with a total of $p = 4$ predictors, X_1, X_2, X_3, and X_4, $2^p = 2^4 = 16$ subset models would need to be estimated, specifically:

- $k = 0 \rightarrow \binom{4}{0} = 1$ null/empty model;
- $k = 1 \rightarrow \binom{4}{1} = 4$ models with 1 predictor: [X_1], [X_2], [X_3], [X_4];
- $k = 2 \rightarrow \binom{4}{2} = 6$ models with 2 predictors: [X_1X_2], [X_1X_3], [X_1X_4], [X_2X_3], [X_2X_4], [X_3X_4];

- $k = 3 \rightarrow \binom{4}{3} = 4$ models with 3 predictors: [$X_1X_2X_3$], [$X_1X_2X_4$], [$X_1X_3X_4$], [$X_2X_3X_4$]; and
- $k = 4 \rightarrow \binom{4}{4} = 1$ full model with all 4 predictors: [$X_1X_2X_3X_4$];

The incremental contribution a predictor makes to a subset model, ΔF, is defined as the difference between the value of the measure of fit (F) for a model that includes the predictor and F for the same model excluding the predictor. For example, the additional contribution that X_1 makes to model [X_2X_3] is computed as:

$$\Delta \mathcal{F}\,[X_1|X_2X_3] \;=\; \mathcal{F}[X_1X_2X_3] \;-\; \mathcal{F}[X_2X_3] \tag{29}$$

In general, let **X** be a set of p predictors where each predictor, X_i, is compared in a pairwise fashion with each other predictor, X_j (where $i, j = 1, \ldots, p-1$ and $i \neq j$), in terms of their additional contribution to the measure of fit F. Also, let F_q be the fit value for model M_q that does not include X_i, where q = model 1, 2 ,…, 2^p and ΔF_{iq} is the change in F_q when X_i is added to M_q.

If X_i contributes more than X_j to all models M_q that do not include both X_i and X_j, then *complete* dominance between X_i and X_j can be established. Even if complete dominance cannot be established, the additional contributions of the predictors can be averaged in different ways to produce what Azen and Budescu (2003) called *conditional* and *general* dominance.

Conditional dominance is determined by first calculating the measure $\overline{\Delta \mathcal{F}}_{iq|k}$ for predictor X_i by averaging its additional contribution to subset models of a given model size k:

$$\overline{\Delta \mathcal{F}}_{iq|k} = \frac{\sum \Delta \mathcal{F}_{iq|k}}{\binom{p}{k}} \tag{30}$$

where $k \leq p$ denotes the number of predictors in the subset model (model size) and $\binom{p}{k}$ is the number of models of size k. Conditional dominance of X_i over X_j is then measured by the difference between their conditional measures at each model size. If the average additional contribution of X_i is larger than the average additional contribution of X_j for all model sizes (i.e., all k), then conditional dominance is established. If conditional dominance cannot be established, further averaging can be performed to try and establish a weaker level of dominance, general dominance.

The general dominance measure associated with X_i, G_i, is an average of all conditional contributions over all model sizes:

$$G_i = \frac{\sum_{k=1}^{p} \overline{\Delta F}_{iq|k}}{p} \tag{31}$$

The general dominance relationship of X_i over X_j is defined as the difference between their general dominance measures:

$$G_{ij} = G_i - G_j \tag{32}$$

The general dominance measure for X_i therefore measures the mean difference (across all model sizes) between the fit of models that include X_i and the fit of the models (of the same sizes) that do not include X_i.

The quantitative measure G_{ij} can also be used to define a qualitative measure of general dominance between X_i and X_j:

$$D_{ij} = \begin{cases} 1, & \text{if } G_{ij} > 0 \ (X_i \text{ generally dominates } X_j) \\ -1, & \text{if } G_{ij} < 0 \ (X_j \text{ generally dominates } X_i) \\ 0, & \text{if } G_{ij} = 0 \ (\text{general dominance cannot be established}) \end{cases}$$

Complete or conditional dominance between X_i and X_j can similarly be defined by a categorical variable (D_{ij}), but this study focuses on general dominance. Note that it is simpler to evaluate dominance at the general level because only one comparison is required to determine general dominance (the overall averaged contributions) for each predictor pair. Complete and conditional dominance can only be established if multiple comparisons (across all models or across all models of different sizes) all consistently point in the same direction (i.e., favor the same predictor over another).

Even though general dominance is the weakest of the three levels of dominance, it possesses some nice qualities that make it an attractive measure to use for evaluating relative importance. It is easy to compute, requiring only one comparison for each pair of predictors, the values G_i for each predictor in the model add up to the full model's measure of fit F providing a simple "decomposition" of the overall model fit, and it can be established in most cases since it is unlikely that two predictors will have exactly the same values (i.e., overall average) in a given data set (Azen, 2013; Luo & Azen, 2013). Despite its nice properties, Azen (2013) recommends that one should not rely only on the general dominance measure to determine relative importance. Conditional and complete dominance should also be computed and reported if they can be established since they provide stronger evidence for relative importance and might also offer other insight into the relationships among the predictors.

Constrained DA

Dominance Analysis can be extended to evaluate the relative importance of predictors in situations where some predictors must always be included in the model. Constrained dominance

analysis can be performed in cases, for example, where a set of predictors need to be included in the model for statistical control or because they are theoretically essential for predicting the outcome (thus excluding them in a subset model might render the model meaningless). The constrained dominance analysis, as described in Azen & Budescu (2003), is performed by comparing the additional contributions of the predictors of interest only to subset models that contain the essential or mandatory predictors. Subset models that do not contain those required predictors are not evaluated. This feature of DA is employed in this study to evaluate the additional contributions of interactions between predictors. In situations where there is an interaction of two or more predictors, it does not make sense to evaluate the additional contribution of the interaction when the corresponding main effects are not present in the model. Therefore, constrained DA is used to evaluate the interactions while controlling for the predictor main effects.

Inference

Inferential procedures for DA is an area that still needs further research. Even in the well-known case of DA for multiple regression using R^2 as a measure of fit, clear inferential procedures have not yet been put forth that unambiguously answer questions of whether dominance relationships are statistically significant (Azen, 2013). Azen & Sass (2008) investigated the power of the asymptotic method for comparing the additional contribution of a predictor to a model's R^2 and found that the procedure demands very high sample sizes to achieve adequate power. Tang (2014) looked at both asymptotic and bootstrap-based confidence intervals for making inference about the differences between general dominance measures and found that the asymptotic and percentile bootstrap confidence intervals seem adequate when the effect size and sample size are large enough. Azen and Traxel (2009) found that the bootstrap confidence interval does not have enough power to detect a nonzero degree of general dominance in logistic regression and stated

the need for more research on inferential procedures for hypothesis testing purposes and sample size recommendations.

A different approach to investigate the stability and generalization of the qualitative dominance relationships, called reproducibility, was proposed by Azen and Budescu (2003) for multiple regression and used successfully in several extensions of the procedure to other statistical methods (e.g., Azen & Traxel, 2009; Azen & Budescu, 2006). Reproducibility is determined through the use of the bootstrap procedure to simulate the process of random sampling. The original sample data set is resampled with replacement a large number of times, B, to create bootstrap samples of the same size as the original sample. DA is then performed for each bootstrap sample and the qualitative measures of dominance are recorded. The reproducibility measure is computed as the proportion of bootstrap samples that match (i.e., reproduce) the dominance pattern observed in the original sample. This measure can also be reported as a percentage, and indicates the estimated probability that the dominance relationship observed in the sample might also be true in the population. The higher the reproducibility rates, the more confident one can be that the observed dominance relationships are also present in the population. Even though there is no clear threshold for the reproducibility values, Azen (2013) reported results from previous studies suggesting that a minimum reproducibility of 70% might be needed to provide a reasonable level of confidence that the dominance relationships detected in the sample are a reflection of the population values.

In this study the bootstrap procedure is used to construct confidence intervals and to calculate the reproducibility of the general dominance difference measures. For inference, asymptotic and bootstrap confidence intervals are calculated to determine the statistical significance of the difference in general dominance measures. This study should thus provide more

insight into the sampling characteristics of the general dominance difference measure and the behavior of confidence intervals for hypothesis testing with these measures. Inferential procedures using the bootstrap method are further described in last section of this chapter.

Summary of Dominance Analysis

In reviews of measures of relative importance, dominance analysis is generally considered a theoretically sound and encompassing approach for determining relative importance and is consistently among the recommended methods unless computation time is prohibitive (Gromping, 2015; LeBreton, Ployhart, & Ladd, 2004; Thomas, Zumbo, Kwan & Schweitzer, 2014). Therefore, this study considers extending DA for use in longitudinal models under the multilevel framework.

It must be noted that the "all subsets" approach of calculating a predictor's contribution to R^2 had been used by several other researchers (Lindeman, Merenda, & Gold, 1980; Kruskal, 1987; Theil & Chung, 1988; Chevan & Sutherland, 1991; Lipovetsky & Conklin, 2001). However, these methods looked mainly at the average contributions over all orderings, equivalent to the general dominance measure in DA. Therefore, none of the other methods provide all the relative importance measures, and corresponding insight, offered by DA. Gromping (2007, 2015) provides a summary of these related methods.

Dominance analysis needs only a measure of model fit to determine the additional contribution of a predictor to a subset model (Azen & Budescu, 2003; Azen & Traxel, 2009). In linear regression models, the R^2 value not only provides an absolute value for the goodness-of-fit of the model, but is also a summary statistic that describes the proportion of the total variance in the outcome explained by the model. However, in multilevel models there is no single definition of R^2 because variance explained can be defined at different levels of the model. Therefore, to

extend dominance analysis to these models, it is necessary to define what a predictor's additional contribution to the prediction model means in these settings and how to measure this contribution. The next section presents a review of measures of fit that have been proposed in the multilevel model literature and could be used for dominance analysis.

Measures of Fit for Multilevel Models

The main requirement for the use of DA with any statistical model is a measure of model fit that allows the determination of the additional contribution of a given predictor to any subset model of interest (Azen & Traxel, 2009; Luo & Azen, 2013). Therefore, for this study, the selection of such a measure is critical to the extension of DA to multilevel models.

Multilevel models present challenges in measuring model fit due to the multiple sources of unexplained variation. Therefore, the concept of explained variance in multilevel models, in the sense of what R^2 represents for the single-level case, is not clear-cut. Several measures have been proposed as proxies for explained variance in multilevel models but they each come with caveats. Some of the measures of model fit that seem more promising for use with dominance analysis will be discussed in this section.

The following criteria are typically applied for defining appropriate R^2 analogues (Kvalseth, 1985; Van den Burg & Lewis, 1988, Azen & Traxel, 2009; Azen & Budescu, 2006) and will also be used here:

- Boundedness: The measure should vary between a minimum of zero, indicating complete lack of fit, and a maximum of one, indicating perfect fit.

- Linear invariance: The measure should be invariant to non-singular linear transformations of the variables (Y's and X's).
- Monotonicity: The measure should not decrease with the addition of a predictor.
- Intuitive interpretability: The measure of fit is intuitively interpretable, in that it agrees with the scale of the linear case for intermediate values (between 0 and 1).

Model fit measures for multilevel models, to be discussed next, may be categorized into three groups: the first group includes measures based on the idea of variance explained, similar to R^2 in linear regression; the second group comprises measures based on the likelihood ratio, similar to some pseudo-R^2 measures proposed for logistic regression; and the third group represents information criteria measures (e.g., AIC and BIC) commonly used for model selection.

Explained variance measures

Because DA was originally developed using R^2, a natural starting point is to look at R^2 equivalent measures for multilevel models. A survey of the literature makes it clear, however, that extending R^2 from linear models to multilevel models is not straightforward (Snijders & Bosker, 1994; Steele, 2013; Nakagawa & Schielzeth, 2013).

One of the earliest measures of variance-explained in multilevel models was proposed by Raudenbush and Bryk (1986, 2002). Their approach computes separate R^2 statistics for each variance component and the measure is defined as the reduction in variance resulting from adding fixed effect predictors to an "empty", intercept-only, model. For the growth curve models studied here, the R^2 measures corresponding to the level-1 residual variance, and level-2 random intercept and random slope variance components, are defined, respectively, as:

$$R\&B\ R_1^2 = \mathrm{PCV}(\sigma^2) = 1 - \frac{\sigma^2}{\sigma^2_{(null)}} \tag{33}$$

$$R\&B\ R_2^2(intercept) = \text{PCV}(u_{0i}) = 1 - \frac{\tau_0^2}{\tau_{0(null)}^2} \quad (34)$$

$$R\&B\ R_2^2(slope) = \text{PCV}(u_{1i}) = 1 - \frac{\tau_1^2}{\tau_{1(null)}^2} \quad (35)$$

where $\sigma_{(null)}^2$ is the level-1 variance, and $\tau_{0(null)}^2$ and $\tau_{1(null)}^2$ are the level-2 intercept and slopes variance components, respectively, from the unconditional growth model (which includes the random intercept and a random slope for time but no covariates); similarly, σ^2 is the level-1, and τ_0^2 and τ_1^2 are the level-2, variance components from the model of interest. One advantage of these measures is that they can be computed for models with any number of hierarchical levels since they look at individual variance components separately. Additionally, Nakagawa and Schielzeth (2013) recommend reporting these measures, which they refer to as proportion change in variance (PCV), along with other more general R^2 measures, because PCV allows researchers to evaluate specific changes to variance components (random effects and residual variance) at different levels that may result from including specific predictors at each level. However, as pointed out by Snijders and Bosker (1994), these measures can decrease or take on negative values when predictors are added at other levels of the model because, for example, adding fixed effects at level-2 may reduce the variance estimate for one component (e.g., the residual variance) while increasing variance for another (e.g., the random intercept) at the same time.

In order to address the issues of negative R^2 arising from looking at reduction in specific variance components, Snijders and Bosker (1994) proposed different R^2 measures for each level of the model, in the context of two-level random intercept models, which they named R_1^2 and R_2^2. These measures look at the proportional reduction in total error of prediction at each level of the model as an estimate of the variance explained at the given level. For level 1, this measure is defined as:

$$S\&B\ R_1^2 = \frac{\text{var}(y_{ij} - \hat{y}_{ij})}{\text{var}(y_{ij})} = \frac{\sigma^2 + \tau^2}{\sigma_0^2 + \tau_0^2} \quad (36)$$

where R_1^2 is the variance explained at level 1 (i.e., the variance among level-1 observations or units, or within-individual variance in longitudinal models). For longitudinal models, y_{ij} is the *i*th response of the *j*th individual, $\hat{y}_{ij}$ is the *i*th predicted value for the *j*th individual, σ^2 is the level-1 residual variance and τ^2 is the level-2 variance (i.e., random intercept variance) in the model of interest, and σ_0^2 and τ_0^2 are the level-1 and level-2 variances in the null model, respectively.

The variance explained at level 2, R_2^2, is defined as reduction in error in predicting individual (level-2) mean values and can be written as:

$$S\&B\ R_2^2 = \frac{\text{var}(\bar{y}_j - \hat{y}_j)}{\text{var}(\bar{y}_j)} = \frac{\frac{\sigma^2}{n^*} + \tau^2}{\frac{\sigma_0^2}{n^*} + \tau_0^2} \quad (37)$$

where $n^* = \frac{M}{\sum_{i=1}^{M} \frac{1}{n_i}}$ is the cluster size, which in an unbalanced design is the harmonic mean of the number n_i of level-1 units in each of the M subjects.

Snijders and Bosker's (S&B) R^2 measures offer the advantage of being able to measure the amount of additional variance explained in both level-1 (observation or time level) and level-2 (person level) when a predictor is added to this type of model. However, these measures might still decrease with the addition of a fixed predictor in larger models. Snijders and Bosker (2012) claim that decreases in R^2 estimates indicate misspecification of the model and can therefore be used as a diagnostic tool. However, Nakagawa and Schielzeth (2013) argue that misspecification is not necessarily the cause of an increase in the amount of unexplained variance in a model. Another criticism of the S&B R^2 measures, made by Nakagawa and Schielzeth, is that extending these measures to more than two levels is not clear, and that even though Gelman and Pardoe (2006)

proposed a solution so that an arbitrary number of levels could be modeled using these measures, the implementation is technically complex and therefore not readily accessible to applied researchers.

The lack of a widely accepted statistic that can summarize how well multilevel models fit the data prompted Nakagawa and Schielzeth (2013) to propose what they call a "general and simple method" for calculating two types of R^2 for both linear and generalized multilevel models (referred to as linear and generalized linear mixed models by the authors). The authors suggest that these two R^2 statistics, named marginal and conditional R^2, can provide measures of variance-explained that are less susceptible to the problems with previously proposed R^2 measures for mixed-effect models, such as negative values, and are easy to compute using standard statistical software. The marginal R^2 ($R^2_{GLMM_{(m)}}$) is the variance explained by the fixed effects as a proportion of the sum of all the variance components. The conditional R^2 ($R^2_{GLMM_{(c)}}$) estimates the variance explained by both fixed and random factors as a proportion of the total variance. Initially proposed for random-intercept models only, these measures have been extended to random-slopes models by P.C. Johnson, (2014).

Formally, the original random-intercept marginal and conditional R^2 statistics for linear mixed models presented by Nakagawa and Schielzeth (2013) were defined as:

$$R^2_{GLMM(m)} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_\varepsilon^2} \tag{38}$$

$$R^2_{GLMM(c)} = \frac{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_\varepsilon^2} \tag{39}$$

where σ_f^2 is the variance explained by the fixed effects component, σ_l^2 is the variance of the *l*th term of the *u* random effects, and σ_ε^2 is the residual (level-1) variance. For generalized

multilevel models, the residual variance is defined on the latent (link) scale as being composed of: (i) multiplicative dispersion (ω), (ii) additive overdispersion variance (σ_e^2), and (iii) distribution specific variance (σ_d^2). For binomial and Poisson distributions in particular, σ_ε^2 is defined as $\sigma_e^2 + \sigma_d^2$. P.C. Johnson (2014) extended this definition to random-slope GLMM by deriving a general formula for the mean random effect variance:

$$\bar{\sigma}_l^2 = \text{Tr}(\boldsymbol{Z\Sigma Z}')/n \tag{40}$$

where $\mathbf{Z}$ is the design matrix of the random effects of a model with *n* rows and *k* columns corresponding to *k* random effects, $\mathbf{\Sigma}$ is the covariance matrix of the *k* random effects, and Tr denotes the trace operation (summing the main diagonal elements). Most recently, Nakagawa, Johnson, and Schielzeth (2017) expanded their proposed version of R^2 to all other non-Gaussian distributions, with special emphasis on negative binomial and gamma distributions, by deriving the observation-level variance σ_ε^2 using three different methods: the delta method, a lognormal approximation, and the trigamma function. The authors indicated that their proposed R^2 framework could also be used for derivation of semi-partial R^2 by using commonality analysis (Nakagawa et al., 2017).

Edwards, Muller, Wolfinger, Qaqish, and Schabenberger (2008) introduced an R^2 statistic based on the *F*-statistic for a Wald test of fixed effects, which they called R_β^2, as a measure of multivariate association between the outcome of interest and the fixed effects. Edwards et al. (2008) posit that, given a model of interest and a null model with only a random intercept and no covariates, the linear mixed model *F* statistic corresponds to a test of the null hypothesis:

$$H_0: \boldsymbol{C\beta} = \mathbf{0}\ for\ \boldsymbol{C} = \left[\mathbf{0}_{(q-1)\times 1}\boldsymbol{I}_{q-1}\right] \equiv H_0: \beta_1 = \beta_2 = \cdots = \beta_{q-1} = 0$$

where $q - 1$ is the numerator degrees of freedom for full rank $\boldsymbol{C}$. The model *F* statistic is then given by:

$$F\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{V}}\right) = \frac{\left(\boldsymbol{C}\widehat{\boldsymbol{\beta}}\right)'\left[\boldsymbol{C}\left(\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{C}'\right]^{-1}\left(\boldsymbol{C}\widehat{\boldsymbol{\beta}}\right)}{\text{rank}(\boldsymbol{C})} \tag{41}$$

where $\widehat{\boldsymbol{V}}$ is the estimated covariance matrix for the outcomes, $Var(\boldsymbol{y_i}) = \boldsymbol{V_i} = \boldsymbol{Z_i G Z_i'} + \boldsymbol{R_i}$. The proposed R^2_β statistic is then calculated using the one-to-one correspondence between R^2 and the *F*-statistic:

$$R^2_\beta = \frac{(q-1)v^{-1}F\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{V}}\right)}{1+(q-1)v^{-1}F\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{V}}\right)} \tag{42}$$

Most recently, Jaeger, Edwards, Das, and Sen (2017) proposed an extension of R^2_β for generalized linear mixed models (GLMM) where the response variable may come from distributions other than the normal. For the GLMM, Jaeger et al. (2017) calculate the *F*-statistic and corresponding R^2_β for the pseudo linear data created by the penalized quasi-likelihood (PQL) estimation procedure, which they call $R^2_{\beta^*}$. The authors show that their proposed measure also generalizes Nakagawa and Schielzeth's (2013) marginal R^2 and can be used for any distribution and with any link function. Jaeger et al. (2017) claim that $R^2_{\beta^*}$ is unique in providing semi-partial correlations for any combination of predictors with the outcome of interest, and indicate that this semi-partial R^2 measure, denoted $R^2_{\beta^*_j}$ with *j* representing an index for the given fixed-effect parameter in the full model, would be able to answer research questions related to the relative importance of predictors. However, due to the unreliable nature of the Wald test for small samples, the authors advise caution when using their proposed R^2 statistics for small data sets, particularly

with logistic multilevel models. Additionally, the authors note that $R^2_{\hat{\beta}^*}$ could decrease with the addition of a fixed effect under an incorrectly specified covariance.

Likelihood ratio measures

A second set of measures of fit that may be used with multilevel models are R^2-analogue measures, originally proposed for generalized linear models such as single-level logistic regression. These measures are based on ratios comparing the likelihood of the data under the null (empty) model and a competing model (with predictors) and are supposed to indicate how well one can predict the outcome variable from the explanatory variables in the model. Commonly used measures are the ones proposed by Cox and Snell (1989), Nagelkerke (1991), and McFadden (1974).

In order to define these measures, let L_0 represent the likelihood of the null (intercept-only) model, L_M represent the likelihood of the model of interest, and n the total sample size. The Cox and Snell measure can be written as:

$$R^2_{C\&S} = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}} \tag{43}$$

An attractive characteristic of the Cox & Snell measure is that it directly corresponds to the usual R^2 in linear regression models and therefore can be thought of as a "generalized" R^2 instead of a pseudo R^2 (Allison, 2013). However, Cox & Snell's R^2 has a maximum value that is less than 1. When the full model perfectly predicts the outcome and thus has a likelihood of 1, Cox & Snell's R^2 would be $1 - (L_0)^{\frac{2}{n}}$, which can be considerably less than one. Therefore, an adjustment can be performed by dividing the $R^2_{C\&S}$ by its upper bound, $1 - (L_0)^{\frac{2}{n}}$, which produces the R^2 attributed to Nagelkerke (1991):

$$R_N^2 = \frac{1-\left(\frac{L_0}{L_M}\right)^{\frac{2}{n}}}{1-(L_0)^{\frac{2}{n}}} \tag{44}$$

However, this adjustment is ad hoc, thus the resulting statistic does not have the theoretical interpretation of the original $R_{C\&S}^2$ (Allison, 2013).

McFadden's (1974) R^2 measure is defined as:

$$R_M^2 = 1 - \frac{\ln(L_M)}{\ln(L_0)} \tag{45}$$

As summarized by Azen and Traxel (2009), McFadden's measure possesses many desirable properties when applied to single-level logistic regression: it is bounded between 0 and 1, does not depend on the units of the variables in the model, is monotonic, and has an intuitive interpretation. However, these features might not hold when the errors are correlated as is the case in multilevel models.

Even though these pseudo-R^2 measures do not have an independent interpretation (like that of a linear model's R^2), and cannot be used for model comparisons across different data sets, they are valid and useful in evaluating a set of models used to predict the same outcome on the same dataset (UCLA: Statistical Consulting Group, n.d.). There are, however, problems in using likelihood-ratio based R^2 measures with mixed models. Nakagawa and Schielzeth (2013) point out that some unresolved obstacles to using these measures are that they only provide R^2 at the lowest level (level 1) and that they can decrease or become negative with the addition of explanatory variables to the model.

Information criteria measures

The third set of measures of fit available for use with multilevel models are information criteria (IC) statistics. Information criteria are based on the likelihood of the data given a fitted model, and have been commonly used with multilevel models for model selection and comparison (Hamaker, van Hattum, Kuiper, & Hoijtink, 2011; Nakagawa & Schielzeth, 2013; Steele, 2013; Wang, Fisher, & Xie, 2011). Information criteria apply some penalty for the number of estimated parameters and/or sample size and can be used to select the "best" or "better" model from a set of candidate models. Model selection based on information criteria aims to find a balance between model fit and parsimony, achieved by maximizing the likelihood function while also penalizing additional complexity.

Commonly used information criteria include the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwartz, 1978). IC take the general form of:

$$IC = -2\log(Likelihood) + penalty\ term \tag{46}$$

The penalty term is based on the complexity or dimension of the model (d) and the sample size (n). For the AIC measure the penalty term is $2d$ and for BIC it is $d \log n$. The AIC depends only on the number of parameters, while other measures depend on both the number of parameters and the sample size. The penalty term of BIC is more stringent than the penalty term of AIC since for $n \geq 8, d \times log(n) > 2d$. Consequently, the BIC tends to favor smaller models compared to the AIC. Hamaker et al. (2011) provide a thorough discussion of how these and other information criteria can be used to make model selection decisions when fitting multilevel models. Even though IC measures are useful for model comparison, they are not ideal as measures of model fit (e.g., for DA) because they do not provide any information about absolute fit or how much variance the

model is able to explain, and do not provide a measure of each predictor's additional contribution in an intuitive manner. Most importantly, because these measures include a penalty for model complexity, they might increase (i.e., get "worse") if predictors that do not improve fit are added to the model. Therefore, these measures are not monotonic with model complexity.

Summary of measures of fit

A summary of the desirable criteria (boundedness, linear invariance, monotonicity and intuitive interpretability) satisfied by each of the measures of fit (R^2 analogues) for multilevel models described here is presented in Table 2. There seems to be, so far, no measure of fit for multilevel models that meets all four criteria. Thus, this study will use measures of fit that meet at least three of the four criteria. The measures that seem most appropriate for use with multilevel models are Nakagawa and Schielzeth (2013) marginal R^2 ($R^2_{GLMM_{(m)}}$), Edwards et. al.'s R^2_β (2008), McFadden's (1974) R^2_M, and Raudenbush and Bryk's R^2 (PCV), which are highlighted in Table 2. Even though Snijders and Bosker's R^2_1 and R^2_2 measures also meet 3 out of 4 criteria, this measure will not be used because computing these measures for the models with random slopes used in this study is non-trivial.

Table 2. Summary of properties (indicated by x) of R^2 analogues for multilevel models.

	Explained Variance Measures					Likelihood Ratio			Information Criteria	
Property	PCV/ R&B R^2	S&B R^2	N&S $R^2_{(Marg)}$	N&S $R^2_{(Cond)}$	R^2_β	$R^2_{C\&S}$	R^2_N	R^2_M	AIC	BIC
Boundedness	x	x	x	x	x		x	x		
Invariance	x	x	x		x	x	x	x	x	x
Monotonicity										
Interpretability	x	x	x	x	x	x		x		
Total Satisfied	3	3	3	2	3	2	2	3	1	1

Bootstrapping for Multilevel Models

Once the dominance relationship between two predictors has been determined in a sample, researchers might be interested in finding out if this relationship can be considered to represent the "true" dominance relationship between the predictors in the population of interest. For instance, researchers might like to determine if the estimated difference between two general dominance measures is significantly different from zero. Therefore, inferential procedures for dominance analysis measures in multilevel models are investigated in this study.

The challenge in devising procedures to test hypotheses about dominance measures is that, other than for the large-sample multiple regression case using R^2 as measure of fit, the theoretical probability distribution of these measures is not known (Budescu, 1993; Azen & Traxel, 2009; Tang, 2014). The bootstrap method (Efron, 1979) estimates the sampling distribution of a statistic of interest empirically; that is, strictly from the sample data. Since the bootstrap does not rely on distributional assumptions, it can be used to estimate the variability of a statistic whose theoretical distribution is unknown (Mooney & Duval, 1993).

Inferential procedures for dominance analysis based on bootstrapping have been investigated for linear regression models (Azen & Budescu, 2003; Tang, 2014), logistic regression models (Azen & Traxel, 2009) and multivariate regression models (Azen & Budescu, 2006). Here, we investigate the use of bootstrapping for making inferences regarding dominance measures in multilevel models.

The bootstrap can also be used for estimating the *reproducibility* of the DA results, a measure for describing how stable the results may be across repeated sampling and how confident we are that we can reproduce the dominance pattern found in the original sample (Azen & Budescu, 2003, 2006). While to date bootstrap confidence intervals have been used for inference

on the quantitative general dominance measure only, reproducibility measures have been examined for all three dominance levels (i.e., complete, conditional, and general dominance).

The basic bootstrap process for estimating a parameter θ involves the following steps:

1) Draw a random sample (the original dataset) from the population and obtain the parameter estimate, $\hat{\theta}$, using that sample.
2) Draw a random (bootstrap) sample, with replacement, of the same size as the original dataset by resampling from the original dataset.
3) Re-estimate the parameter of interest for this bootstrap sample *b* to obtain $\hat{\theta}_b$.
4) Repeat steps 1 and 2 a large number (B) of times to obtain the distribution of the $\hat{\theta}_b$ values, which represents the empirically estimated sampling distribution of the parameter estimate, $\hat{\theta}$.

As presented in Efron and Tibshirani (1993), the mean of the bootstrap estimates of the parameter is given by $\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$ and the standard deviation by $\hat{\sigma}^* = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \hat{\theta}^*)^2}$, and these values can be used to derive confidence intervals. A bootstrap estimate for the bias of the parameter of interest can be computed as:

$$Bias(\hat{\theta}) = (\hat{\theta}^* - \hat{\theta}) \tag{47}$$

And its bias-corrected estimate is:

$$\hat{\theta}_{bc} = \left(\hat{\theta} - Bias(\hat{\theta})\right) = (2\hat{\theta} - \hat{\theta}^*) \tag{48}$$

Bootstrap confidence intervals can be constructed in different ways depending on how well the distribution of the bootstrap estimates can be approximated by the normal distribution. Two common methods of constructing a confidence interval (Efron & Tibshirani, 1993) are:

1) **Asymptotic Normal CI**: Use the standard deviation of the bootstrap distribution, $\hat{\sigma}^*$, as the estimated standard error and construct a bootstrap confidence interval based on a standard normal distribution. If the sampling distribution is approximately normal, the 100(1-α)% CI can be computed as $CI_{100(1-\alpha)\%} = \hat{\theta}^* \pm z_{\alpha/2}\hat{\sigma}^*$. In this study the 95% CI is used, so α=0.05 and $z_{\alpha/2} = 1.96$ is the value from the standard normal distribution corresponding to the two-sided 95% confidence level.
2) **Percentile CI**: If normality of the bootstrap sampling distribution cannot be established, we can find the middle 100(1-α)% of the distribution by sorting the B bootstrap estimates from smallest to largest and selecting the values corresponding to the $100(\alpha/2)^{th}$ and the $100(1-\alpha/2)^{th}$ positions as the lower and upper confidence limits, $\hat{\theta}^{*100(\alpha/2)}$ and $\hat{\theta}^{*100(1-\alpha/2)}$, respectively. The 95% CI is used here, so the 2.5^{th} and the 97.5^{th} positions are the lower and upper confidence limits, corresponding to $\hat{\theta}^{*2.5}$ and $\hat{\theta}^{*97.5}$.

For multilevel models, however, the basic bootstrap algorithm described above is inadequate since it assumes independent and identically distributed (i.i.d) responses (Goldstein, 2010). The bootstrap might not make assumptions about the specific distribution of the data, but it assumes that the sampling properties of the statistic of interest are preserved in the resampling distribution. Additionally, the bootstrap method is supposed to follow the same probabilistic mechanism that is assumed to have generated the data (van der Leeden, Meijer & Busing, 2008; Goldstein, 2011; Hox & van de Schoot, 2013). Therefore, alternative approaches to obtain the

bootstrap estimate $\hat{\theta}^*$ have been proposed for multilevel data. These bootstrap methods are summarized below and follow the approaches described in Goldstein (2010, 2011) and van der Leeden et al. (2008). There are three general bootstrap approaches for multilevel modeling: parametric residual bootstrap, nonparametric residual bootstrap, and (nonparametric) case resampling.

For the description of the different bootstrap procedures for multilevel models presented below, consider the following two-level model with a random intercept and a random slope for predictor *x*, where $j = 1,..., J$ groups, $i = 1,.., n_j$ individuals per group, and N is the total number of observations (i.e., $\sum_{j=1}^{J} n_j = N$):

$$y_{ij} = \gamma_{00} + \gamma_{10}(x_{ij}) + u_{0j} + u_{1j}(x_{ij}) + e_{ij} \tag{49}$$

Here y_{ij} is the outcome value for person i in group j, γ_{00} and γ_{10} are the fixed intercept and slope, respectively, u_{0j} is the random intercept coefficient, u_{1j} is the random slope for x and e_{ij} is the level-1 residual. Also, assume $\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix}\right]$, $e_{ij} \sim N(0, \sigma_e^2)$, and let $\boldsymbol{G} = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix}$ and $\boldsymbol{R} = \sigma_e^2 \boldsymbol{I}_{\boldsymbol{n_j}}$.

Parametric residual bootstrap. The parametric bootstrap uses the parametrically estimated distribution function of the data to generate bootstrap samples. This method quantifies the design-specific sampling variance by simulating response values based on the estimated distribution of the residuals at each level followed by a re-estimation of the statistic of interest. Assuming the model and its error distributions are correctly specified, the variance among replicated simulations represents the sampling uncertainty of the estimate. This method also assumes that the predictors are fixed. The parametric bootstrap makes the strongest assumptions

of all three methods. For the model presented above, the parametric residual bootstrap is constructed as follows:

1) Draw *N* elements $\hat{e}_{ij}^{*}$ from the estimated distribution of level-1 residuals, $\hat{F}_e \sim N(0, \hat{\sigma}_e^2)$;
2) Draw *J* vectors of elements $\hat{u}_{0j}^{*}$ and $\hat{u}_{1j}^{*}$from the estimated distribution of the random effects $\hat{F}_u \sim N\left(0, \hat{G}\right)$.
3) Generate the bootstrap responses as $y_{ij}^{*} = \hat{\gamma}_{00} + \hat{\gamma}_{10}\left(x_{ij}\right) + \hat{u}_{0j}^{*} + \hat{u}_{1j}^{*}\left(x_{ij}\right) + \hat{e}_{ij}^{*} \quad \forall i, j.$
4) Refit the model and compute the bootstrap value $\hat{\theta}_b$ on the generated sample and store its value.
5) Repeat steps 1−4 B times as to obtain B sets of bootstrap replications of the parameter(s).

The parametric bootstrap is not robust with respect to any deviation from the distributional assumption of the random terms and, therefore, it should be used with caution. Goldstein (2010) observes that an advantage of the parametric bootstrap procedure is that it can be extended straightforwardly to discrete response models. Nakagawa and Schielzeth (2010) indeed recommend the use of the parametric bootstrap to calculate uncertainty estimates for the ICC (which they call repeatability) in GLMM; however, they warn that the parametric bootstrapping may fail when non-Gaussian data exhibit underdispersion.

Non-parametric residual bootstrap. The non-parametric residual bootstrap (Carpenter, Goldstein & Rasbash, 1999, 2003; Goldstein, 2010, 2011) consists of randomly drawing residuals with replacement from transformed (centered and reflated) residuals obtained from the model-

estimated "crude" residuals. The need for "reflating" the residuals comes from the fact that in multilevel models both the level-1 and level-2 residuals are "shrunk" towards zero so that the true variability of the residuals is not reproduced in the bootstrap samples (Carpenter, Goldstein, & Rasbash, 1999). The non-parametric residual bootstrap method does not make assumptions about the distribution of the errors, but it assumes that the explanatory variables are fixed and that the model is correctly specified. These assumptions might make sense if the model is theoretically justified. For the model in Equation (49), the procedure involves the following steps:

1) Use the original sample to fit the multilevel model under study and save the level-1 and level-2 raw residuals.
2) Center both the level-1 and level-2 residuals so that they have a mean of 0.
3) Reflate[1] the (centered) residuals.
4) Draw, with replacement, *J* vectors of elements $\hat{u}_{0j}^{*}$ and $\hat{u}_{1j}^{*}$ from the set of reflated level-2 residuals.
5) Draw with replacement *N* elements $\hat{e}_{ij}^{*}$ from the set of reflated level-1 residuals.
6) Generate the bootstrap responses as $y_{ij}^{*} = \hat{\gamma}_{00} + \hat{\gamma}_{10}(x_{ij}) + \hat{u}_{0j}^{*} + \hat{u}_{1j}^{*}(x_{ij}) + \hat{e}_{ij}^{*} \quad \forall i, j$
7) Refit the model, compute the bootstrap value $\hat{\theta}_b$ on the generated bootstrap sample, and store its value.

[1] The procedure to reflate the residuals is illustrated here for the level-2 residuals but it can be applied to all levels. This is the method described in Goldstein (2010, pp. 99-101) and reproduced here.
First, rewrite model in Equation (49) as $y_{ij} = (X\beta)_{ij} + (ZU)_j + e_{ij}$; where $U^T = \{U_0, U_1, \dots\}$. After fitting the model, residuals are calculated by $\hat{U} = \{\hat{u}_0, \hat{u}_1, \dots\}$. Then write the empirical covariance matrix of the estimated residuals at level-2 in $U^T = \{U_0, U_1, \dots\}$ as $S = (\hat{U}^T\hat{U})/J$ and the corresponding covariance matrix of the level-2 random coefficients estimated from the model as R. Then transform the residuals using $\hat{U}^* = \hat{U}A$, where A is an upper triangular matrix of order equal to the number of random coefficients at level-2, such that $\left(\hat{U}^{*T}\,\hat{U}^*\right)/J = A^T\hat{U}^T\hat{U}A = A^TSA = R$. The new set of residuals $\hat{U}^*$ now have covariance matrix equal to that estimated from the model. The set of residuals in step (4) are then re-sampled from $\hat{U}^*$. A similar procedure is carried out for the level-1 residuals in e_{ij}.

8) Repeat steps 1−7 B times as to obtain B sets of bootstrap replications of the parameter(s).

Van der Leeden et al. (2008) caution about the use of the residual bootstrap, especially when residuals are not estimated in a satisfactory way or are not independent of the predictors in the model. On the other hand, the nonparametric residual bootstrap is robust with respect to non-normality of the error processes since it does not make any assumptions about the error distribution. It also seems to provide better confidence interval coverage compared with the parametric bootstrap when the underlying distribution of the data is non-normal (Carpenter et al., 2003).

Case resampling bootstrap. The case-resampling bootstrap is a nonparametric approach where samples (cases) are randomly drawn before fitting the model. This method is the one that most closely resembles the basic bootstrap algorithm. It has the least restrictive assumptions of all three bootstrap approaches considered here. Specifically, it only assumes that the nested structure in the data is correctly specified and that all explanatory variables are random variables.

If we consider the two-level model in Equation (49), there are different approaches to selecting cases to create a bootstrap sample, each with its advantages and drawbacks depending on the nature of the data at hand (Roberts & Fan, 2004):

(i) Draw a sample of N observations with replacement, ignoring the nested data structure.

(ii) Draw a bootstrap sample of n_j observations with replacement from each and every group in the sample data.

(iii) Bootstrap J groups with replacement while selecting all n_j observations in each bootstrapped group.

(iv) Drawn a bootstrap sample of J groups with replacement, then, from each sampled group, draw a bootstrap sample of n_j observations with replacement.

Approach (i) samples level-1 units directly and retains the overall number of observations in the sample, N, but it leads to variable numbers of groups and observations per group and may alter the correlation structure in the dataset. Approach (ii) also provides a consistent sample size of N for each bootstrap iteration, and it might make sense if the level-2 unit is not a random sample from a population of groups. Approaches (iii) and (iv) will lead to variable number of observations N in the bootstrap samples. The third approach retains the nested structure of the data and makes sense if the level-1 units are not exchangeable, like repeated measures within a patient. Finally, method (iv) might be appropriate if both levels can be considered random samples from the population.

Other bootstrap methods. Another bootstrap method that has been adapted for multilevel models is the wild bootstrap originally proposed by Wu (1986). This approach resamples residuals from an external distribution satisfying certain specifications and is supposed to obtain consistent estimators for the model when the errors are heteroscedastic. Modugno and Giannerini (2015) proposed a modified version of wild bootstrap for multilevel models that, similarly to the case resampling method, does not require homoscedasticity and makes no assumptions about the distribution of the error processes. The authors compared this new procedure with the traditional methods presented above in a simulation study and concluded that the wild bootstrap is preferred under heteroscedasticity and if sample sizes are large. However, since this study is specifically interested in cases where the sample sizes at the lowest levels are not large, this method will not be investigated further.

Summary of bootstrapping for multilevel models

Van der Leeden et al. (2008) observe that cases bootstrap estimators are usually less efficient than those from parametric residuals bootstrap exactly because they work under weaker assumptions. The authors mention that, for instance, the cases bootstrap method is consistent under heteroscedasticity. Therefore, it provides robustness at the expense of efficiency. The authors mention different scenarios under which it makes sense to resample units from all levels of the model or from only level 1 or 2. According to van der Leeden and colleagues, two main factors will determine the best approach for case resampling: the degree of randomness of the sampling at both levels and the average sample size at each level (2008, pp. 413-414). Roberts and Fan (2004) argue that case-resampling approaches (i) and (ii) listed above are preferred because the sampling distribution is defined for a specific sample size, hence a consistent sample of size N is needed to construct an empirical sampling distribution for a statistical estimator of interest.

Goldstein (2010) contends that, if model assumptions are plausible, the parametric bootstrap is preferred, particularly if models are complex. Van der Leeden et al. (2008) hypothesize that the cases bootstrap might be more sensitive to the problematic effects of small sample size but also that it is the most attractive due to relying on the least number of assumptions and leading to consistent estimators if the cases resampling scheme is appropriate for the data.

The primary goal of the bootstrap method is to generate a distribution that is a close approximation of the true distribution of the original sample. To accomplish this goal, the resampling procedure should closely replicate the "true" data generating process. In this study, the outcome variable is continuous and the level-1 units are repeated measures which cannot be considered random realizations. Therefore, the cases bootstrap will be used where the level-2 units (e.g., persons) are randomly selected but all corresponding level-1 units are included in the sample.

CHAPTER 3. METHODS

Study Overview

A simulation study was conducted to evaluate the performance of dominance analysis in determining the relative importance of predictors in multilevel models for longitudinal data. Specifically, this study investigated the suitability of DA for linear growth models where individuals are assumed to differ in their initial status of the (continuous) outcome variable. Dominance analysis was used to assess the relative incremental contribution of both time-varying (level-1) and time-invariant (level-2) predictors using various measures of model fit for multilevel models. Longitudinal models can be used to depict change over time in either continuous (e.g., student achievement, blood pressure, weight) or categorical (e.g., student proficiency, high blood pressure, obesity) outcome measures, and researchers might want to use DA for rank-ordering the factors that influence the direction and rate of change of such outcomes. This dissertation focuses on continuous responses; therefore, simulation conditions were used to generate continuous longitudinal data and the performance of DA was evaluated for the corresponding linear multilevel models.

The investigation of inferential procedures in this study focused on determining the *general* dominance relationships among *p* predictors in multilevel models for longitudinal data, similarly to Azen and Traxel (2009). The general dominance measure of a predictor, X_i, is denoted by G_i and reflects an overall (weighted) average of the additional contribution of the predictor across all subset models of interest. This is a quantitative measure of dominance that can be easily understood and will almost always allow the establishment of a dominance relationship between two predictors, making it a convenient and informative summary statistic of the relative importance of

one predictor over another. The difference between the quantitative general dominance measures of two predictors (X_i and X_j) is defined as:

$$G_{ij} = G_i - G_j \tag{50}$$

where G_i is the general dominance measure for predictor X_i and G_j is the general dominance measure for predictor X_j with $i \neq j = 1, 2, \ldots, p$. As described earlier, each general dominance measure is an average of the additional contributions of a predictor to the fit of all relevant subset models (by model size). Additionally, the general dominance relationships between X_i and X_j can be defined qualitatively by D_{ij} such that:

$$D_{ij} = \begin{cases} 1, & \text{if } G_i > G_j \\ -1, & \text{if } G_i < G_j \\ 0, & \text{if } G_i = G_j. \end{cases}$$

This categorical indicator of the general dominance relationship will be used to investigate the reproducibility of the general dominance results. The reproducibility provides an indication of how confident one can be that the dominance relationship found in the sample reflects the population dominance relationship.

Research Questions

The simulations investigated the performance of DA for longitudinal multilevel models with continuous outcomes. The general research questions investigated are what effects do different levels of (i) model complexity, (ii) predictor coefficients, (iii) sample sizes, (iv) collinearity, (v) covariance structure misspecification, and (vi) measures of model fit, have on:

1. Rank-ordering of the predictors in terms of their relative importance;

2. Inferential results for the quantitative dominance measure (G_{ij}), including type I error, power, and accuracy of estimation, using asymptotic normal (standard error) and percentile confidence intervals;
3. Reproducibility of the qualitative dominance measure (D_{ij}), including the expected level of reproducibility for a given population dominance effect.

In the sections that follow, details are provided on the conditions investigated as well as the evaluation of and expectations regarding the above outcomes.

Simulation Conditions

Two-level models, representing measurement occasions at level 1 and individuals at level 2, were used to generate the data. All models reflect a linear effect of time on the outcome through a time main effect variable. A simple linear trend was used because it represents a basic growth model and a large proportion of applications of growth models use linear models (Kwok et al., 2008). Non-linear time trends modeled through the inclusion of higher-order functions of the time variable (i.e., quadratic: time^2, cubic: time^3) could also be used. However, since time and its functions are held constant in the models when performing DA, the inclusion of the functions of time should not affect the DA results. That is, DA is used to compare predictors other than time in these models.

The conditions manipulated include model complexity, predictor effects (coefficients), number of level-1 units (measurement occasions), number of level-2 units (e.g., students), and amount of collinearity between the predictors. The choice of models and covariance structures represented in this simulation study was driven by their prevalence and parsimony. The models reflect commonly used growth models and are also simple enough so that this initial evaluation of dominance analysis can be performed within a set of well-understood models. Collinearity is

introduced in the models by allowing the predictors to correlate with each other using a correlation parameter ρ_{jk} (set at the same value for all pairs of predictors j,k where $j \neq k$), which is varied according to the *collinearity* simulation condition. Additionally, the impact of misspecification of the covariance structure (of the level-1 residuals) on DA results is investigated. Data for all models are generated assuming a linear time trend that can vary across individuals, in terms of both its intercept and slope, and correlation between the repeated measures is modeled by generating the level-1 residuals using a first-order autoregressive (AR(1)) covariance structure. The various simulation conditions are discussed in detail in this section.

Model complexity. Three models of increasing complexity were used to investigate the relative importance of both time-invariant (level-2/between-subjects) and time-varying (level-1/within-subjects) predictors. The number of predictors in each model was chosen to allow for different combinations of pairwise comparisons among predictors at different levels of analysis (person or time) and different magnitudes of effect size. Additionally, models with four predictors are commonly found in program evaluation research for example, while models with larger number of predictors can be found in exploratory research using data from large, federally funded longitudinal studies. Table 3 lists the combined equations for each model complexity condition. All models contain the terms $+ u_{0i} + e_{ti}$ for the random effects (not explicitly shown in Table 3), and the models vary in terms of the fixed effects (predictors) at levels 1 and 2. The description and rationale for each model are presented next.

Table 3 Model complexity conditions.

	Equations for $y_{ti}\ \ (+ u_{0i} + e_{ti})$				
Model	**Intercept + Time effect**	**Level-1 predictors**	**Level-2 predictors**	**Interaction terms**	**Random Slope effect**
Predictors of random intercept (Model 1)	$\gamma_{00} + \gamma_{10}Time_{ti} +$		$\sum_{h=1}^{4} \gamma_{0h} w_{hi}$		$+u_{1i}Time_{ti}$
Predictors of Time effect (Model 2)	$\gamma_{00} + \gamma_{10}Time_{ti} +$		$\sum_{h=1}^{4} \gamma_{0h} w_{hi}$	$+ \sum_{h=1}^{4} \gamma_{1h} w_{hi} Time_{ti}$	$+u_{1i}Time_{ti}$
Time-varying predictors (Model 3)	$\gamma_{00} + \gamma_{10}Time_{ti} +$	$\sum_{g=2}^{5} \gamma_{g0} x_{(g-1)ti} +$	$\sum_{h=1}^{4} \gamma_{0h} w_{hi}$		$+u_{1i}Time_{ti}$

- **Model 1:** The *growth model with (time-invariant) predictors of the random intercept* includes four level-2 (student-level, time-invariant) explanatory variables as (fixed) predictors of the random intercept, and fixed and random slopes for the effect of time on the outcome. The presence of a random effect (slope) of time means that the rate of change (effect of time on the outcome) is allowed to vary across individuals. This model was used to investigate DA for longitudinal models where interest is in the relative importance of explanatory variables that are time-invariant (i.e., variables measured at baseline).

- **Model 2:** The *growth model with (time-invariant) predictors of the random intercept and of the time effect* includes person-level (time-invariant) variables as predictors of both the random intercept and of the effect (slope) of time, the latter represented in the model as cross-level interactions between the person-level predictors and time. DA for this model compared the relative importance of person-level predictors when modeled as main effects

and as predictors of the time effect. Cross-level interactions are included in this model to represent situations where one is interested in investigating whether the effect of time (rate of change) on the outcome can be predicted by characteristics of the student such as gender or baseline SES. The dominance relationships of these cross-level interactions are determined using constrained DA to compare the relative importance of the interaction terms after controlling for the main effects. Therefore, comparisons between a cross-level interaction and its corresponding main effect were not considered.

- **Model 3:** The *growth model with time-varying predictors* includes four time-varying (level-1) predictors, which change with time, in addition to the time-invariant (level-2) predictors of the intercept. The time-varying predictors are added to reflect situations where there is interest in including predictors that vary across time and in evaluating their relative importance. Following the example previously presented, the researcher might want to investigate the effects of predictors that could change over time, such as the number of books read in the past year, time-varying social skills, expressive vocabulary skill, and verbal memory, on reading comprehension scores after accounting for the effect of time (and the other time-invariant predictors).

Table 4 lists the number of general dominance pairwise comparisons (G_{ij}) examined in each model complexity condition. The comparisons are categorized according to the level of the predictors in the pair: L2 represents a person-level predictor, L1 represents a time-varying predictor, and IN represents the cross-level interaction between a level-2 predictor and the Time trend effect. The number of pairwise comparisons in a model with p predictors is $p \times (p - 1)$.

However, for model 2, the (four) comparisons between a main effect and their own interaction terms were not considered; therefore, this model has $\big(p \times (p-1)\big) - 4$ G_{ij} pairs.

Table 4 Number of population general dominance difference measures (G_{ij}) by model complexity and predictor type.

	Predictor Type					
	L2-L2	**L2-IN**	**IN-IN**	**L2-L1**	**L1-L1**	**Total**
Model 1	6					6
Model 2	6	12	6			24
Model 3	6			16	6	28

Covariance structure misspecification. The data for the simulation were generated using the models just described and the AR(1) structure for the residual covariance matrix. To evaluate any potential effects of misspecification at the estimation stage (which often occurs in practice), the models were estimated (i.e., by fitting the data) under two different covariance structures:

1. Growth model and AR(1) residual structure (correct specification): the random intercept and random slopes of time are modeled with an unstructured covariance matrix, $\boldsymbol{G_i} = \begin{bmatrix} \tau_0^2 & \tau_{01}^2 \\ \tau_{01}^2 & \tau_1^2 \end{bmatrix}$, and the level-1 residuals are modeled with a first-order autoregressive structure: $\boldsymbol{R_i} = \sigma^2[\boldsymbol{AR}(\mathbf{1})]$. This model correctly specifies both the level-1 and level-2 covariance structures and will be referred to as "GAR" in this study.
2. Standard growth model structure (misspecification): the random intercept and random slopes of time are modeled with an unstructured covariance matrix, $\boldsymbol{G_i} = \begin{bmatrix} \tau_0^2 & \tau_{01}^2 \\ \tau_{01}^2 & \tau_1^2 \end{bmatrix}$, but the level-1 residuals are assumed to have an identity structure, $\boldsymbol{R_i} = \sigma^2 \boldsymbol{I_{n_i}}$. This model misspecifies the level-1 covariance structure. This specification is equivalent to

what Singer and Willett (2003) call the "standard multilevel model for change" and will be referred to as "SGR" in this study.

Predictor effects. Predictor fixed effects (model coefficients) were chosen to reflect a range from small (0.1) to large (0.8) as well as no effect (0). These are essentially the values of the standardized coefficients since all predictor variables are generated from a standard normal distribution. Differences in the magnitude of coefficients were chosen to investigate the sensitivity of the DA procedure for ordering predictors according to their absolute and relative effects. Combinations of within- and between-subjects effects (after controlling for the linear effect of time), as shown in Table 5, were investigated for a total of 9 model type and predictor effect combinations. The population fixed intercept, γ_{00}, was set to 1 for all conditions to represent the value of the outcome when time and all other predictors are zero. The fixed effect of time (γ_{10}) was fixed at 0.5 for all conditions to represent a moderate linear effect of time on the outcome, and is similar to values chosen by other longitudinal simulation studies (Jaeger et al., 2017). The variance of the level-1 random intercept (i.e, intercepts of individual growth models) was set to V(u_{0i}) = τ_0^2 = 0.4, the variance of the random slope of time (*Time*$_{ti}$, or the linear growth trends of individual growth models) was set to V(u_{1i}) = τ_1^2 = 0.2, and the covariance between the individual intercepts and linear growth trends (i.e., slope) was set to Cov(u_{0i}, u_{1i}) = τ_{01}= 0.1. These values were chosen following a study by Kwok, West, and Green (2007) to represent a strong clustering effect and a large variation among subjects in terms of both their outcome measures at time zero (arbitrarily chosen as the intercept) and the linear growth trend (slope of Time variable) in addition to an appropriate covariance between the two. For all models, the level-1 errors e_{ti} were generated with a first-order autoregressive model such that $e_{ti} = \phi_i e_{(t-1)i} + w_{ti}$, where w_{ti} is the independent and

identically distributed (*i.i.d.*) error, distributed $N(0,\sigma^2)$, with σ^2 set to 1 following similar simulation studies (Ferron, Dailey, & Yi, 2002; Kwok et al., 2007). The autocorrelation between observations measured at time *t* and *t*-1 is modeled by the first order autoregressive parameter ϕ, which was set at 0.3 at the population level as this represents a moderate level of autocorrelation (Ferron et al, 2002).

Table 5 Model complexity by predictor effect conditions.

Model	**Effect**	**Person-level (L2)**				**Time (L1)**	**Time-Varying (L1)**				**Interaction (IN=L2*$Time_{ti}$)**			
		w_{1i}	w_{2i}	w_{3i}	w_{4i}	$Time_{ti}$	x_{1ti}	x_{2ti}	x_{3ti}	x_{4ti}	$w_{1i}(T_{ti})$	$w_{2i}(T_{ti})$	$w_{3i}(T_{ti})$	$w_{4i}(T_{ti})$
Model 1:		γ_{01}	γ_{02}	γ_{03}	γ_{04}	γ_{10}	γ_{20}	γ_{30}	γ_{40}	γ_{50}	γ_{11}	γ_{12}	γ_{13}	γ_{14}
Predictors	Base	.3	.3	.1	.1	.5								
of random	Small	.5	.45	.4	.3	.5					n/a			
intercept	Large	.8	.6	.4	.2	.5								
Model 2:	Base-Base	.3	.3	.1	.1	.5					.3	.3	.1	.1
Predictors of	Base-Large	.3	.3	.1	.1	.5		n/a			.8	.6	.4	.2
Time effect	Large-Large	.8	.6	.4	.2	.5					.8	.6	.4	.2
Model 3:	Base-Base	.3	.3	.1	.1	.5	.3	.3	.1	.1				
Time-varying	Base-Large	.3	.3	.1	.1	.5	.8	.6	.4	.2				
predictors	Large-Large	.8	.6	.4	.2	.5	.8	.6	.4	.2				

In Table 5, the base(line) condition, where some but not all pairs of predictors have the same fixed effects, was designed to try and investigate both type I error and power rates for testing the general dominance relationships (i.e., testing the null hypothesis H_0: $G_{ij} = 0$). Since the general dominance measures cannot be simulated directly, predictor fixed effects were used here as a way to manipulate the general dominance effect. Therefore, the predictor effects were varied to different extents to investigate how the power of the DA procedure may change under small to large differences in dominance effects for different combinations of fixed effects and their interaction with collinearity. It is expected that the power to detect dominance will increase with the dominance effect size, which corresponds to differences in predictor effects when the predictors are independent but can vary substantially when collinearity is present. Actual population general dominance values corresponding to the fixed effects and collinearity conditions are presented in the results section.

Collinearity. Since relative importance measures are particularly informative when predictors are correlated, three degrees of collinearity (correlation ρ_{jk} among the predictors j and k, with $j \neq k$) are considered: no collinearity ($\rho_{jk} = 0$); medium collinearity ($\rho_{jk} = 0.5$); and high collinearity ($\rho_{jk} = 0.8$). The high collinearity condition is investigated only for model 1 as it is not expected that this condition will be prevalent in practice. When predictors are correlated, importance measures that are calculated by "holding all other predictors constant" might be misleading when comparing predictors to each other since each predictor's effect will be affected by both the predictors included as well as the predictors excluded from the model (Azen & Budescu, 2003). Therefore, in order to identify the impact of collinearity on DA results, predictors were simulated to have different degrees of collinearity. The no-collinearity condition is unrealistic and used here

as a baseline. The collinearity value of 0.5 is an amount of collinearity that can be expected to be found in real datasets. The presence of collinearity is expected to impact dominance in the sense that dominance relationships might not be as clear cut and will not be directly related to predictor coefficients as in the case with uncorrelated predictors.

Sample size. To account for situations typical of longitudinal designs, the level-1 sample size, n_{L1}, was set at 4 and 8 to reflect the number of measurement occasions or waves in longitudinal data. Sample size at level 2 (n_{L2}) was set at 50 and 200 for models 2 and 3, and a larger sample size of 1000 was investigated for model 1. These sample sizes reflect datasets from small to large in terms of number of subjects and were based on values used in previous simulation studies of longitudinal multilevel data (Matuszewski, 2011; Jaeger, 2017; Jaeger et al., 2017). Sample size conditions were fully crossed within each model, producing 4 sample size combinations for models 2 and 3, and 6 combinations for model 1. DA results were expected to be more accurate and less biased when sample sizes were larger, especially at the person level, but even under small samples DA was expected to produce accurate measures of relative importance.

Table 6 provides a summary of all simulation conditions per model complexity. The covariance structure misspecification is not listed in this table because it is not technically a design factor in the sense that only one matrix (growth model with AR(1) residuals) was used for data generation. However, each individual condition combination was estimated twice, once assuming the true (GAR) covariance structure and once assuming the simplified (SGR) structure, which effectively doubled the number of simulation conditions listed in Table 6.

Table 6 Summary of all simulation conditions and levels by model complexity.

Model	Predictor effects			Occasions (n_{L1})		Subjects (n_{L2})			Collinearity (ρ_{jk})			Number of conditions
	1	**2**	**3**	**4**	**8**	**50**	**200**	**1000**	**0.0**	**0.5**	**0.8**	
1	x	x	x	x	x	x	x	x	x	x	x	54
2	x	x	x	x	x	x	x		x	x		24
3	x	x	x	x	x	x	x		x	x		24
All												102

Simulation Study – Procedure

The simulation study procedure consisted of the following steps:

(1) Generate a pseudo-population according to the model complexity, predictor effects and collinearity conditions, and record the obtained parameters;

(2) Obtain S = 100 simple random samples (SRS) from each pseudo-population;

(3) Obtain *B* = 300 bootstrap samples for the S=100 randomly selected (parent) samples from each pseudo-population;

(4) Perform dominance analysis on all of these samples;

(5) Collect relevant outcome measures; and

(6) Evaluate these measures in the context of the simulation conditions.

As steps (2) and (3) indicate, two different sampling methods were used to obtain samples for each condition: simple random sampling (SRS) and bootstrap sampling. SRS was used to evaluate the dominance analysis procedure more directly because in this study the population is known and random samples can be selected from it. However, this is not a realistic or feasible situation in practice. Therefore, SRS is used only to demonstrate the procedure's theoretical applicability. In practice, researchers will most likely have only one, presumably random, sample

from the population. Therefore, the bootstrap method is used to study the performance of DA and related inferential procedures when only one random sample is obtained from the population. The number of bootstrap samples was determined based on empirical considerations. A subset of the simulation conditions was run with both larger and smaller number of bootstraps and it was determined that 300 subsamples produced sufficiently accurate results. In order to evaluate the confidence intervals created by the bootstrap procedure, S=100 replications of the bootstrapping method were performed and averages or proportions across all replications were calculated. Pseudo-population generation, the two sampling methods, and the evaluation methods are described next.

Generating the pseudo-population. Model complexity, predictor effects, and collinearity conditions were used to generate a pseudo-population with number of level-1 and level-2 units equal to N_{L1}= 4 or 8 (number of measurement occasions) and N_{L2} = 100,000 (number of cases or individuals), respectively. The number of measurement occasions in the sample is the same as in the population. The model used to generate the data follows from equations (5) to (7), which are repeated below:

$$\underset{(n_i \times 1)}{\boldsymbol{y_i}} = \underset{(n_i \times (p+1))}{\boldsymbol{X_i}} \underset{((p+1) \times 1)}{\boldsymbol{\beta}} + \underset{(n_i \times (q+1))}{\boldsymbol{Z_i}} \underset{((q+1) \times 1)}{\boldsymbol{u_i}} + \underset{(n_i \times 1)}{\boldsymbol{e_i}} \tag{5}$$

$$\begin{pmatrix} \boldsymbol{u_i} \\ \boldsymbol{e_i} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R_i} \end{pmatrix}\right] \tag{6}$$

$$Var(\boldsymbol{y_i}) = \boldsymbol{V_i} = \boldsymbol{Z_i G Z_i'} + \boldsymbol{R_i} \tag{7}$$

In these equations, $1 \leq i \leq N_{L2}$ and $t = 1, \ldots, (n_i = n_{L1})$. The simulation conditions described previously determined the parameters corresponding to the elements in $\boldsymbol{X_i}$, $\boldsymbol{\beta}$, $\boldsymbol{Z_i}$, **G**, and

$\mathbf{R}_i$. Matrix $\boldsymbol{X_i}$ contains the values of all fixed predictors, matrix $\boldsymbol{\beta}$ contains the fixed coefficients obtained from the *predictor effects* conditions (i.e., γ values in Table 5), matrix $\boldsymbol{Z_i}$, contains the design matrix of the random effects (in this case the random intercept and the random slope of *Time*), $\boldsymbol{G}$ is the covariance matrix of the (level-2) random effects, and $\boldsymbol{R_i}$ is the covariance matrix of the level-1 residuals. Data for all models were generated with a first-order autoregressive covariance structure for the level-1 residuals to reflect the realistic scenario where measurements that are closer to each other in time are more correlated than measurements that are farther apart in time. For each of the models in Table 3, data for a continuous outcome *y* for student *i* was generated using the configurations detailed below, shown using the configuration for the lowest level-1 sample size condition ($n_{L1} = 4$). That is, for model 1 **(**growth model with predictors of the random intercept), with 4 level-2 predictors and 4 measurement occasions, the matrices would be:

$$\boldsymbol{X_i} = \begin{bmatrix} 1 & Time_1 & w_{1i} & w_{2i} & w_{3i} & w_{4i} \\ 1 & Time_2 & w_{1i} & w_{2i} & w_{3i} & w_{4i} \\ 1 & Time_3 & w_{1i} & w_{2i} & w_{3i} & w_{4i} \\ 1 & Time_4 & w_{1i} & w_{2i} & w_{3i} & w_{4i} \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{01} \\ \gamma_{02} \\ \gamma_{03} \\ \gamma_{04} \end{bmatrix}$$

$$\boldsymbol{Z_i} = \begin{bmatrix} 1 & Time_{1i} \\ 1 & Time_{2i} \\ 1 & Time_{3i} \\ 1 & Time_{4i} \end{bmatrix}; \quad \boldsymbol{u_i} = \begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix}; \quad \boldsymbol{e_i} = \begin{bmatrix} e_{1i} \\ e_{2i} \\ e_{3i} \\ e_{4i} \end{bmatrix}$$

$$\boldsymbol{G} = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix}; \qquad \boldsymbol{R_i} = \sigma^2 \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{bmatrix}$$

The w_1, w_2, w_3, w_4 predictors were generated from a multivariate normal distribution with mean zero, standard deviation of one and correlation among predictors equal to the *collinearity* condition values $\rho(w_j, w_k) = (0, 0.5, 0.8),\ \forall\, j, k \in 1, 2, 3, 4\ with\ j \neq k$. The $Time_{ti}$ variables

were generated by setting $Time_{ti} = t - 1$ for t = 1, 2, …,n_{L1} for all i. The level-1 residuals, e_{ti}, were generated following a first-order autoregressive process, $e_{ti} = \phi_i e_{(t-1)i} + v_{ti}$, where $v_{ti} \sim N(0, \sigma^2)$, corresponding to $\boldsymbol{R_i} = Var(e_{ti}) = \sigma^2[\boldsymbol{AR(1)}]$ with σ^2 =1 and autoregressive parameter ϕ = 0.3. The random effects $\boldsymbol{u_i}$ were generated from a multivariate normal distribution with mean zero and covariance matrix **G**:

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \tau_0^2 = .4 & \tau_{01} = .1 \\ \tau_{01} = .1 & \tau_1^2 = .2 \end{bmatrix}\right)$$

Similarly, for model 2 (growth-model with predictors of the random intercept and of the time effect), the matrices would be:

$$\boldsymbol{X_i} = \begin{bmatrix} 1 & Time_1 & w_{1i} & w_{2i} & w_{3i} & w_{4i} & w_{1i}Time_1 & w_{2i}Time_1 & w_{3i}Time_1 & w_{4i}Time_1 \\ 1 & Time_2 & w_{1i} & w_{2i} & w_{3i} & w_{4i} & w_{1i}Time_2 & w_{2i}Time_2 & w_{3i}Time_2 & w_{4i}Time_2 \\ 1 & Time_3 & w_{1i} & w_{2i} & w_{3i} & w_{4i} & w_{1i}Time_3 & w_{2i}Time_3 & w_{3i}Time_3 & w_{4i}Time_3 \\ 1 & Time_4 & w_{1i} & w_{2i} & w_{3i} & w_{4i} & w_{1i}Time_4 & w_{2i}Time_4 & w_{3i}Time_4 & w_{4i}Time_4 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{01} \\ \gamma_{02} \\ \gamma_{03} \\ \gamma_{04} \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{14} \end{bmatrix}$$

The $w_1,\ldots, w_4$, $Time_{ti}$ predictors, $\boldsymbol{Z_i}$, $\boldsymbol{u_i}$, $\boldsymbol{e_i}$, matrices and the $\boldsymbol{G}$ and $\boldsymbol{R_i}$ covariance structures were generated as the previous model. The $w_1Time,\ldots, w_4Time$ predictors were generated by multiplying the $w_1,\ldots, w_4$ and *Time* variables.

Finally, for model 3 (growth-model with time-varying predictors), the matrices are:

$$\boldsymbol{X_i} = \begin{bmatrix} 1 & Time_1 & x_{11i} & x_{21i} & x_{31i} & x_{41i} & w_{1i} & w_{2i} & w_{3i} & w_{4i} \\ 1 & Time_2 & x_{12i} & x_{22i} & x_{32i} & x_{42i} & w_{1i} & w_{2i} & w_{3i} & w_{4i} \\ 1 & Time_3 & x_{13i} & x_{23i} & x_{33i} & x_{43i} & w_{1i} & w_{2i} & w_{3i} & w_{4i} \\ 1 & Time_4 & x_{14i} & x_{24i} & x_{34i} & x_{44} & w_{1i} & w_{2i} & w_{3i} & w_{4i} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{20} \\ \gamma_{30} \\ \gamma_{40} \\ \gamma_{50} \\ \gamma_{01} \\ \gamma_{02} \\ \gamma_{03} \\ \gamma_{04} \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{14} \end{bmatrix}$$

The $w_1, \ldots, w_4$, $Time_{ti}$ predictors, $\boldsymbol{Z_i}$, $\boldsymbol{u_i}$, $\boldsymbol{e_i}$, matrices and the $\boldsymbol{G}$ and $\boldsymbol{R_i}$ covariance structures were generated as in model 1. The x_{jti} predictors were generated from a multivariate normal distribution with mean zero, standard deviation of one and correlation among predictors equal to $\rho(x_{jti}, x_{kti}) = (0, 0.5, 0.8),\ \forall\, j, k \in 1, 2, 3, 4\ with\ j \neq k.$

Note that, since in this study the lower level of the model (level-1) corresponds to time points, the number of level-1 measurements in the sample can be assumed to be the same as the number in the population (except for missing values). This is consistent with van der Leeden et al. (2008), who do not consider repeated measures as being random for purposes of bootstrapping. Once the data were generated, parameter values and population, or "true", dominance results (G_{ij} and D_{ij} values for each pair of predictors) were obtained from the pseudo-population.

Simple random sampling. For each pseudo-population, S=100 simple random samples (SRS) were obtained with level-1 and level-2 sample sizes determined by the n_{L1} and n_{L2} sample size conditions, respectively. To start, n_{L2} persons were randomly drawn from the pseudo-population and then all corresponding n_{L1} measurements were obtained for these individuals. The samples were drawn without replacement. In these samples, all $\binom{N_{L2}}{n_{L2}}$ samples have an equal

probability of being selected from the population. For each SRS sample *s*, the dominance measures G_{ij}^{s} and D_{ij}^{s}, were computed for each pair ($i \neq j = 1, 2, \ldots, p$) of predictors so they can be compared to those obtained in the population.

Bootstrap sampling. Since in real-world settings researchers do not have access to the population data, the use of bootstrap methods was investigated for making inferences regarding dominance analysis measures based on one random "parent" sample. However, in this study, since we want to evaluate the performance of bootstrap-based inferential procedures, this step was replicated S=100 times.

The bootstrap is a nonparametric approach for statistical estimation and inference based on intensive computer-based resampling that does not make any assumptions regarding the distribution of the data (Efron, 1979). In this study, the parent sample for each condition was created by following the SRS procedure described above to draw one simple random sample from the pseudo-population according to the sample sizes specified by n_{L1} and n_{L2}. Subsequently, B=300 bootstrap samples were drawn randomly and with replacement from the parent sample, treating person as random. The lowest level, level-1, is not considered random since in this study this level corresponds to repeated measurements (van der Leeden et al., 2008). Therefore, once a level-2 unit (person) was sampled, all corresponding level-1 units (measurement occasions for that person) were also included in the bootstrap sample.

The case-resampling bootstrap method was used for all simulation conditions; that is, the entire response vector (outcome and predictor values) was drawn with replacement from the level-2 units. In each bootstrap sample *b* (corresponding to a given parent sample), the sample estimate values of G_{ij}^{b} and D_{ij}^{b} were computed for each pair ($i \neq j = 1, 2, \ldots, p$) of predictors.

Estimation. For each generated data set (pseudo-population, SRS, or bootstrap sample), all subset multilevel models were fit using SAS PROC MIXED to obtain model estimates and measures of fit. The models were estimated using the FIML method (METHOD=ML in SAS) since the DA procedure is comparing the fit of models with different fixed parameters.

Measures of fit. The dominance relationships among predictors (i.e., the G_{ij} and D_{ij} values) were determined using four R^2 analogue measures: Nakagawa and Schielzeth's (2013) marginal R^2 (henceforth N&S R^2), Edwards et. al.'s (2008) R^2_β (henceforth R^2 Beta), McFadden's (1974) R^2_M (henceforth McFadden R^2), and the proportion change in variance (PCV) proposed by Raudenbush and Bryk (2002). Since the PCV is calculated separately for each random component, the random intercept PCV(u_{0i}) is used for the models with predictors of the random intercept and the cross-level interaction (models 1 and 2; henceforth R&B2 R^2), and the level-1 residual PCV(σ^2) is used for the time-varying predictors model (model 3; henceforth R&B1 R^2).

Relative importance results for person-level (level-2) predictors of the random intercept should be adequate when using proportional change in variance of the random intercept, i.e., R&B2 R^2. The use of this measure for time-varying and cross-level comparisons might be more problematic unless these predictors also help explain part of the variability in the random intercept. For the model with time-varying predictors, global R^2 measures such as the N&S R^2 and the likelihood-based McFadden R^2 might provide more relevant information and more accurate and precise dominance results.

Dominance analysis evaluation measures. The performance of DA was evaluated by calculating ranking accuracy, bias, 95% confidence intervals, and reproducibility, as described

below. These measures were evaluated for samples obtained under both the simple random and (replicated) bootstrap sampling methods.

Ranking accuracy. Evaluation of whether dominance analysis accurately rank-orders predictors in terms of relative importance was determined by computing:

1) the proportion of samples where there is agreement between the population and simple random or bootstrap samples as to:
   i. the predictor identified by DA as *most* important, and
   ii. the predictor identified by DA as the *least* important.

These outcomes were calculated using the general dominance measure G_i for each predictor in each bootstrap and simple random sample. First, the predictors were rank ordered by the value of G_i in each sample. The predictor with the highest G_i value was considered the most important predictor, and the predictor with the lowest G_i value was considered the least important predictor. Two or more predictors could be tied as most or least important. In this case they were saved as a set of most or least important predictors for that sample. These predictors were then compared to the predictor(s) rank ordered by the G_i measure as most or least important in the population. If the predictor ranked as most/least important in the bootstrap sample was the same as in the population, or, in the case of a tie, at least one of the predictors in the most/least important set in the bootstrap or SRS matched one or more predictors in the population most/least important set, the agreement outcome for most/least important predictor was recorded as 1 for that sample, otherwise it was recorded as zero. The overall agreement measures in terms of most/least important predictors were then calculated as the proportion of bootstrap and SRS samples that agreed with the population ranking. For bootstrap samples these values were calculated

for each parent sample as well as averaged across all parent samples. These agreement values ranged from 0%, when none of the samples agreed with the population result, to 100%, when all samples agreed with the population result.

2) Kendall rank correlation coefficient (Kendall's tau-b) between the predictor ranking produced by DA in the bootstrap or SRS sample and the population ranking.

Kendall's tau (Kendall's τ; Kendall, 1938) is a nonparametric measure of ordinal association based on the number of concordances and discordances in paired observations or rankings. Kendall (1945) proposed an adjustment for ties, usually called tau-b, which is the statistic used here. To define this coefficient, let (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n) be a set of observations of the joint random variables X and Y respectively. Two pairs of observations (X_i , Y_i) and (X_j , Y_j), where $i \neq j$, are *concordant* if they are in the same order with respect to each variable, that is, if $X_i < Y_i$ and $X_j < Y_j$; or if $X_i > Y_i$ and $X_j > Y_j$. The pairs are *discordant* if they are in the reverse ordering for X and Y, that is, if $X_i < Y_i$ and $X_j > Y_j$; or if $X_i > Y_i$ and $X_j < Y_j$. The pair is *tied* if $X_i = X_j$ and/or $Y_i = Y_j$. The total number of pairs that can be constructed for a sample size of n is $\binom{n}{2} = n(n-1)/2$.

Kendall's tau-b is calculated by:

$$\tau_b = \frac{C-D}{\sqrt{(C+D+T)\times(C+D+U)}} \tag{51}$$

where C is the number of concordant pairs, D the number of discordant pairs, T the number of ties only in X, and U the number of ties only in Y. If a tie occurs for the same pair in both X and Y, it is not added to either T or U. Values of tau-b close to 1 indicate strong agreement, values close to -1 indicate strong disagreement, and a value of zero indicates lack of association.

Kendall's tau-b was calculated for each bootstrap and SRS sample. In this study, X would be the general dominance value G_i^b or G_i^s for each variable X_i in each bootstrap sample or SRS, respectively, and Y would be the general dominance value G_i in the population. The data is double sorted by ranking observations according to values of the first variable (X) and reranking the observations according to values of the second variable (Y). Kendall's tau-b is computed from the number of interchanges of the first variable and corrects for tied pairs (pairs of observations with equal values of X or equal values of Y). Let X be the values of the general dominance measures for all predictors in a bootstrap sample, noted as G_i^b, and Y be the general dominance values in the population, noted as G_i. If the observation (in this case predictors) with the smallest and second smallest dominance values in the bootstrap sample is in the same order as the predictors in the population (e.g., $G_1^b < G_2^b$ and $G_1 < G_2$), then the pair is counted as concordant. If the ordering is reversed, i.e., either $G_1^b < G_2^b$ and $G_1 > G_2$, or $G_1^b > G_2^b$ and $G_1 < G_2$, the pair is counted as discordant. If $G_1^b = G_2^b$, but $G_1 \neq G_2$, then the pair is counted as a tie for X (the bootstrap), if the variables are reversed ($G_1^b \neq G_2^b$, but $G_1 = G_2$), it would count as a tie for Y (the population). For models with four predictors, the total number of pairs is $n(n-1)/2 = 4(4-1)/2 = 6$.

For bootstrap samples, the ranking accuracy evaluation also consisted of how often the DA results across bootstrap samples agreed with the parent sample. Specifically, the indices above were calculated by comparing the bootstrap rankings to the parent sample rankings instead of the population. Average agreement values across all S=100 replications were also calculated. Agreement in the rank-ordering of predictors is expected to improve with larger sample sizes and

larger differences between predictor fixed effects (presuming these factors lead to larger general dominance differences, or G_{ij} values).

Bias. To examine the accuracy of the general dominance estimates, standardized bias was calculated as the difference between a parameter and its sample estimate, standardized by an appropriate standard deviation measure. For the population parameter, G_{ij}, or the parent sample estimate, G_{ij}^{PS}, standardized bias was calculated as:

$$Bias_{SRS,Pop} = \frac{\overline{\mathrm{G}_{ij}^{s}} - \mathrm{G}_{ij}}{sd\left(\mathrm{G}_{ij}^{s}\right)} \tag{59}$$

$$Bias_{Boot,Pop}^{r} = \frac{\overline{\mathrm{G}_{ij}^{b}} - \mathrm{G}_{ij}}{sd\left(\mathrm{G}_{ij}^{b}\right)} \tag{60}$$

$$Bias_{Boot,PS}^{r} = \frac{\overline{\mathrm{G}_{ij}^{b}} - \mathrm{G}_{ij}^{PS}}{sd\left(\mathrm{G}_{ij}^{b}\right)} \tag{61}$$

where $\overline{\mathrm{G}_{ij}^{s}}$ or $\overline{\mathrm{G}_{ij}^{b}}$ represent the average of the general dominance estimates across all relevant simple random or bootstrap samples, respectively, and $sd(G_{ij})$ is the corresponding standard deviation of the estimates. More formally:

$$\overline{\mathrm{G}_{ij}^{s}} = \frac{1}{S}\sum_{s=1}^{S} \mathrm{G}_{ij}^{s} \text{ and } sd\left(\mathrm{G}_{ij}^{s}\right) = \sqrt{\frac{1}{S-1}\sum_{s=1}^{S}(\mathrm{G}_{ij}^{s} - \overline{\mathrm{G}_{ij}^{s}})^2} \tag{62}$$

$$\overline{\mathrm{G}_{ij}^{b}} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{G}_{ij}^{b} \text{ and } sd\left(\mathrm{G}_{ij}^{b}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\mathrm{G}_{ij}^{b} - \overline{\mathrm{G}_{ij}^{b}})^2} \tag{63}$$

For the bootstrap samples, the bias measures were averaged across all S=100 replications:

$$\overline{Bias}_{Boot,Pop}^{r} = \frac{1}{R}\sum_{r=1}^{R} Bias_{Boot,Pop}^{r} \tag{52}$$

$$\overline{Bias}_{Boot,PS}^{r} = \frac{1}{R}\sum_{r=1}^{R} Bias_{Boot,PS}^{r} \tag{53}$$

The standardized bias is used here to determine if the dominance values produced by the bootstrap procedure are close to their corresponding population parameters. This measure was proposed by Rosenbaum and Rubin (1985) in the context of propensity score methodology and is similar to Cohen's *d* because it calculates the standardized mean difference between estimates of general dominance produced by the bootstrap procedure or simple random sampling and the population parameter. Bias values were compared across simulation conditions. When bias is small, we can assume the dominance results adequately reflect the population and/or parent sample values.

Inference. Statistical inference for the estimated difference between the general dominance measures (i.e., G_{ij} values) was carried out by using both asymptotic normal (standard error) and percentile confidence intervals (CI). For the purposes of this study, the SRS CIs were used only as a check on the bootstrap CI. Only the bootstrap CIs were evaluated in terms of coverage, width, type I error, and power rates across all S=100 replications.

Asymptotic normal confidence intervals (ANCI) are usually adequate when the distribution of the parameter of interest is normal or the sample size is sufficiently large. Since this method is straightforward, requiring minimal computation, its suitability for making inferences regarding the difference in magnitude of two general dominance measures was investigated. The asymptotic normal 95% CI for the G_{ij} parameter was constructed for each sample as:

$$CI_{95\%} = \hat{G}_{ij} \pm Z_{.05}s \tag{54}$$

where $\hat{G}_{ij}$ is the general dominance difference estimate for each pair of predictors averaged across all (SRS or bootstrap) samples (i.e., either $\overline{G_{ij}^{s}}$ or $\overline{G_{ij}^{b}}$), $Z_{.05}$ is the value from the standard normal distribution corresponding to the 95% confidence level (i.e., $Z_{.05} = 1.96$), and s is the

standard deviation of all $\hat{G}_{ij}$ (i.e., either $sd(G_{ij}^{s})$ or $sd(G_{ij}^{b})$). This method assumes that the studied statistic, general dominance difference in this case, is normally distributed. Therefore, if the general dominance measures cannot be assumed to be normally distributed, it might not be appropriate (and not perform well) for making inferences about these measures.

The percentile confidence interval (PCI) estimates the percentile points of the confidence interval empirically from the observed distribution of the statistic. Percentile 95% confidence intervals were constructed by ranking the estimated general dominance values, G_{ij}^{s} or G_{ij}^{b} , obtained from all samples (either SRS or bootstrap) and selecting the values corresponding to the 2.5th and 97.5th percentiles as the end points of the confidence interval. This method does not assume that the studied statistic is normally distributed. Bootstrap-based percentile confidence intervals estimate the percentile points of the confidence interval empirically from the observed bootstrap distribution of the statistic, and SRS percentile confidence intervals estimate the percentile points from the observed SRS distribution.

Each of the two types of confidence intervals were used to evaluate the following indicators of estimation accuracy and inferential performance for the bootstrap samples:

1. *CI Coverage.* Confidence interval coverage evaluates whether the constructed 95% CI actually contains the population value about 95% of the time. Coverage was calculated by the proportion of S=100 replications in which the 95% confidence interval for the dominance difference included the true (population) value; that is, $CI_{upper} < G_{ij} < CI_{lower}$. A confidence interval is first-order accurate if the actual one-sided rejection probabilities differ from the nominal values by O($n^{-1/2}$); that is, it covers the true parameter with probability $(100 - \alpha)\% + O(n^{-1/2})$. It is second-order accurate if the differences are O(n^{-1}). The confidence intervals should be at least first-order accurate

(Efron, & Tibshirani, 1993). A second-order accurate interval means that the error in the probability (α/2) of not covering the true value of the parameter from above or below tends to zero at a rate that is inversely proportional to the sample size. On the other hand, first-order accuracy means that the error tends to zero more slowly, at a rate inversely proportional to the square root of the sample size.

2. *CI Width.* The confidence interval width was computed for all simulation conditions by subtracting the lower bound of the confidence interval from its upper bound for each replication and averaging across all S=100 replications:

$$\overline{Width_{CI}} = \frac{1}{R}\sum_{r=1}^{R}(CI_{Upper} - CI_{Lower}) \quad (55)$$

Even if the intervals contain the population G_{ij} value 95% of the time, the method might not be useful if intervals are very wide. It is desirable that the width of a confidence interval be as narrow as possible for a given level of coverage (e.g., 95%). If CI coverage is above $(100 - \alpha)\% + \mathrm{O}(n^{-1/2})$, then the CI is too wide.

3. *Type I Error.* Type I error was calculated as the percentage of the S=100 bootstrap confidence intervals that did not contain zero when the population G_{ij} value was zero. Since 95% confidence intervals were constructed, it is expected that about 5% or the intervals will not include zero when the true value is zero (which is the case for some of the simulation conditions). Type I error should be close to nominal rates but is expected to deviate from it most under the lowest sample size conditions.

4. *Power.* The power of the inferential procedure was calculated as the percentage of the S=100 bootstrap confidence intervals that did not contain zero when the population G_{ij} value was non-zero. Generally, power rates greater than 0.8 are considered adequate. Small sample sizes and small predictor effect differences are expected to result in lower

power (everything else being equal). General dominance difference values (G_{ij}) that are non-zero but small in the population might not be accurately detected by the CI procedures. Bootstrap results was also expected to be impacted by sampling error, especially when the bootstrap parent sample is small and/or differs substantially from the population.

Reproducibility. The SRS reproducibility rate was calculated as the proportion of SRS in which the sample qualitative dominance relationship result, D_{ij}^{s}, agreed with the population value of D_{ij} for each pair of predictors. Similarly, bootstrap reproducibility was calculated as the proportion of bootstrap samples in which the sample qualitative dominance relationship result, D_{ij}^{b}, agreed with the parent sample D_{ij}^{PS} and/or the population D_{ij}. For both sampling methods, the higher the reproducibility the stronger the evidence for the stability and robustness of the dominance relationship result. The reproducibility rates under the no dominance conditions (i.e., when effects are the same for the two predictors being compared, or $D_{ij} = 0$) were used as a baseline to evaluate the conditions of known dominance effects. Reproducibility of the (non-zero) population dominance relationship (i.e., $D_{ij} = 1$ or -1) should be adequate (i.e., above baseline rates) under the bootstrap procedure. Larger sample sizes and larger differences in predictor effects should translate into higher reproducibility. Level of collinearity may also impact reproducibility but perhaps not as severely as it impacts regression coefficients. Reproducibility results under SRS were used as a check for the bootstrap results because, as with CI measures, reproducibility results for bootstrap samples might be negatively impacted if the parent sample does not accurately reflect the population.

CHAPTER 4. RESULTS

Results from the simulation study are presented in this chapter. First, DA results for the pseudo-populations are presented in order to inform the remaining analyses. Since dominance effects cannot be derived directly from the simulation conditions in multilevel models, the population DA results are the actual parameters evaluated in this study. Second, an example of the DA procedure is presented to demonstrate the application of dominance analysis with longitudinal multilevel models in practice. Finally, results from analyzing the simulation data are presented and summarized. The simulation results section includes an examination of the rate of non-positive definite (also referred to in this paper as "npd") random effects covariance matrices (i.e., the G-matrix containing the variances and covariances of the random intercept and random slopes) across conditions. A non-positive definite covariance matrix occurs when the variances within these matrices are estimated to be zero or negative, signaling that there was not enough variation in the response to attribute any variation to the random effect after controlling for all other effects in the model. In instances when npd G-matrices are found, the resulting variance component estimates are not reliable and, therefore, the measures of model fit that depend on these estimates should not be used. Simulation factors that impact the rates of npd G-matrices were also examined.

Analyses of variance (ANOVAs) were conducted to examine the effects of the design factors (simulation conditions) on outcome measures that were produced at the replication level (ranking accuracy, bias, and reproducibility). For each of these outcome measures, an overall factorial ANOVA was conducted for factors that were fully crossed across models (nSubjects=50, 200; nTimePoints=4, 8; Collinearity=0.0, 0.5; Predictor Effects=Baseline, Small, Large; Measures of Fit (R^2)=McFadden, N&S, Beta, R&B). Note that throughout the results section, "nSubjects"

will be used as shorthand for number of level-2 units and "nTimePoints" as shorthand for the number of level-1 units. Since some design factors included in model 1 were not included in models 2 and 3, a separate ANOVA was conducted only for the factors within this model. The effect size (η^2) of each combination of the design factors, representing the proportion of outcome variance explained by each factor combination, was used to determine practical significance. The statistical significance of the ANOVA tests was not used because the sample sizes were large and therefore all effects were significant. The effect size is defined here as $\eta^2 = SS_{factor}/SS_{total}$, where SS_{factor} is the variation corresponding to the factor of interest and SS_{total} is the total variation in the outcome variable (Maxwell & Delaney, 2004). Combinations of factors (i.e., effects) that explained five percent or more of the total outcome variance were further investigated, as this value corresponds to a moderate effect size as suggested by Cohen (1988). Outcome measures related to inferential analyses (confidence interval coverage, width, type I error rates, and power) were analyzed descriptively as these results were reported as proportions or averages of the number of occurrences across replications. Specifically, only one CI was computed per replication. Therefore, the *rates* of coverage, type I error and power were calculated as the proportion of all S replications where these occurred. For instance, the type I error rate for each (population) zero-valued D_{ij} measure was the proportion of all S=100 CIs where the CI erroneously rejected the null hypothesis of no dominance.

Population DA parameters

To put the simulation results into context, the pseudo-population general dominance results are presented first. The pseudo-populations were generated according to model complexity, sample size at level-1 (number of time points), collinearity and fixed effects conditions.

Model 1 included four level-2 explanatory variables as predictors of the random intercept (denoted w_1 - w_4). This model was crossed with three levels of collinearity, i.e., the correlation among predictors: no collinearity ($\rho = 0$); medium collinearity ($\rho = 0.5$); and high collinearity ($\rho = 0.8$). Model 1 was also crossed with three levels of the predictor fixed-effects condition: a "baseline" level where the first two predictors (w_1 and w_2) and the last two predictors (w_3 and w_4) had the same regression coefficients and was theorized to produce similar general dominance for these pairs of predictors (γ_{w1} =.3, γ_{w2} =.3, γ_{w1} =.1, γ_{w4} =.1); a "small effects" level where coefficients had an ordering from large to small but differed by small values (γ_{w1} =.5, γ_{w2} =.45, γ_{w1} =.4, γ_{w4} =.3); and a "large effects" level where coefficients had a larger magnitude and differed by larger values (γ_{w1} =.8, γ_{w2} =.6, γ_{w1} =.4, γ_{w4} =.2).

Model 2 contained 8 fixed effects: four person-level variables as predictors of the random intercept (similarly to model 1, denoted w_1 - w_4), and the same variables as predictors of the effect of time, i.e. cross-level interactions between the person-level predictors and time (denoted w_1T - w_4T). Model 2 was also crossed with three levels of predictor fixed-effects, a baseline level (denoted base-base) where each of the four main effects (w_i) and the cross-level interactions (w_iT) had coefficients as in the baseline level for model 1; a base-large effects level where the four main effects had baseline coefficients but the four cross-level interactions had coefficients as the "large

effects" level in model 1; and a large-large level where both the main effects and cross-level interactions had "large effects" coefficients.

Model 3 also contained eight fixed effects: the same four time-invariant predictors of the random intercept seen in models 1 and 2 (w_1 - w_4) and four time-varying predictors at level-1 (denoted x_1 - x_4). The levels of the predictor fixed effects condition for model 3 was similar to model 2, with base-base, base-large and large-large effects, however the first set of effects (before the hyphen) refers to the level-2 predictor coefficients and the second set of effects refers to the level-1 predictor coefficients. Models 2 and 3 were crossed with two levels of collinearity: no collinearity ($\rho = 0$) and medium collinearity ($\rho = 0.5$). All models were generated for each of the two level-1 sample size conditions, nTimePoints=4 and nTimePoints=8, and included fixed and random slopes for the effect of time on the outcome. Therefore, a total of 42 pseudo-populations were generated, 18 for model 1, 12 for model 2 and 12 for model 3.

Table 7 lists the overall maximum, minimum and average values of the measures of fit and general dominance differences across conditions based on the population data. The general dominance difference (G_{ij}) produced using the McFadden R^2 had the narrowest range among all measures, varying from a minimum of -0.11 to a maximum of 0.09, followed by the G_{ij} produced using the N&S R^2 with a range between -0.22 and 0.14. The general dominance comparisons based on R^2 Beta and R&B1 R^2 measures had similar G_{ij} ranges, with minimums around -0.30 and maximums around 0.30. The dominance measures using the R&B2 R^2 measure had the widest range, going from a minimum of -0.84 to a maximum of 0.37. Also listed in Table 7 are the maximum and minimum values that each of the R^2 measures obtained when fitting all subset models to the pseudo-population data. These results indicate that both R&B R^2 measures are not bounded between 0 and 1, which is an undesirable feature for an R^2 measure.

Table 7 Values of population parameters for the model fit (R^2) and general dominance difference measures (G_{ij}) across simulation conditions.

	Statistic	McFadden	Beta	N&S	R&B 1	R&B 2
R^2	Avg	0.09	0.34	0.31	0.23	0.36
	Min	0.00	0.00	0.02	-0.01	-2.97
	Max	0.29	0.87	0.93	0.72	0.89
G_{ij}	Avg(Abs)	0.02	0.06	0.05	0.09	0.10
	Min	-0.11	-0.31	-0.22	-0.29	-0.84
	Max	0.09	0.30	0.14	0.27	0.37

Population general dominance values for each predictor (G_i) in model 1 are displayed in Figure 1. The exact G_i values and the corresponding population rank ordering of predictors by relative importance for model 1 are listed in Table 23 and Table 25 of the Appendix, respectively. Model 1 contains only predictors of the random intercept, which vary between subjects. According to the results presented in Figure 1, the G_i values seem to reflect the rank ordering that would be expected based on the fixed effects condition (i.e., regression coefficients), although the differential additional contribution of predictors gets smaller as collinearity increases; that is, the G_i values become closer to each other. The G_i values produced with the R&B2 R^2, which is the proportional change in variance of the random intercept, have the largest magnitude of all measures for this model. The G_i values using McFadden's R^2 have the smallest magnitude among all measures of fit and do not vary much among predictors, even when fixed effects are large and collinearity is zero, indicating that this measure might not be able to detect much variability at level-2, but it still produced rankings consistent with what would be expected based on the fixed-effects and collinearity conditions. The G_i values using N&S R^2 also seem to display the low level-2 variability issue, but to a lesser extent. The number of time points did not seem to have a large influence on the pattern of dominance for this model.

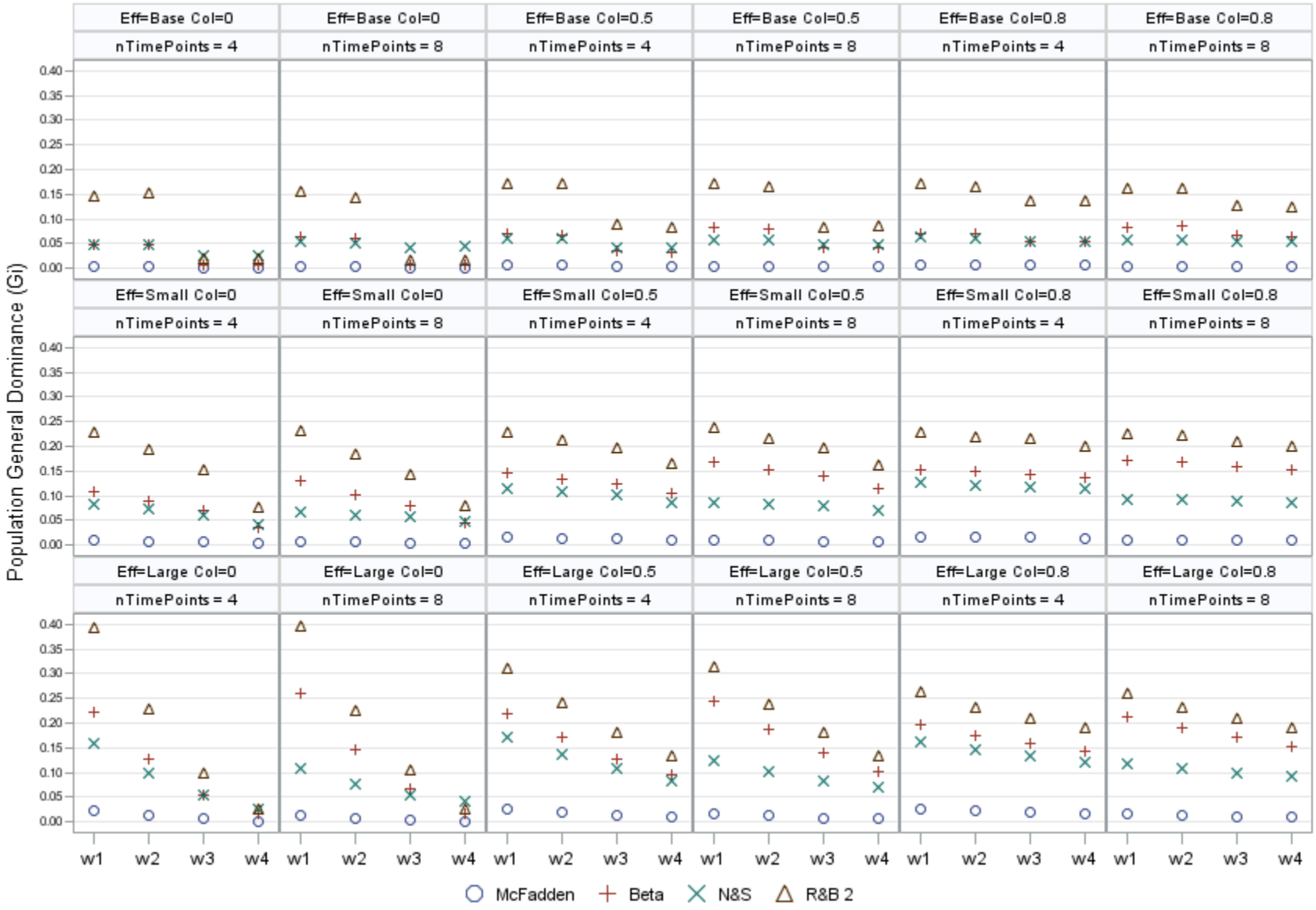

Figure 1 Model 1 population general dominance (G_i) values for all conditions.

Population general dominance values for each predictor (G_i) in model 2 are displayed in Figure 2. The exact G_i values and the corresponding population rank ordering of predictors by relative importance for model 2 are listed in Table 26 and Table 27 of the Appendix, respectively. This model, which contains interaction terms between the level-2 predictors and the time trend variable, shows a more pronounced effect of collinearity on the dominance values, especially for G_i values using the N&S and R&B2 R^2 measures. Here again the effect of collinearity was to flatten the distribution of G_i values. Additionally, in model 2 some G_i measures using the R&B2 R^2 had negative values (these negative values are outside of the chart area in Figure 2 but are listed in Table 26 of the Appendix). This is not a desirable outcome but is in line with what is known about this measure; namely, that it might decrease when level-1 predictors that explain variability at level-2 are added to the model (Snijders & Bosker, 1994). Unlike model 1, in model 2 the number of time points seemed to influence the pattern of dominance, especially for G_i values using the R&B2 and the Beta R^2. As number of time points increase, the magnitude of the additional contribution of the predictors seem to get dampened. Since this model contains interaction terms between the level-2 predictors and the time trend variable, and time trend arguably has the largest effect on the measures of fit, it is likely that as number of time points increase the total variance to be explained after accounting for the effect of time and some of the interactions is lower for conditions with more time points.

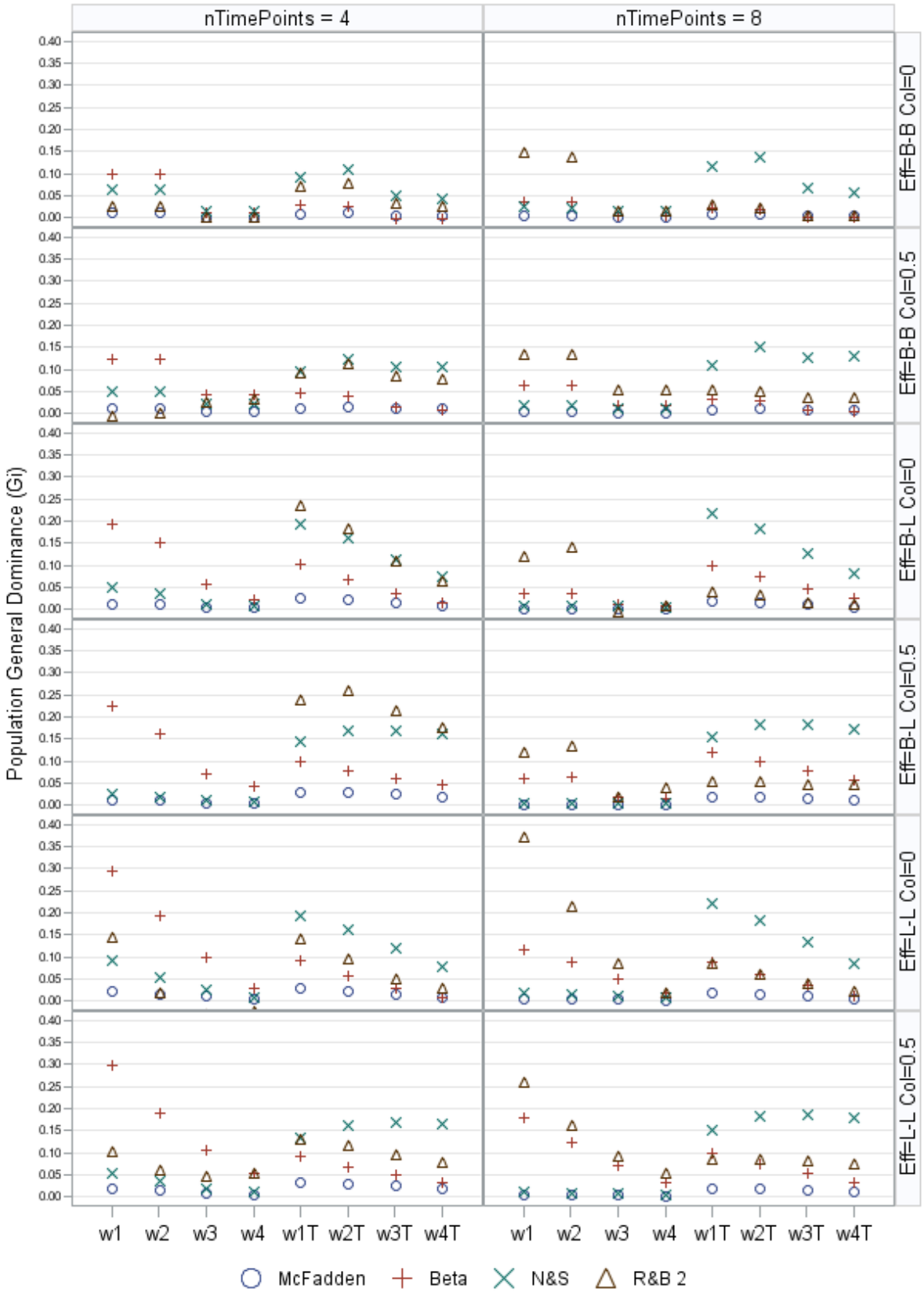

Figure 2 Model 2 population general dominance (G_i) values for all conditions.

Population general dominance values for each predictor (G_i) in model 3 are displayed in Figure 3. The exact G_i values and the corresponding population rank ordering of predictors by relative importance for model 3 are listed in Table 28 and Table 29 of the Appendix, respectively. From Figure 3, we can see that the G_i values distribution using the R&B1 R^2 is flat for all level-2 predictors in model 3 (i.e., w_1-w_4), indicating that this measure is not able to detect additional contributions of level-2 predictors. This result is expected since the R&B1 R^2 measure is the proportional change in variance at the residual level and can only detect variation at level-1. Therefore, the R&B1 R^2 should not be used to compare level-2 predictors. The G_i values using the other measures are mostly consistent with the patterns that would be expected based on fixed effect condition for this model and are not affected by collinearity. The number of time points does not seem to influence the pattern of dominance in this model.

In general, across models, the dominance values for the predictors became closer to each other as collinearity increased. Additionally, the dominance values calculated using McFadden's R^2 were much lower in magnitude than the other measures. Inspection of the complete dominance results for these models (not shown) indicate that the McFadden and N&S R^2 did not decrease as additional predictors were added to the model (i.e., these measures seem to be monotonic with model complexity based on these results). However, the R^2 Beta and the R&B R^2 did decrease with more predictors for some models, indicating that monotonicity does not hold for these measures. Specifically, these measures decreased in model 2 when cross-level interaction terms were added to models with main effects (level-2 predictors). The complete dominance results from model 3 also showed that R^2 Beta decreased when time varying predictors (level-1) were added to subset models with level-2 predictors. Therefore, R^2 Beta is only adequate for comparing predictors within the same level of analysis.

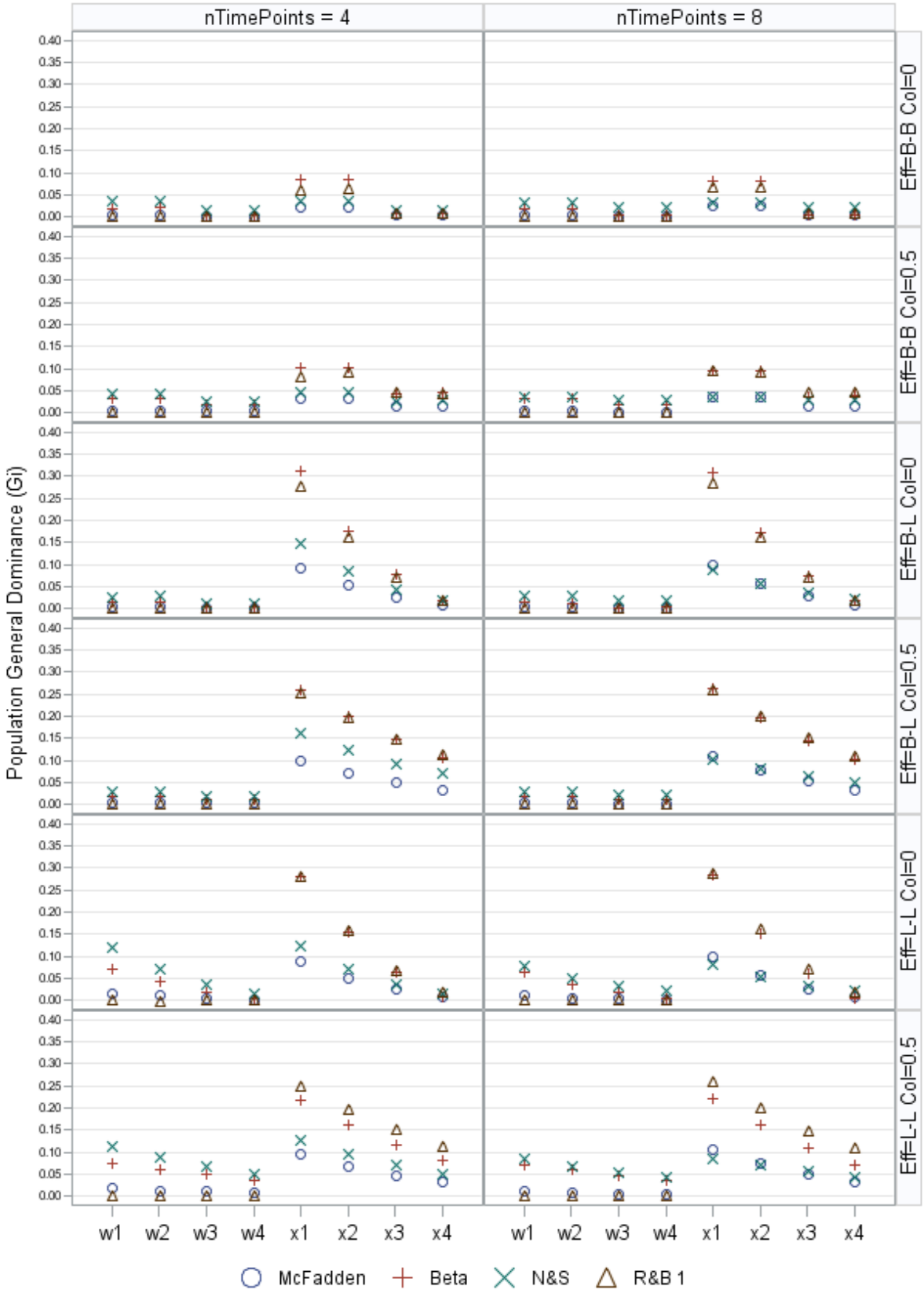

Figure 3 Model 3 population general dominance (*Gi*) values for all conditions.

The population general dominance difference measures (G_{ij}) can be derived from the G_i values by subtracting the additional contribution of one predictor from that of another, i.e., $G_{ij} = G_i - G_j$. Pairs of predictors in conditions and measures where the G_i values are the same, and thus $G_{ij} = 0$, were used for calculating type I error rates. All non-zero G_{ij} measures were used for calculating power. Exact values of population G_{ij} for model 1 are listed in Table 24 of the Appendix. For models 2 and 3, the number of G_{ij} measures in each pseudo-population is large and therefore are not shown, but can be derived from the G_i values listed in Table 26 and Table 28 of the Appendix respectively. Values of $G_{ij} = 0$ (i.e., equally important predictors) in the population occurred for the most part in conditions where predictors had the same regression coefficients. Specifically, this was the case for G_{12} and G_{34} in the baseline effect condition for model 1 as well as the base-base and base-large effects for models 2 and 3. However, not all predictors with the same fixed effects had zero population dominance difference values and not all predictor comparisons with zero-valued population dominance measures had the same fixed effects, which goes to show that predictor coefficients should not be confused with dominance effects, particularly when predictors are correlated. Table 8 shows the number of dominance measures considered for the calculation of type I error and power across conditions.

Table 8 Number of population general dominance difference measures (G_{ij}) used for Type I (T1) and Power (P) rates evaluation across conditions summed over level-1 sample size.

	Effect	**Baseline/ Base-Base**						**Small/ Base-Large**						**Large/ Large-Large**					
	Collinearity	0.0		0.5		0.8		0.0		0.5		0.8		0.0		0.5		0.8	
Model	**R^2**	T1	P	T1	P	T1	P	T1	P	T1	P	T1	P	T1	P	T1	P	P	
1	McFadden	3	9	4	8	4	8		12		12	1	11		12		12	12	
	Beta	1	11	1	11		12		12		12		12		12		12	12	
	N&S	2	10	3	9	2	10		12		12		12		12		12	12	
	R&B 2		12		12	1	11		12		12		12		12		12	12	
2	McFadden	7	41	12	36			2	46	2	46			1	47	1	47		
	Beta	2	46	3	45				48		48			1	47		48		
	N&S	3	45	4	44			1	47	3	45				48		48		
	R&B 2	3	45	2	46				48	2	46				48	1	47		
3	McFadden	7	49	7	49			4	52	3	53				56		56		
	Beta	6	50	3	53			2	54	4	52				56		56		
	N&S	15	41	11	45			2	54	3	53			3	53	1	55		
	R&B 1	14	42	8	48			12	44	9	47			7	49	8	48		

DA Example

To illustrate the application of dominance analysis with longitudinal multilevel models, an example is presented based on one simulation parent sample. This parent sample is drawn from a population created with model 1; therefore, there are four predictors of the random intercept at level-2. The sample has 200 subjects with 4 time points and a medium level of collinearity among predictors ($\rho = 0.5$). For this specific example, the sample was estimated with the SGR covariance structure.

Table 9 shows the dominance analysis results using both the McFadden's (the shorthand MF is used here) and N&S R^2 measures. The columns labeled with the R^2 measure show the values

of these measures associated with each subset model, and the remaining columns show the additional contribution of each predictor to each subset model. For example, the model containing w_1 produces MF R^2 = .050 and N&S R^2 = .359 (see entry in the w_1 row under the McFadden and N&S columns). If w_3 is added to this model, the MF R^2 increases by .010 and the N&S R^2 increases by .044, and this increase is the entry in the w_1 row and w_3 column in the table for each measure. This means that the model $w_1\, w_3$ will result in a MF R^2 = .050 + .010 = .060 and in a N&S R^2 = .359 + .044 = .403, which are the values under each R^2 measures in the w_1, w_3 row. If w_4 is added to the w_1 model, the MF R^2 increases by .006 and the N&S R^2 increases by .030, as shown in the w_1 row and w_4 column of the table. Therefore, the model $w_1\, w_4$ will result in a MF R^2 = .050 + .006 = .056 and in a N&S R^2 = .359 + .030 = .389, which are the values in the w_1, w_4 row under the respective measure. According to these results, w_3 dominates w_4 when added to the w_1 subset model because the additional contribution of w_3 to this model (MF=.010, N&S=.044) is larger than the additional contribution of w_4 to the same model (MF=.006, N&S=.030). If the additional contribution of w_3 was larger than the additional contributions of w_4 to all subset models (i.e., in every row in this table), we would say that w_3 completely dominates w_4. However, we can see that for the subset model containing only w_2, the additional contribution of w_3 (MF=.003, N&S=.015) is actually smaller than that of w_4 (MF=.007, N&S=.034), thus we say that complete dominance cannot be established between w_3 and w_4. When this occurs, the "average" rows of the table can be used to determine conditional dominance. For instance, the average additional contribution of w_3 to models of size 1 is computed as (.010 + .003 + .009) / 3 =.007 for MF R^2 and as (.044 + .015 + .053) / 3 = .037 for N&S R^2. + .003)/3 = .004. Conditional dominance is established by comparing the average additional contribution across all model sizes. In the case of w_3 and w_4, we can see that for models of size 3 (k = 3), the average additional contribution of w_3 (MF=.001,

N&S=.002) is the same as that of w_4; therefore, conditional dominance also cannot be established between these two predictors. Hence, the last and least restrictive form of dominance is investigated, namely general dominance. General dominance corresponds to the "Overall Average" row in Table 9 and is computed by averaging all the conditional dominance measures, or all the "k = n Average" rows in the table. For example, in Table 9 the general dominance measure for w_4 is computed as G_4 = (.028 + .008 + .003 + .001) / 4 = .010 for MF and G_4 = (.249+ .046+ .012+ .002) / 4 = .077 for N&S. In this example, w_4 is said to generally dominance w_3 because the overall average for w_4 (MF G_4 =.010, N&S G_4 =.077) is greater than the overall average for w_3 (MF G_3 =.009, N&S G_3 =.069).

The example presented in Table 9 also demonstrates a desirable characteristic of the general dominance measures, namely, that they add up to the given measure of model fit of the full model (which contains all predictors, w_1 w_2 w_3 w_4 in this example), allowing for a direct decomposition of the full model's measure of fit across all predictors. In this example, we can see that $\sum_{i=1}^{p} G_i = .031 + .027 + .009 + .010 = .077$ for MF R^2, and $\sum_{i=1}^{p} G_i = .165 + .161 + .069 + .077 = .472$ for N&S R^2.

The general dominance (G_i) results shown in Table 9 are used to calculate the general dominance difference values between pairs of predictors: $G_{ij} = G_i - G_j$. Here, for w_3 and w_4, the measure would be $G_{34} = G_3 - G_4 = .009 - .010 = -.001$ for MF R^2 and $G_{34} = G_3 - G_4 = .069 - .077 = -.008$ for N&S R^2. The negative values indicate that the dominance relationship is in the opposite direction of how the predictors are listed in the measure, so in this example a negative G_{34} indicates that w_4 dominates w_3 at the general level.

Table 9 Dominance Analysis example for a parent sample from condition: nSubjects=200, nTimePoints=4, Fixed Effects=Large, Collinearity=0.5, Covariance Structure=SGR.

		Additional Contribution of					Additional Contribution of			
Subset model	McFadden R^2	w_1	w_2	w_3	w_4	N&S R^2	w_1	w_2	w_3	w_4
k = 0 Average	0	.050	.047	.025	.028	0	.359	.365	.228	.249
w_1	.050		.025	.010	.006	.359		.107	.044	.030
w_2	.047	.028		.003	.007	.365	.101		.015	.034
w_3	.025	.035	.026		.012	.228	.176	.152		.074
w_4	.028	.028	.027	.009		.249	.140	.150	.053	
k = 1 Average		.030	.026	.007	.008		.139	.137	.037	.046
w_1, w_2	.075			.001	.001	.467			.004	.004
w_1, w_3	.060		.016		.002	.403		.067		.010
w_1, w_4	.056		.020	.006		.389		.082	.024	
w_2, w_3	.051	.025			.005	.380	.090			.023
w_2, w_4	.055	.021		.001		.399	.072		.005	
w_3, w_4	.037	.025	.019			.302	.112	.102		
k = 2 Average		.024	.018	.003	.003		.091	.083	.011	.012
w_1, w_2, w_3	.076				.001	.470				.002
w_1, w_2, w_4	.076			.001		.470			.002	
w_1, w_3, w_4	.062		.015			.413		.059		
w_2, w_3, w_4	.056	.021				.403	.069			
k = 3 Average		.021	.015	.001	.001		.069	.059	.002	.002
w_1, w_2, w_3, w_4	.077					.472				
Overall Average (G_i^S)		.031	.027	.009	.010		.165	.161	.069	.077
Population Parameter (G_i)		.026	.019	.013	.009		.173	.137	.107	.082

Table 10 shows all the G_{ij} values for this specific parent sample, the corresponding population parameters and their bootstrap estimates, and the percentile and asymptotic confidence intervals obtained using the bootstrap samples. For example, for the MF R^2, G_{14} = .017 in the population. In the parent sample, and, in this case, when estimated as an average across bootstrap samples, G_{14} = .021. The positive values indicate that w_1 dominates w_4 in both the population and

parent sample. On the other hand, G_{34} = .004 in the population but G_{34} = -.001 in the parent sample, indicating that while in the population w_4 dominates w_3, in the parent sample the opposite is true. In general, the bootstrap G_{ij} estimates will be closer in value to the parent sample than to the population, and since there is a mismatch between the parent sample and the population, the bootstrap estimates will likely also display a similar discrepancy. In this example in particular, the bootstrap estimates with the MF measure are exact estimates of the parent sample, so they will reflect the same bias in relation to the population that was present in their corresponding parent sample.

The last part of Table 10 shows the qualitative dominance measures, D_{ij}, obtained from the population and the parent sample, as well as the reproducibility of these values over the B=300 bootstrap samples. For example, in the parent sample D_{14} = 1 since $G_1 > G_4$ and therefore G_{14} is positive. Conversely, D_{34} = -1 since $G_3 < G_4$ and therefore G_{34} is negative. If there were any measures for which $G_i = G_j$, D_{ij} would be 0. The last two rows of Table 10 show the reproducibility of the population and of the parent sample D_{ij} values over the bootstrap samples. The second to last row shows the reproducibility values corresponding to the proportion of bootstrap samples that replicated the D_{ij} values found in the population, and the last row indicates the proportion of bootstrap samples that replicated the D_{ij} values found in the parent sample. These values will match if the D_{ij} values are the same between the population and parent sample, but they will be different if there is a mismatch, as is the case with D_{34}. In the case of D_{34}, 55.3% of the bootstrap samples agreed with the parent sample dominance pattern (D_{34} = -1), 32.3% of the bootstrap samples agreed with the population dominance pattern (D_{34} = 1), and the remaining 12.4% of bootstrap samples found an indeterminate dominance pattern (D_{34} = 0). In general, the magnitude of the reproducibility will be proportional to the magnitude of the G_{ij} values. In this example we can see

this is the case with D_{12} and D_{34} having much lower reproducibility than the other pairs of predictors. In particular, when comparing the measures of fit, we can see that D_{34} had a mismatch between population and parent sample for both measures, but for the N&S R^2 the absolute magnitude of the G_{34} measure was higher than that of G_{12}, so the D_{34} reproducibility was higher than that of D_{12}, while for the MF R^2 the G_{12} value was larger in absolute terms than the G_{34} value, and thus D_{34} showed a smaller reproducibility value than D_{12}. It is worth noting that reproducibility values for the dominance relationships other than D_{12} and D_{34} were either 1 or very close to 1, indicating that all bootstrap samples replicated the dominance patterns found in the population and parent sample for these dominance relationships.

Table 10 also presents the percentile and asymptotic normal confidence intervals for the parent sample G_{ij} measures. We can see that the upper and lower bounds are very similar between the two CI types. Additionally, the CIs for G_{13}, G_{14}, G_{23} and G_{24} do not contain 0, indicating that general dominance is well established (and null hypothesis of no dominance is rejected) between the predictors in these pairs. The inferential results also mirrored the reproducibility results, where a larger magnitude of the dominance general difference was linked to a higher degree of confidence that the sample dominance relationship was actually present in the population. In this example both G_{12} and G_{34} are very small in the population and in the parent samples, and both the reproducibility and CI results indicate that we should not place much certainty in the fact that w_1 is more (or less) important than w_2 and that w_3 is more (or less) important than w_4 in explaining the outcome. On the other hand, based on these results we can be fairly certain that w_1 and w_2 are relatively more important than both w_3 and w_4.

Table 10 Population general dominance measures, estimates and CIs (using McFadden R^2) for condition: nSubjects=200, nTimePoints=4, Fixed Effects=Large, Collinearity=0.5, Covariance Structure=SGR.

	McFadden R^2						N&S R^2					
Result	G_{12}	G_{13}	G_{14}	G_{23}	G_{24}	G_{34}	G_{12}	G_{13}	G_{14}	G_{23}	G_{24}	G_{34}
Population parameter	.007	.013	.017	.006	.010	.004	.036	.066	.091	.030	.055	.025
Parent sample estimate	.004	.022	.021	.018	.017	-.001	.004	.096	.088	.092	.084	-.008
Bias (parameter-parent sample estimate)	.003	-.009	-.004	-.012	-.007	.005	.032	-.030	.003	-.062	-.029	.033
Bootstrap samples average estimate	.004	.022	.021	.018	.017	-.001	.004	.093	.085	.090	.082	-.008
Bias (parameter-bootstrap estimate)	.003	-.009	-.004	-.012	-.007	.005	.032	-.027	.006	-.060	-.027	.033
Bias (parent-bootstrap estimate)	.000	.000	.000	.000	.000	.000	.000	.003	.003	.002	.002	.000
Percentile CI	-.009, .017	.010, .032	.011, .032	.006, .030	.006, .027	-.008, .008	-.054, .062	.042, .144	.039, .133	.037, .144	.027, .131	-.045, .039
Asymptotic Normal CI	-.009, .017	.011, .033	.011, .031	.006, .030	.007, .027	-.009, .007	-.054, .062	.045, .147	.039, .137	.040, .144	.034, .134	-.050, .034
Result	D_{12}	D_{13}	D_{14}	D_{23}	D_{24}	D_{34}	D_{12}	D_{13}	D_{14}	D_{23}	D_{24}	D_{34}
Population D_{ij}	1	1	1	1	1	1	1	1	1	1	1	1
Parent sample D_{ij}	1	1	1	1	1	-1	1	1	1	1	1	-1
Reproducibility of population D_{ij}	.737	.997	1.0	.997	1.0	.323	.557	.997	1.0	1.0	1.0	.360
Reproducibility of parent sample D_{ij}	.737	.997	1.0	.997	1.0	.553	.557	.997	1.0	1.0	1.0	.630

Simulation Results

Rate of non-positive definite (npd) random components covariance matrices

Due to the complexity of the models estimated here, the proportion of non-positive definite G-matrices was very high for some combinations of the design factors (i.e., simulation conditions). In instances when non-positive definite G-matrices are present, the resulting variance component estimates are questionable and may not be used (Schoeneberger, 2016). These estimation issues impacted the choice of measures of model fit that were used for further analyses. Of the measures of model fit considered in this study - Nakagawa and Schielzeth's marginal R^2 (N&S), Edwards et al.'s R^2_β (Beta), McFadden's R^2 (McFadden), and Raudenbush and Bryk's pseudo-R^2 measures representing the proportional change in variance of the residual (R&B 1) and of the random intercept (R&B 2) - the first (N&S) and the last (R&B 2) measures rely on level-2 random effects estimates and should not be used for conditions where the covariance matrices were found to be non-positive definite.

For models estimated using the GAR covariance structure, an average of 38% of the bootstrap samples per replication across simulation conditions had at least one subset model estimate that resulted in a npd covariance matrix. This rate decreases to just under 7% for models estimated using the SGR covariance structure. At the model level, model 1 had the lowest rate of npd matrices, with an average of just under 14% of bootstrap sample estimates having a npd G-matrix. This rate went up to 21% for model 2 and 43% for model 3.

In order to investigate what combination of factors were most strongly related to non-positive definite random effect covariance matrices, a factorial ANOVA was conducted. As expected, the covariance matrix specification was the factor with the largest effect size ($\eta^2 = .30$),

followed by model complexity ($\eta^2 = .10$), sample size at level-2 ($\eta^2 = .09$), and sample size at level-1 ($\eta^2 = .07$). None of the other studied factors or their interactions were found to have a relevant effect on the rate of non-positive definite matrices. Figure 4 shows how these factors impact the rate of npd G-matrices in bootstrap samples. From this figure we see that the GAR covariance structure produced npd G-matrices for every condition, even at the highest sample sizes. The problem is considerably less severe for the SGR covariance structure, since it is a simpler structure that estimates a smaller number of parameters. However, the simulation data were generated using the GAR structure. Figure 5 shows the rate of non-positive definite G-matrices in the simple random (parent) samples. The problem is less severe in general for SRS, implying that the bootstrapping procedure aggravated this issue.

Due to the high rate of non-positive definite covariance matrices found when estimating models with the GAR covariance structure, the remaining analyses presented here will focus on results estimated with the SGR structure. For the conditions where this problem was not as severe (i.e., Model=1, nSubjects=1000, nTimePoints=8), a comparison of the results between these two covariance structures was performed to determine whether this factor was indeed influential.

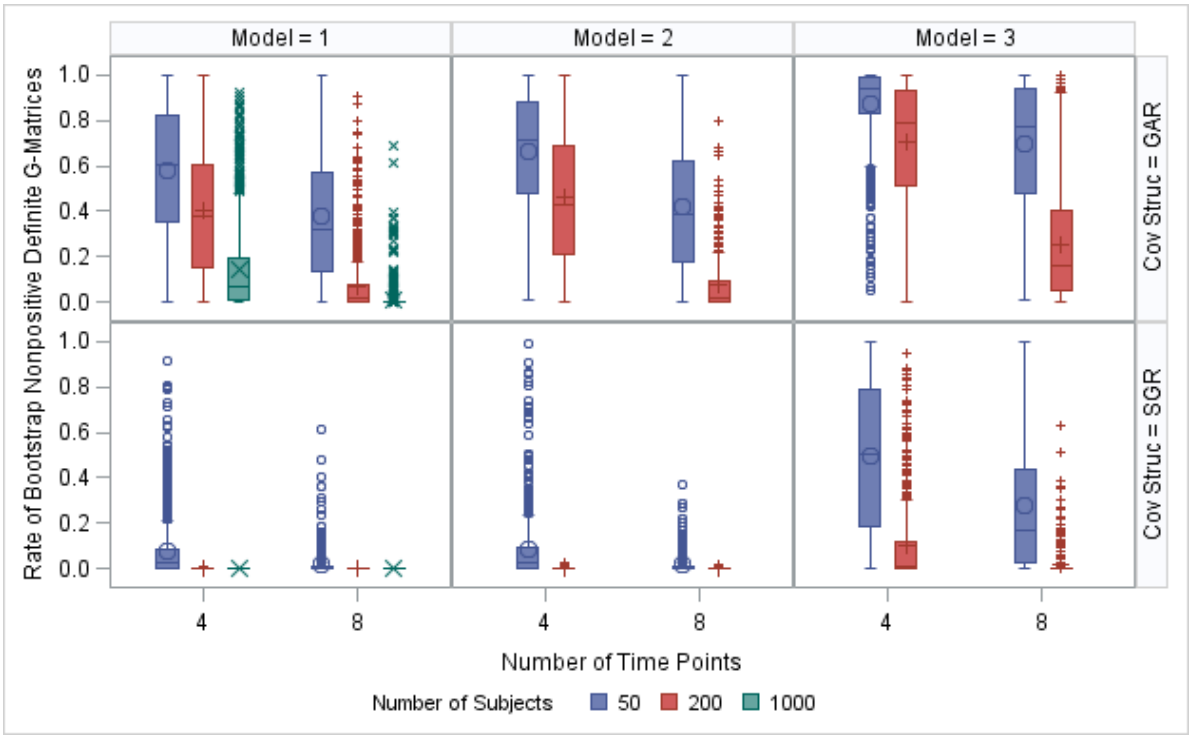

Figure 4 Distribution of rate of non-positive definite G-matrices for bootstrap samples across simulation conditions and replications.

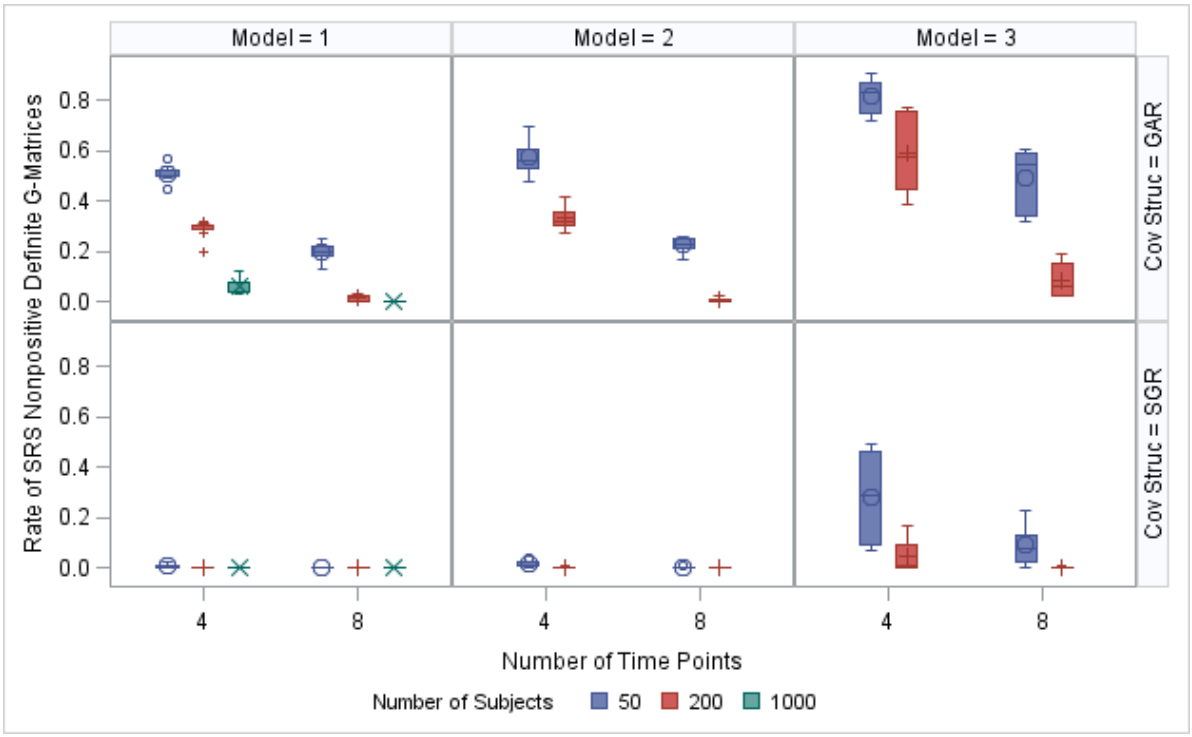

Figure 5 Distribution of rate of non-positive definite G-matrices for simple random (parent) samples.

Ranking accuracy

To evaluate the factors associated with an accurate rank-ordering of predictors (by relative importance) using dominance analysis, ANOVAs were conducted using the simulation conditions as design factors. Results from the overall factorial ANOVA using factors that were fully crossed will be presented first, followed by results of the ANOVA conducted for model 1 only, which included additional levels of sample size and collinearity factors. The results for any combination of factors found to have a moderate effect ($\eta^2 \geq .05$) on the outcome will be emphasized.

Predictor identified by DA as most important

The proportion of bootstrap samples that agreed with the population on the predictor identified as most important by DA was computed for each condition. An examination of the factors related to this outcome indicated that the most influential factors were sample size at level-2 ($\eta^2 = .05$), the interaction between model complexity and measure of fit ($\eta^2 = .06$), and the interaction between model complexity and predictor fixed effects condition ($\eta^2 = .05$).

Results from the ANOVA of model 1 results indicated that, within this model, the most influential factors associated with top predictor agreement were the predictor fixed effects ($\eta^2 = .13$) and the sample size at level-2 ($\eta^2 = .13$). Based on the descriptive data, provided in Table 11, we see that rates of agreement increased with an increase in level-2 sample size. Additionally, for model 1, when data were simulated using small effect sizes the rates of agreements were lower than with the baseline effects and the large effects. This is expected given the population values presented in Figure 1. The general dominance values were very close to each other for this condition which might result in unstable rank orderings due to sampling error. On average, the McFadden R^2 measure seemed to produce the highest rates of agreement, followed by the R^2 Beta and the N&S R^2 respectively. The N&S R^2 seems to have performed especially poorly in model 3

under the baseline and large effect conditions even though it performed comparably with other measures in models 1 and 2. This measure is constructed based on variance components estimates and was most severely affected by rates of non-positive definite matrices, which may help explain its poor performance in model 3 (the model with the highest rates of non-positive definite matrices).

Table 11 Percentage of bootstrap samples that agree with the population on the predictor ranked most important by DA.

% Agree	Top Predictor	Baseline/ Base-Base			Small/ Base-Large			Large/ Large-Large			All
Model	R^2 / nSubjects	50	200	1000	50	200	1000	50	200	1000	
1	McFadden	64	80	94	36	47	63	56	78	96	68
	Beta	35	41	50	34	43	57	55	75	95	54
	N&S	56	71	83	34	43	58	55	75	94	63
	R&B 2	39	46	58	33	42	55	53	73	94	55
2	McFadden	44	70		57	78		55	75		63
	Beta	67	81		46	73		57	78		67
	N&S	46	67		56	71		47	64		58
	R&B 2	31	48		33	54		38	57		43
3	McFadden	92	100		88	99		88	99		94
	Beta	57	75		88	99		86	99		84
	N&S	35	43		65	88		34	44		51
	R&B 1	59	64		86	98		85	98		82

Figure 6 compares the rates of agreement in terms of the predictor ranked as most important between (1) the bootstrap samples and population or (2) the bootstrap samples with their parent sample. The agreement between SRS and population is also shown and used as a check on the sampling procedure. We can see from this figure that the pattern of results among comparison types is similar, indicating that results obtained with the bootstrap samples accurately reflect the

patterns in the population. In general, the agreement rates between bootstrap and population were of an order of magnitude smaller than the agreement between the bootstrap and their parent sample rankings. On average, the true rate of agreement between bootstrap and population was about 20 percentage points lower than the observed agreement between bootstrap and parent sample.

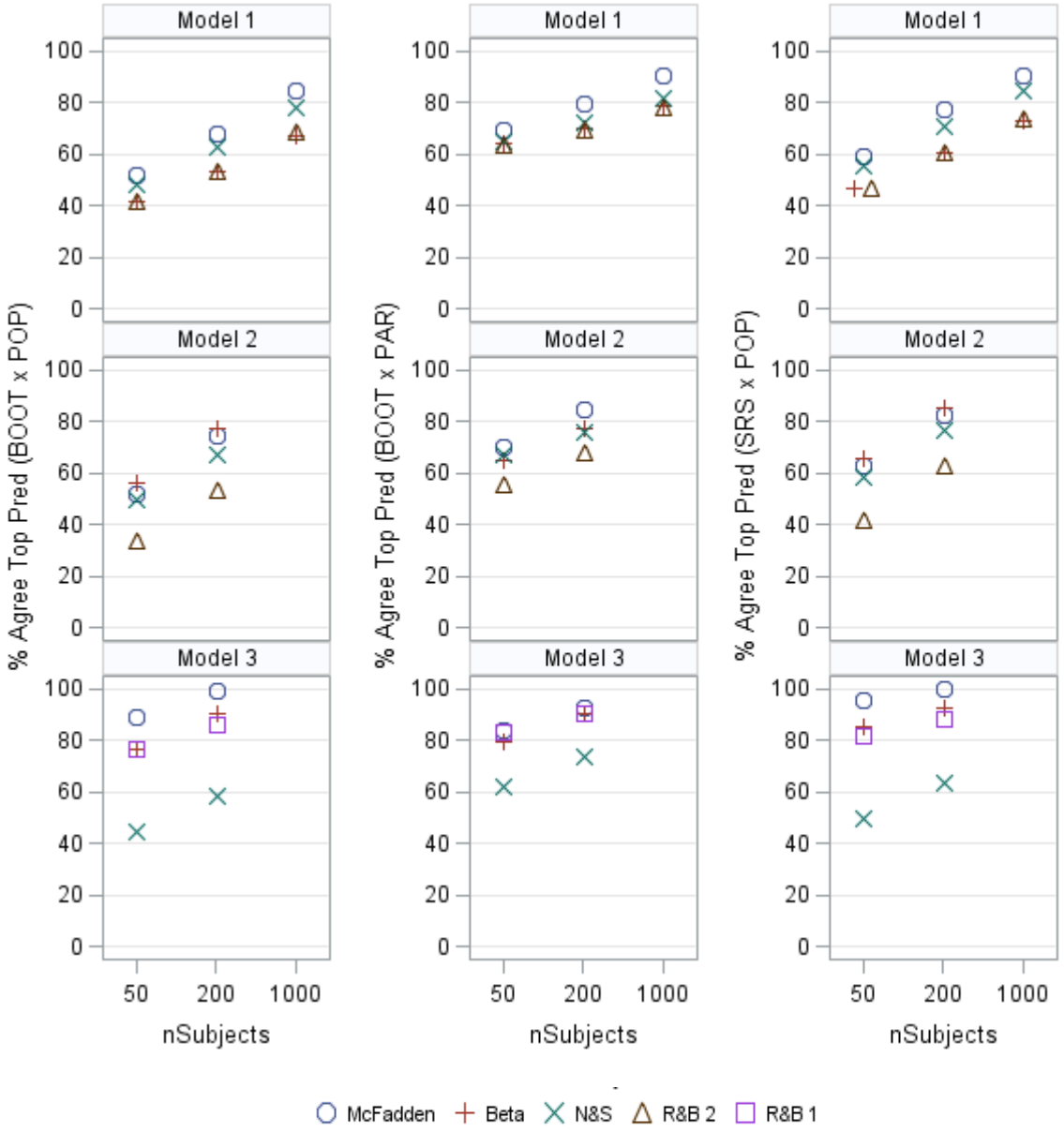

Figure 6 Average agreement rates in terms of the predictor ranked most important by DA when compared between bootstrap sample and population (left), bootstrap and parent sample (middle), and SRS and population (right).

Predictor identified by DA as the least important

The proportion of bootstrap samples that agreed with the population on the predictor identified as least important by DA was analyzed by simulation condition. An examination of the factors related to this outcome indicated that the most influential factors were measure of fit (η^2 = .16) and the sample size at level-2 (η^2 = .09). Results from the ANOVA of model 1 results indicated that, within this model, the most influential factors associated with least important predictor agreement were sample size at level-2 (η^2 = .19), measure of fit (η^2 = .07), and the predictor fixed effects condition (η^2 = .06).

Based on the descriptive data presented in Table 12, we see that rates of agreement increased with an increase in level-2 sample size across all models and measures of fit. Additionally, for model 1, as was the case for the agreement of the top predictor, when data were simulated using small effect sizes the rates of agreements were lower than with the baseline effects and the large effects. The same explanation applies here, where population dominance measures that are too close (but not zero) might create instability in the samples and a decreased likelihood of good agreement with the population. On average, the McFadden R^2 measure seemed to produce the highest rates of agreement, followed by the N&S R^2 and the R^2 Beta respectively, except for model 3 where the R&B1 R^2 (proportion change in variance of the level-1 residual) performed better than all the other measures.

Figure 7 compares the rates of agreement in terms of the predictor ranked as least important when the comparison is done between the bootstrap samples and population, as well as the rates produced when comparing the bootstrap with its parent sample. The comparison between SRS and population is again used as a check on the sampling procedure. We can see from this figure that the pattern of results is very similar across the comparison types, indicating that results obtained

with the bootstrap samples accurately reflect the patterns in the data. On average, as expected, the agreement rates between bootstrap and population were of an order of magnitude smaller than the agreement between the bootstrap and their parent sample rankings, although not as far apart as was the case for the most important predictor. On average, the agreement proportion for the least important predictor was about 5-15% points lower when comparing the bootstrap with the population than when comparing the bootstrap with their parent sample.

Table 12 Percentage of bootstrap samples that agree with the population on the predictor ranked last by DA.

% Agree	Last Predictor	Baseline/ Base-Base			Small/ Base-Large			Large/ Large-Large			All
Model	R^2 / nSubjects	50	200	1000	50	200	1000	50	200	1000	
1	McFadden	74	90	99	44	60	82	55	79	96	75
	Beta	45	57	69	36	49	71	51	74	94	61
	N&S	52	65	77	36	50	72	51	74	93	63
	R&B 2	32	41	51	35	49	70	49	72	92	55
2	McFadden	78	96		68	85		65	85		79
	Beta	54	69		40	58		70	84		62
	N&S	77	96		63	80		70	89		79
	R&B 2	24	36		42	63		29	45		40
3	McFadden	68	90		76	93		58	85		78
	Beta	26	47		50	72		43	63		50
	N&S	54	77		51	81		52	76		65
	R&B 1	79	95		87	97		77	96		89

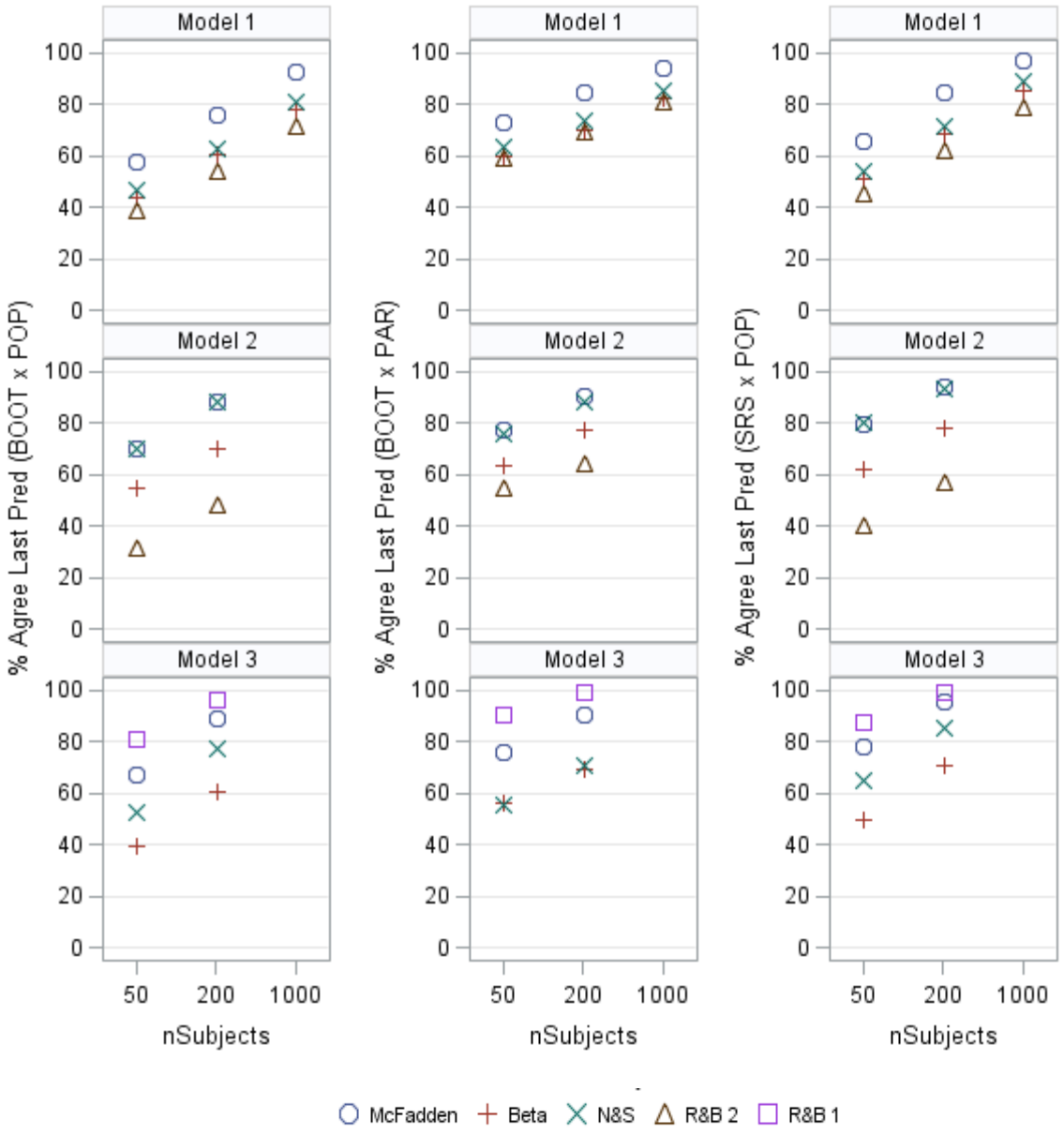

Figure 7 Average agreement rates in terms of the predictor ranked least important by DA when compared between bootstrap sample and population (left), bootstrap and parent sample (middle), and SRS and population (right).

Kendall's tau-b

The Kendall rank correlation coefficient (Kendall tau) for the correlation between the population and bootstrap DA predictor rankings was computed for each condition. Examination of the factors related to this outcome indicated that the most influential factors were model complexity (η^2 = .13), sample size at level-2 (η^2 = .11), measure of fit (η^2 = .08), the predictor fixed effects condition (η^2 = .07), the interaction between model complexity and predictor fixed effects condition (η^2 = .06), and the interaction between model complexity and measure of fit (η^2 = .06). The ANOVA results for model 1 indicate that, within this model, the most influential factors associated with predictor ranking agreement were the sample size at level-2 (η^2 = .28), the predictor fixed effects (η^2 = .18), and the level of collinearity (η^2 = .07).

Descriptive results from the factors that seemed to more strongly impact the rank correlations are listed in Table 13. Correlations below .3, considered a low correlation value, were highlighted in red; correlations above .5, indicating moderate to strong correlations, were highlighted in green. Based on the descriptive data presented in Table 13, we see that Kendall tau correlations increased with an increase in level-2 sample size. The lowest correlations occurred when sample size at level-2 was 50. For model 1, correlations above .5 were only reached when sample size at level-2 was 1000 for the baseline and small effects conditions. Additionally, for model 1, when data were simulated using small effect sizes, the Kendall tau correlations were lower than with the baseline effects and the large effects. For models 2 and 3, the large fixed effect condition was not consistently associated with higher Kendall tau correlations. The instability of the rankings for these conditions might be related to population general dominance values that are too close for some measures (see Figure 2 and Figure 3). For model 1, all measures of model fit performed similarly well in terms of Kendall tau correlation. For model 2, the N&S R^2 measure

was the best performing followed by McFadden's R^2. For model 3, the R&B1 R^2 (proportion change in variance of the level-1 residual) performed better than all the other measures, with McFadden's R^2 performing second best. Table 14 contains the model 1 Kendall tau averages by factors in this model. We can see that there is a noticeable increase in average correlations with an increase in sample size. Conversely, the increase in collinearity negatively impacts the Kendall tau correlations. At the highest collinearity level (0.8), the average tau correlation is above .5 only for the highest sample size and fixed effects conditions.

Table 13 Kendall's tau rank correlation between population and bootstrap DA predictor rankings.

	Effect	**Baseline/ Base-Base**			**Small/ Base-Large**			**Large/ Large-Large**			**All**
Model	**R^2 / nSubjects**	**50**	**200**	**1000**	**50**	**200**	**1000**	**50**	**200**	**1000**	
1	McFadden	.21	.47	.75	.16	.34	.57	.45	.73	.94	.51
	Beta	.19	.39	.63	.16	.33	.55	.44	.72	.94	.48
	N&S	.20	.43	.69	.16	.33	.56	.44	.72	.94	.49
	R&B 2	.17	.36	.60	.15	.31	.54	.41	.69	.92	.46
2	McFadden	.51	.74		.71	.87		.68	.86		.73
	Beta	.50	.72		.51	.75		.66	.83		.66
	N&S	.66	.83		.77	.88		.79	.90		.80
	R&B 2	.25	.49		.41	.66		.36	.61		.46
3	McFadden	.61	.78		.80	.91		.80	.93		.80
	Beta	.40	.65		.72	.85		.68	.85		.69
	N&S	.24	.49		.57	.79		.44	.68		.54
	R&B 1	.77	.88		.89	.97		.86	.94		.89

Table 14 Kendall's tau rank correlation between population and bootstrap DA predictor rankings for model 1 by collinearity, level-2 sample size and predictor fixed effect conditions.

Effect	**Baseline**			**Small Effect**			**Large Effect**			**All**
Collinearity	50	200	1000	50	200	1000	50	200	1000	
0	.22	.54	.75	.23	.42	.76	.56	.84	.99	.59
0.5	.19	.44	.70	.16	.36	.60	.43	.73	.95	.51
0.8	.16	.26	.55	.09	.20	.30	.31	.58	.87	.37

The Kendall tau correlations between bootstrap and population rankings are compared to the correlations between the bootstrap and parent sample rankings in Figure 8. The agreement levels between SRS and population is again used as a check on the sampling procedure. As was the case with the previous ranking outcome measures, the pattern of results is effectively the same across comparison types, indicating that results obtained with the bootstrap samples accurately reflect the patterns in the data. On average, Kendall tau correlations for the agreement between bootstrap and the (true) population predictor rankings were about .20 points (20 percentage points) lower than between bootstrap and parent sample. Model complexity seemed to have a positive effect on Kendall tau; that is, holding sample size constant, the more complex the model (larger number of predictors) the larger the Kendall tau correlation values.

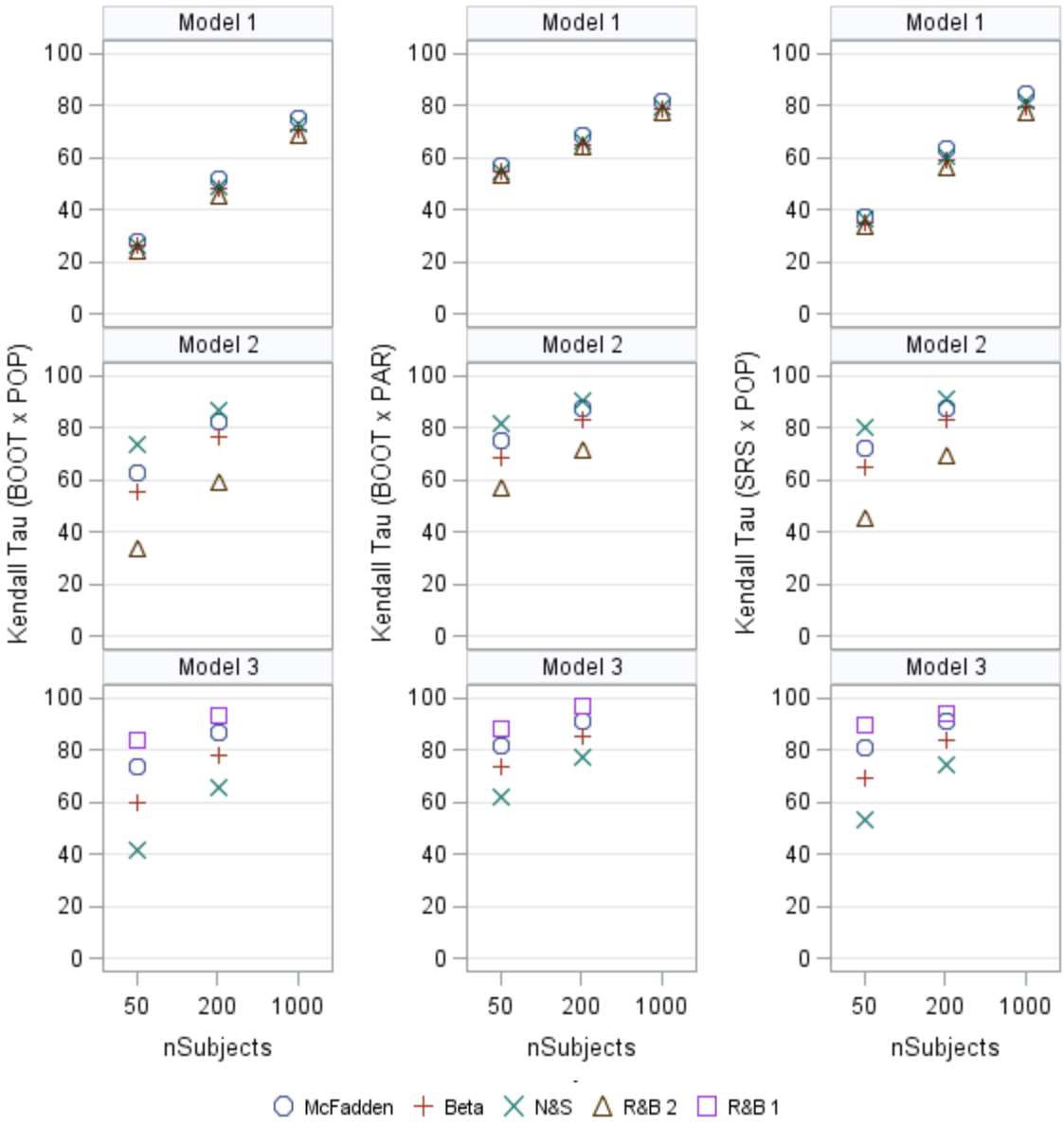

Figure 8 Average agreement rates in terms of the Kendall tau rank order correlation between bootstrap sample and population (left), bootstrap and parent sample (middle), and SRS and population (right).

Bias

Standardized bias was computed to compare the bootstrap DA estimates to their corresponding population parameters by condition. Examination of the factors related to bias indicated that the most influential factors were measure of fit (η^2 = .08) and model complexity (η^2 = .05). Results from the ANOVA of model 1 results did not find any factors with a meaningful effect on standardized bias; any individual factor or interaction explained no more than 3% of the variance in standardized bias. Of special interest, the covariance structure factor had only a small effect on standardized bias.

Average bootstrap standardized bias values are shown in Table 15 and Figure 9 for the factors deemed influential by the ANOVA. On average, standardized bias values were low for all but the R&B measures, indicating that bootstrap DA measures did not deviate much from their corresponding population values.

Table 15 Average standardized bias between bootstrap and population DA measures.

	Model 1			**Model 2**		**Model 3**		
R^2 / nSubjects	**50**	**200**	**1000**	**50**	**200**	**50**	**200**	**All**
McFadden	-0.09	-0.10	-0.13	-0.05	-0.03	0.06	0.08	-0.05
Beta	-0.09	-0.06	-0.05	0.18	0.08	0.21	0.12	0.03
N&S	-0.20	-0.21	-0.36	0.01	0.12	0.08	0.03	-0.11
R&B 1						-0.31	-0.53	-0.42
R&B 2	-0.24	-0.43	-0.87	-0.10	-0.05			-0.38

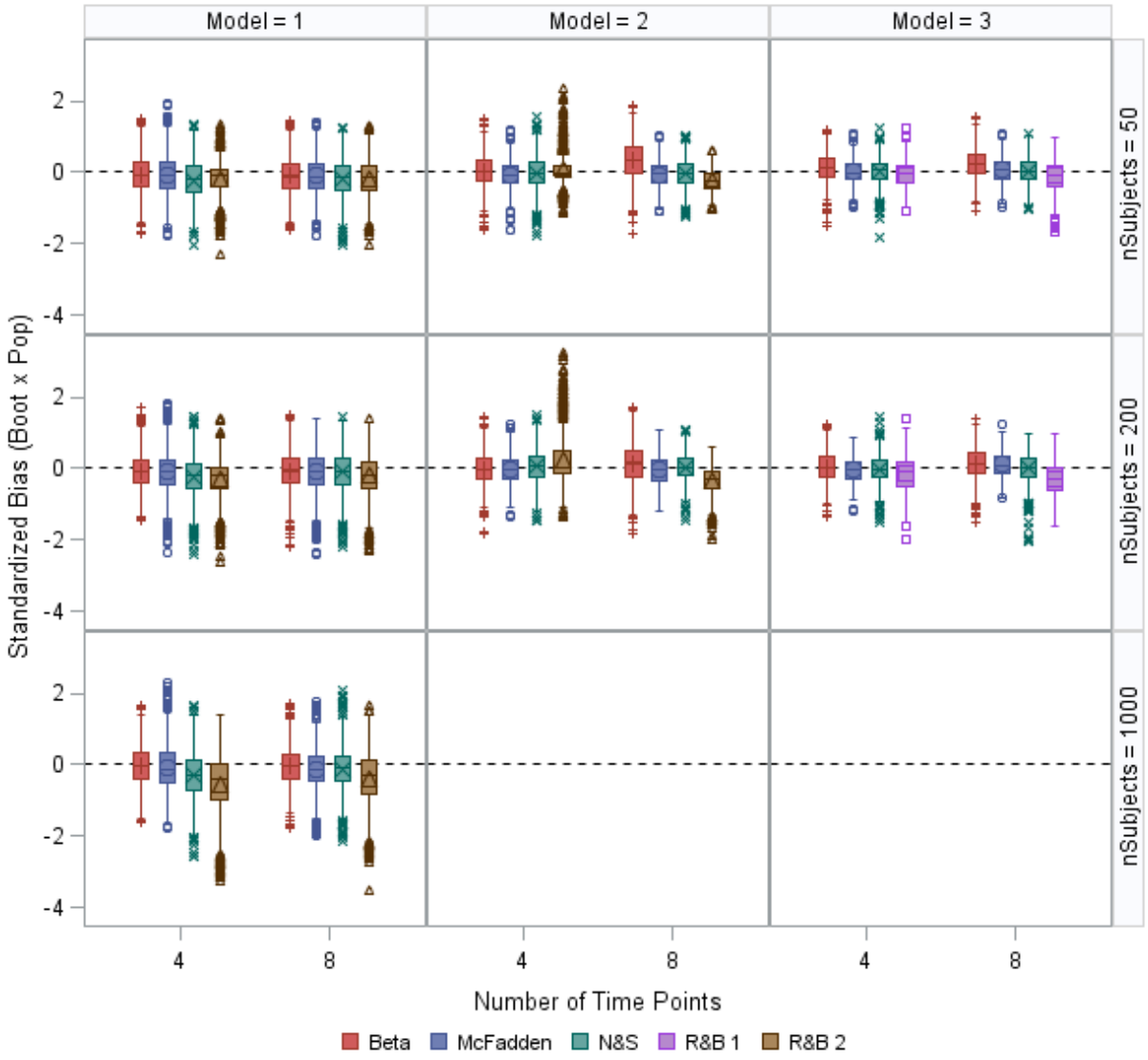

Figure 9 Average standardized bias for bootstrap vs population DA measures.

Figure 10 shows the distribution of the standardized bias values of the general dominance measures estimated in the bootstrap samples compared to the corresponding population and parent sample values. The bias between the SRS and population measures is again used as a check on the sampling procedure. As expected, bias between the population DA measures and bootstrap sample estimates were larger on average than between bootstrap and parent samples. The pattern of bias comparing SRS and population values matches that of the bias comparing bootstrap and population

values, indicating that the bootstrapping procedure did not introduce a significant amount of bias in the measures. Overall, except for R&B R^2 measures, standardized bias was small (within 0.2 standard deviations of the population values).

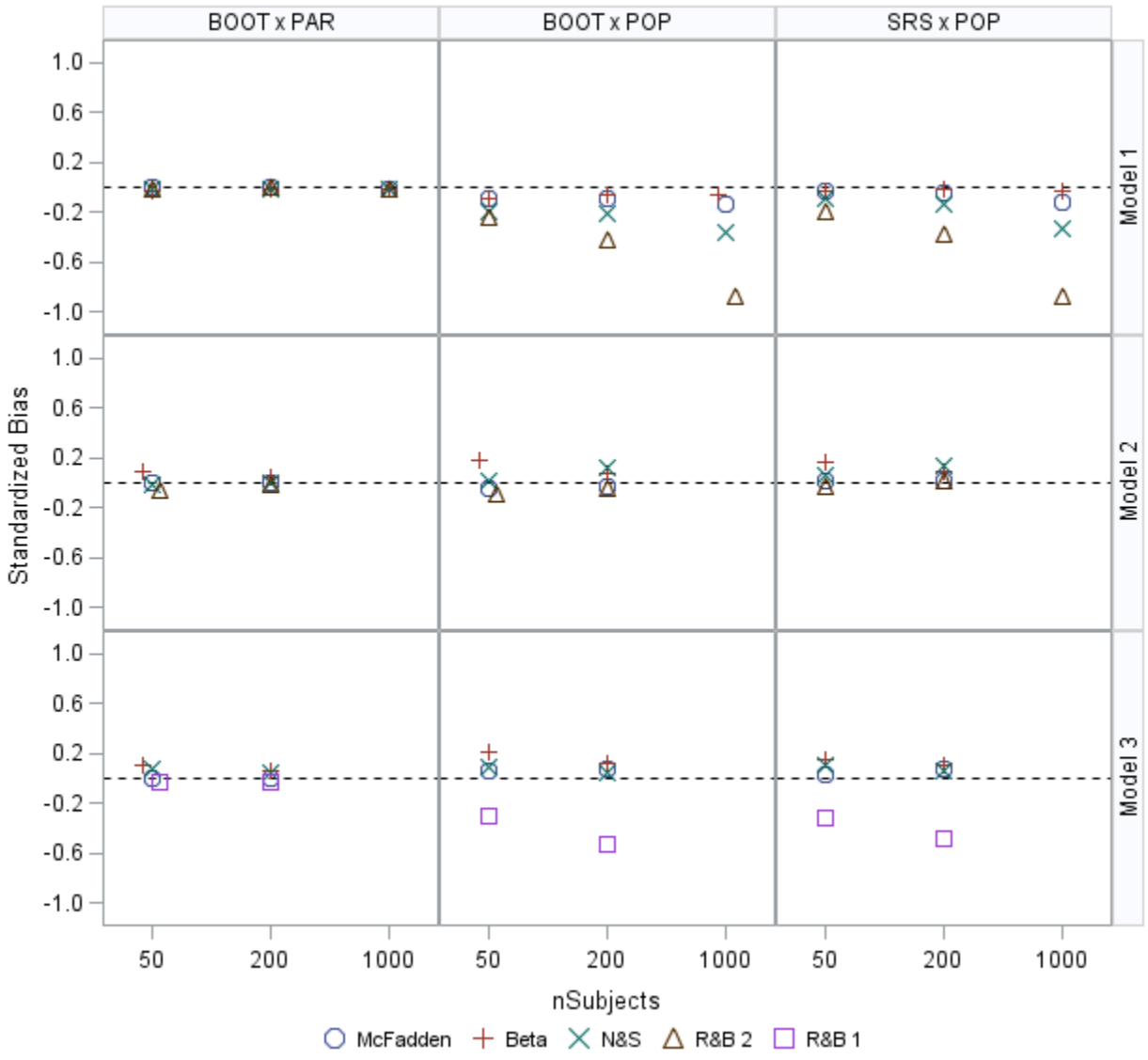

Figure 10 Average standardized bias values for the general dominance measures estimated by the bootstrap vs parent sample (left), bootstrap sample vs population (middle) and SRS vs population (right).

Inference

Statistical inference for the estimated difference between the general dominance measures (i.e., G_{ij} values) was carried out by using two types of confidence intervals: asymptotic normal (ANCI) and percentile (PCI). The asymptotic normal 95% CI for the G_{ij} parameter is constructed for each sample as: $CI_{95\%} = \hat{G}_{ij} \pm Z_{.05}s$, where $\hat{G}_{ij}$ is the general dominance difference estimate for each pair of predictors averaged across all samples (SRS or bootstrap), $Z_{.05} = 1.96$, and *s* is the standard deviation of all $\hat{G}_{ij}$ from all (SRS or bootstrap) samples. Percentile 95% confidence intervals are constructed by ranking the estimated general dominance values, G_{ij}^{s} or G_{ij}^{b} , obtained from all samples (either SRS or bootstrap) and selecting the values corresponding to the 2.5th and 97.5th percentiles as the end points of the confidence interval.

1. *CI Coverage.*

Confidence interval coverage was averaged across all G_{ij} dominance measures by simulation condition and measure of fit to obtain an overall coverage rate per condition. The coverage rate is the number of intervals that contained the corresponding population parameter, converted to a proportion out of the S=100 replications. Additionally, since coverage is calculated for each G_{ij} measure, to obtain a single rate per condition the coverage for all G_{ij} measures were averaged within each condition. For instance, model 3 has 28 G_{ij} dominance pairs since it has 8 predictors; therefore, the average coverage rate for each simulation condition combination within model 3 will be an average of the 28 G_{ij} coverage rates produced by that condition. Average confidence interval coverage was close to the .95 (95%) nominal rate for most conditions for both the asymptotic normal (ANCI) and the percentile (PCI) confidence intervals (Table 16). Since both CI methods produced very similar results in terms of coverage, only the asymptotic normal CI is analyzed in depth. Figure 11 shows the coverage rates by model, sample size and measures of fit,

averaged over the other simulation conditions. Coverage rates were close to the 95% nominal levels across sample size and model complexity conditions for all but the R&B R^2 measures.

Table 16 Confidence interval coverage rates for general dominance measures by sample size combination, confidence interval type, and measure of model fit.

Coverage (%)		**nSubjects - nTimePoints**						
CI Type	**R^2**	**50 - 4**	**50 - 8**	**200 - 4**	**200 - 8**	**1000 - 4**	**1000 - 8**	**Mean**
Percentile	McFadden	95	95	95	96	96	97	95
	Beta	94	93	94	94	94	94	94
	N&S	93	94	92	94	90	94	93
	R&B	93	93	79	89	78	83	88
Asymptotic Normal	McFadden	96	95	95	94	94	94	95
	Beta	94	94	95	94	95	95	94
	N&S	94	94	92	94	90	93	93
	R&B	94	94	79	89	77	84	88

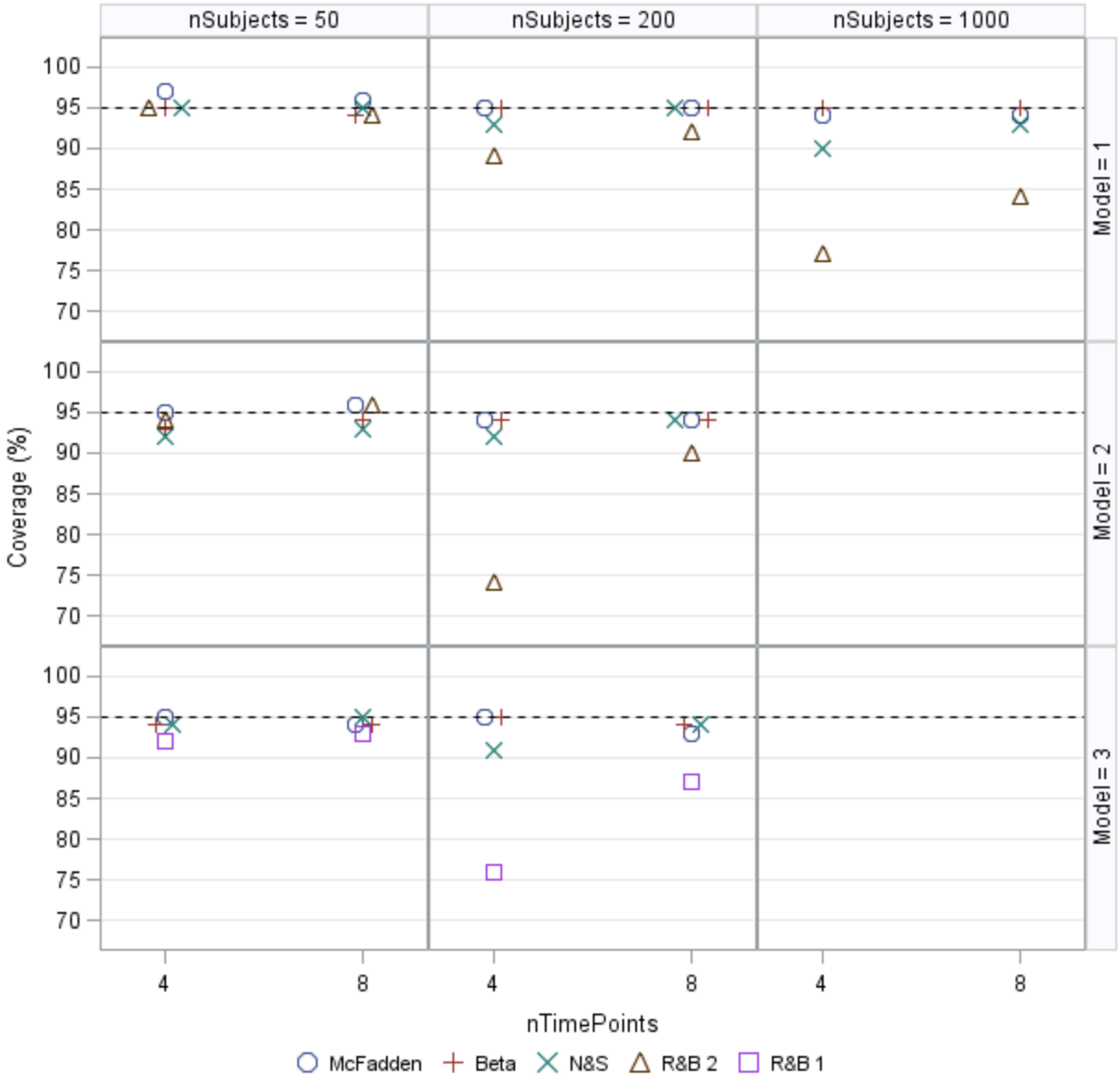

Figure 11 Asymptotic normal confidence interval coverage averaged across collinearity and predictor effects condition.

2. *CI Width.*

The width of the confidence interval was computed as the range of the CI, averaged across all bootstrap or SRS samples for each condition. Additionally, the CI width for the G_{ij} measures were averaged within each condition. The mean estimates for 95% confidence interval widths for

the general dominance measures appeared to be reasonable when considering the population range of the G_{ij} measures (see Table 7) and were very similar (within .005) of SRS CI width results (not shown). CI width values yielded intuitive results as they related to the factors explored in this study. That is, as expected, CI width generally decreased with an increase in sample size at both the subject and the observation (time) levels. For models 2 and 3, the R&B2 R^2 measure, which corresponds to the proportional change in variance in the random intercept variance component, produced extreme values of confidence interval width and were removed from the analysis. This issue might be related to the presence of non-positive definite matrices in some models. The average confidence interval width results for the other measures of model fit (R^2) are presented in Table 17. Confidence interval width results are consistent across the percentile and asymptotic normal methods, showing minimal variation across these methods.

Table 17 Average confidence interval width for the general dominance measures by sample size combination, confidence interval type and measure of model fit.

CI Width		**nSubjects - nTimePoints**						
CI Type	**R^2**	**50 - 4**	**50 - 8**	**200 - 4**	**200 - 8**	**1000 - 4**	**1000 - 8**	**Mean**
Percentile	McFadden	.036	.021	.017	.010	.007	.005	.018
	Beta	.226	.209	.111	.102	.062	.068	.145
	N&S	.141	.076	.068	.038	.034	.019	.071
	R&B 1	.134	.078	.068	.039			.080
Asymptotic Normal	McFadden	.036	.021	.017	.010	.007	.005	.018
	Beta	.224	.208	.111	.102	.062	.069	.144
	N&S	.140	.076	.068	.038	.034	.019	.071
	R&B 1	.134	.078	.068	.040			.080

3. *Type I Error.*

Type I error in this study refers to the (false) detection of a general dominance relationship between two predictors when the population dominance difference G_{ij} is zero (i.e., the null hypothesis of no dominance is true but is rejected). False detection rates were evaluated at the 0.05 alpha level (5% rate) and were averaged across all the null dominance measures, which are the population G_{ij} measures formed by G_i values that are equal to each other in Figure 1 (model 1), Figure 2 (model 2), and Figure 3 (model 3). If the 95% CI in any of these cases did not include 0 then the null hypothesis was considered to be (falsely) rejected.

Using Bradley's (1978) liberal criterion of robustness, a test can be considered robust, and thus acceptable, if the empirical type I error rate is within the interval $\alpha \pm 0.5\alpha$, which, for $\alpha = .05$, implies a range between 2.5 and 7.5%. Table 18 lists the average type I error rates by CI type, measure of fit, and leve-2 sample size. A large proportion of type I error values were below the nominal 5% level, especially under the lower level-2 sample size condition. As sample size at level-2 increased, type I error rates approached the nominal level with a few exceptions, most notably under the McFadden measure, which got more conservative as sample size increased in the percentile CI type. As can be seen in Figure 12, very few instances went over the upper limit of the acceptable range. For the asymptotic standard error interval, dominance measures using the N&S R^2 and the R&B R^2 were overly conservative when the sample size at level-2 was 50.

Table 18 Average type I error rate by CI type, R^2 measure and level-2 sample size.

	nSubjects	50	200	1000
CI	R^2	Average Type I Error		
PCI	McFadden	3.2	2.7	1.3
	Beta	3.6	4.5	7.5
	N&S	3.6	3.9	5.1
	R&B 2	3.0	3.7	3.0
ANCI	McFadden	2.7	3.9	4.7
	Beta	3.3	3.5	5.0
	N&S	2.1	3.4	5.0
	R&B 2	1.3	2.8	4.0

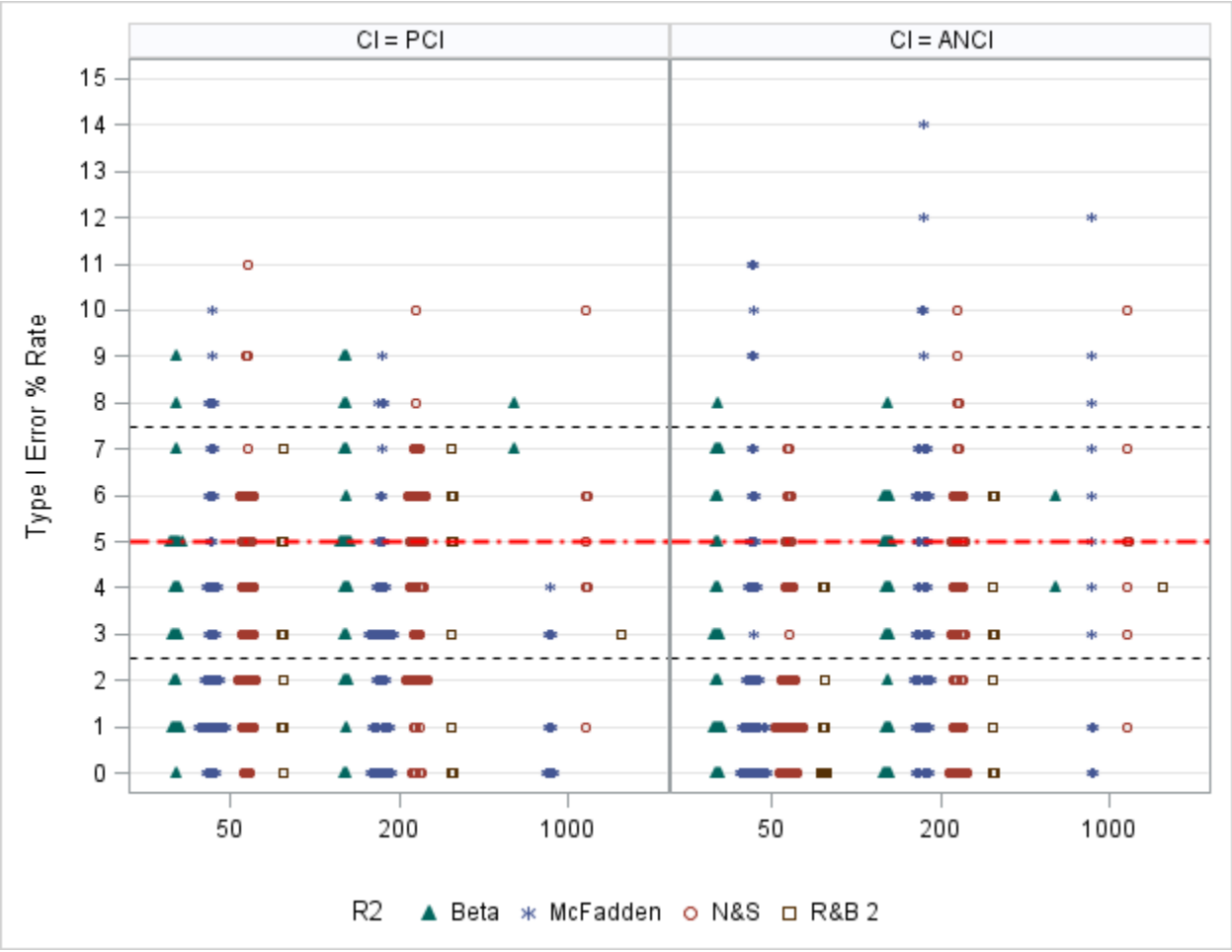

Figure 12 Type I error rate across all G_{ij} measures with a population value of zero.

[Note: The straight, horizontal black dashed lines represent acceptable Type I errors rates between 2.5% and 7.5%. The red line represents the nominal 5% rate.]

4. *Power.*

The power of the inferential procedure was calculated as the proportion of replications where the non-null population dominance relationships ($G_{ij} \neq 0$, rounded to two decimal points) was detected by the 95% confidence interval (i.e., zero was not included in the interval so the null hypothesis of zero dominance was rejected). In general, power rates of 80% or above are considered adequate.

Figure 13 shows the power rates by values of the population dominance absolute effect G_{ij} for each sample size and predictor level combination. The collinearity factor did not seem to impact power as much as the predictor level (related to model complexity) and sample size, so the results here were averaged across the collinearity factor. Since the results from the asymptotic normal and the percentile confidence intervals were very similar, here only the former (ANCI) is presented. It can be seen from Figure 13 that, when comparing the relative importance of level-2 predictors, a sample size of at least 200 subjects is needed to obtain adequate power. Additionally, the McFadden and N&S R^2 measures seem to result in higher power for these comparisons. A list of the minimum effect size of the general dominance difference (G_{ij}) needed to obtain 80% power by sample size, predictor type and measure of fit is provided in Table 19.

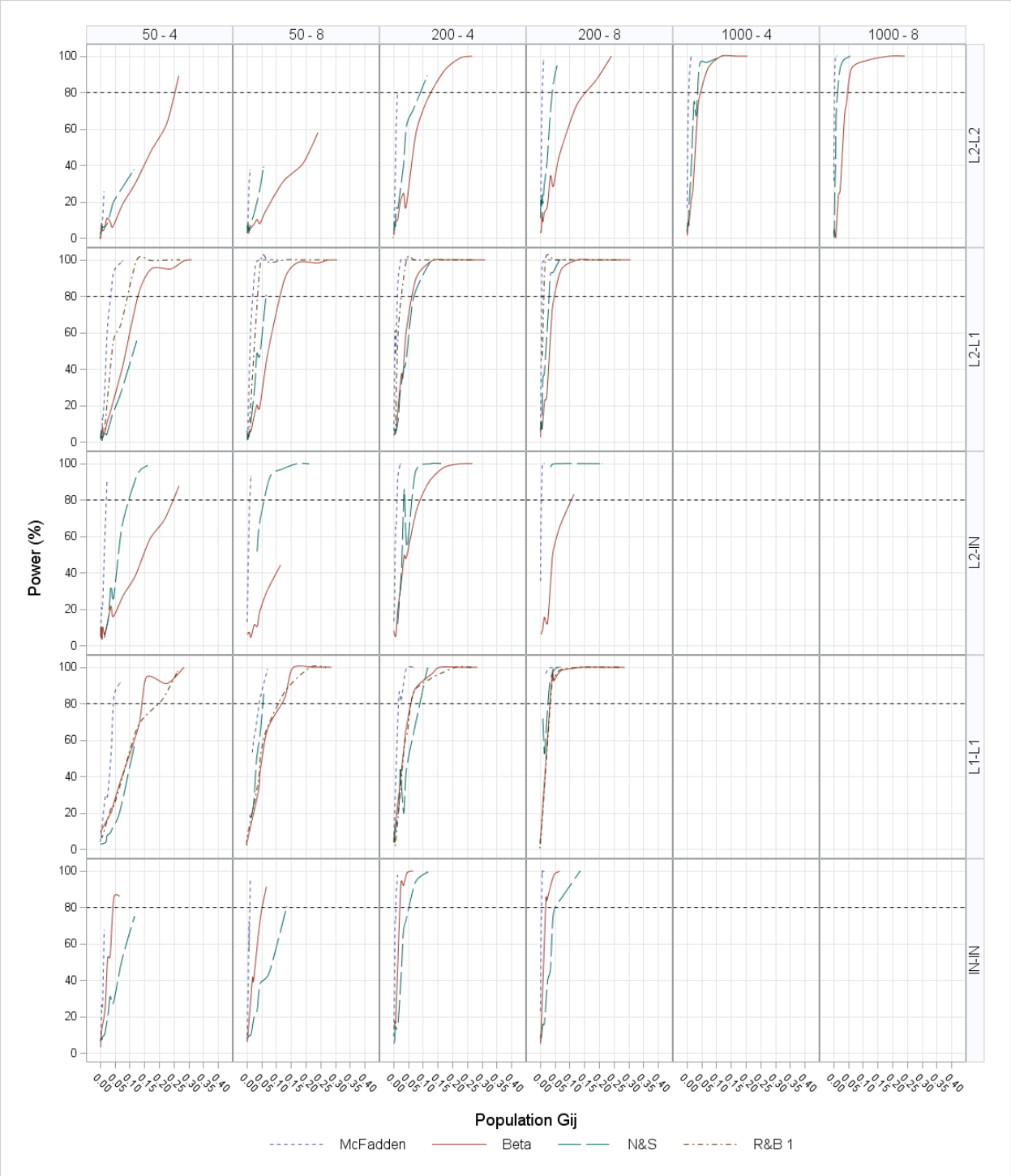

Figure 13 Power rates obtained with the asymptotic standard error CI by the absolute value of the population general dominance difference (G_{ij}) across sample sizes (columns), predictor level (rows), and measure of fit (lines).

Table 19 Minimum general dominance (G_{ij}) effect size to achieve 80% power per measure of fit, sample size and predictor level.

	nSub-nTimePts:	50 - 4	50 - 8	200 - 4	200 - 8	1000 - 4	1000 - 8
R2	**Preditor Level**						
McFadden	IN-IN	0.01	0.01	0.01	0.004		
	L1-L1	0.05	0.03	0.02	0.02		
	L2-IN	0.02	0.01	0.01	0.003		
	L2-L1	0.03	0.01	0.01	0.01		
	L2-L2			0.01	0.004	0.004	0.002
Beta	IN-IN	0.05	0.05	0.02	0.02		
	L1-L1	0.14	0.08	0.06	0.04		
	L2-IN	0.19	0.09	0.05	0.04		
	L2-L1	0.09	0.07	0.06	0.05		
	L2-L2	0.27		0.07	0.06	0.03	0.04
N&S	IN-IN		0.14	0.03	0.04		
	L1-L1		0.06	0.07	0.01		
	L2-IN	0.07	0.04	0.03	0.04		
	L2-L1		0.06	0.05	0.02		
	L2-L2			0.02	0.01	0.02	0.01
R&B 1	L1-L1		0.14	0.26	0.05		
	L2-L1	0.2	0.11	0.11	0.05		
R&B 2	IN-IN				0.04		
	L2-IN				0.19		
	L2-L2				0.18	0.11	0.04

Note: The predictor level refers to the level of predictors in the pair, where L2 represents a person-level predictor, L1 represents a time-varying predictor, and IN represents the cross-level interaction between a level-2 predictor and the Time trend effect.

Figure 14 shows a comparison of the power rates for the different measures of fit, for both confidence interval types, disaggregated by model and sample size. This figure shows that for model 1, only the McFadden R^2 is able to achieve reasonable power and only for 200 subjects and 8 time points, or 1000 subjects. For model 2, 80% power was only achieved by the McFadden and the N&S R^2 measures at the highest sample size combination (nSubjects=200 and nTimePoints=8).

For model 3, which contains more predictors and a wider range of effects, 80% power was achieved by the McFadden and the R&B R^2 (change in variance of the level-1 residuals) with 4 time points when the number of subjects was 200 and, with 8 time points, for both 50 and 200 subjects.

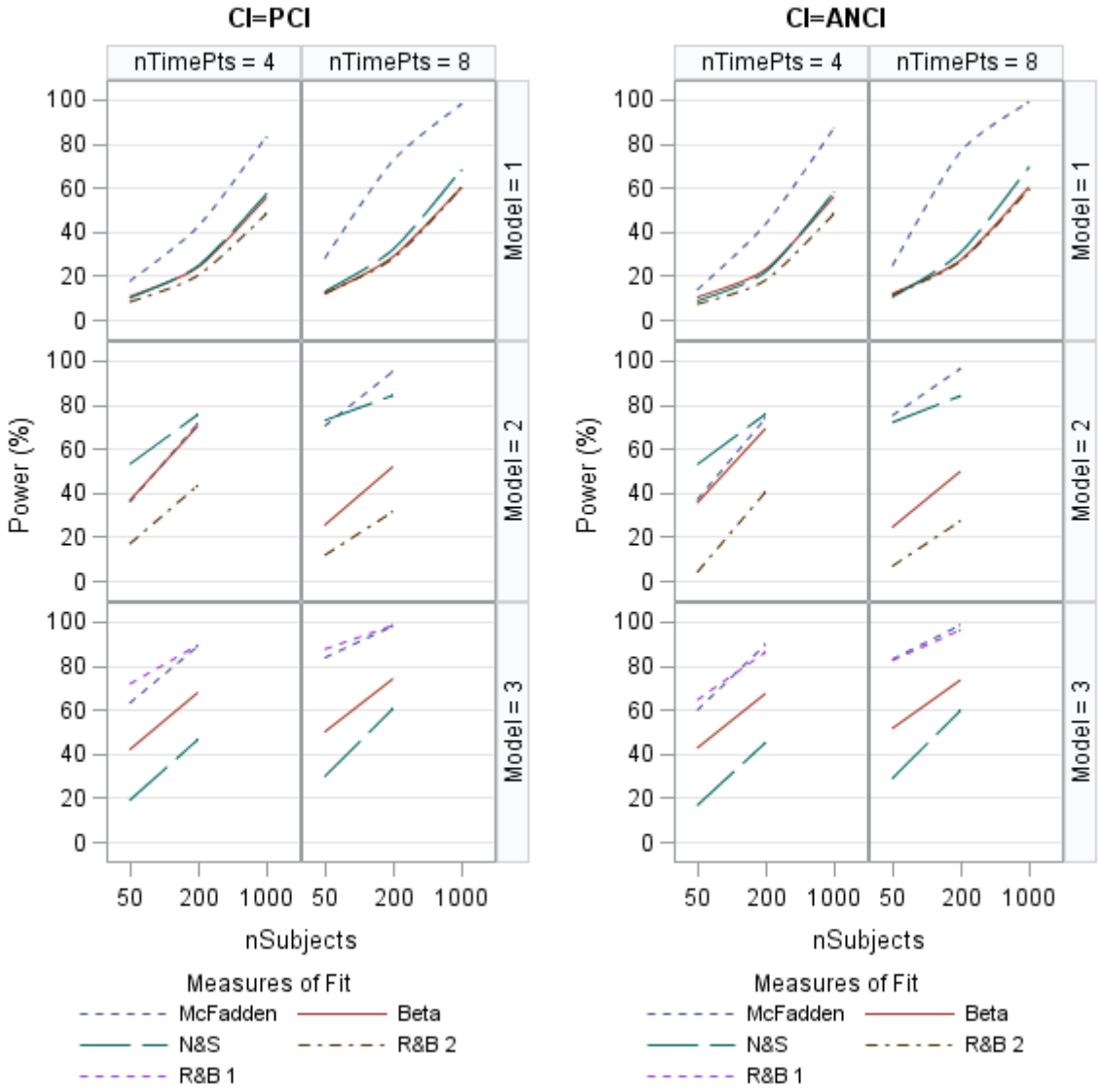

Figure 14 Average power rates (%) of the non-null dominance measures by sample size and model for each measure of fit averaged across all dominance measures and collinearity conditions.

Reproducibility

In order to investigate the factors associated with bootstrap reproducibility rates (proportion of bootstrap samples where the qualitative dominance measure D_{ij} agreed with its corresponding population value), reproducibility of all D_{ij} values were averaged across simulation conditions. Then the relationship between average reproducibility and the magnitude of the population general dominance relationship G_{ij} was also investigated.

Examination of the factors related to the proportion of bootstrap samples that reproduced the population dominance relationship indicated that the most influential factors were the predictor fixed effects condition (η^2 = .15), the model complexity (η^2 = .11), sample size at level-2 (η^2 = .10), and the interaction between measure of fit and model complexity (η^2 = .08). The ANOVA of model 1 results indicated that, within this model, the most influential factors associated with average reproducibility were the predictor fixed effects condition (η^2 = .25), sample size at level-2 (η^2 = .24), and the level of collinearity (η^2 = .06).

Reproducibility rates for the factors deemed influential are presented descriptively in Table 20. Reproducibility was on average greater than .70 across measures of fit and sample size conditions. Reproducibility increased with an increase in level-2 sample size and with an increase in effect size. For model 1, all measures of model fit performed similarly well in terms of reproducibility. For model 2, the N&S R^2 measure was the best performing, and for model 3, the R&B R^2 (proportion change in variance of the level-1 residual) performed better than all the other measures, with McFadden R^2 being second best. Table 21 contains the model 1 reproducibility averages for factors deemed influential in this model's ANOVA results. There is a noticeable increase in average reproducibility with an increase in sample size. On the other hand, the increase in collinearity between predictors negatively impacts reproducibility.

The relationship between reproducibility and the magnitude of the general dominance measures in the population can be visualized in Figure 15. When parent sample reproducibility was lower than 0.8, it seems to indicate that the magnitude of the dominance relationship in the population is very low or zero.

Table 20 Average reproducibility rates of the population general dominance relationship.

		Baseline/ Base-Base			**Small Effect/ Base-Large**			**Large Effect/ Large-Large**			
Model	**R^2 /nLvl2**	**50**	**200**	**1000**	**50**	**200**	**1000**	**50**	**200**	**1000**	**All**
1	McFadden	.43	.55	.70	.54	.61	.70	.70	.84	.95	.67
	Beta	.56	.66	.77	.58	.66	.77	.72	.86	.97	.73
	N&S	.48	.58	.70	.57	.65	.76	.71	.85	.96	.70
	R&B 2	.56	.66	.78	.57	.65	.76	.70	.84	.96	.72
2	McFadden	.63	.75		.81	.89		.81	.89		.80
	Beta	.72	.82		.75	.87		.81	.90		.81
	N&S	.78	.87		.85	.91		.89	.95		.87
	R&B 2	.59	.71		.69	.82		.67	.80		.71
3	McFadden	.71	.79		.82	.90		.89	.95		.84
	Beta	.65	.77		.79	.88		.83	.92		.81
	N&S	.48	.60		.73	.86		.69	.81		.70
	R&B 1	.82	.89		.87	.97		.89	.92		.89

Table 21 Model 1 average reproducibility rate of the population general dominance relationship.

Reproducibility	**Baseline**			**Small Effect**			**Large Effect**			**All**
Collinearity	50	200	1000	50	200	1000	50	200	1000	
0	.53	.68	.79	.60	.69	.86	.77	.91	.99	.76
0.5	.50	.62	.75	.57	.66	.78	.71	.86	.97	.71
0.8	.49	.54	.67	.53	.57	.61	.65	.78	.91	.64

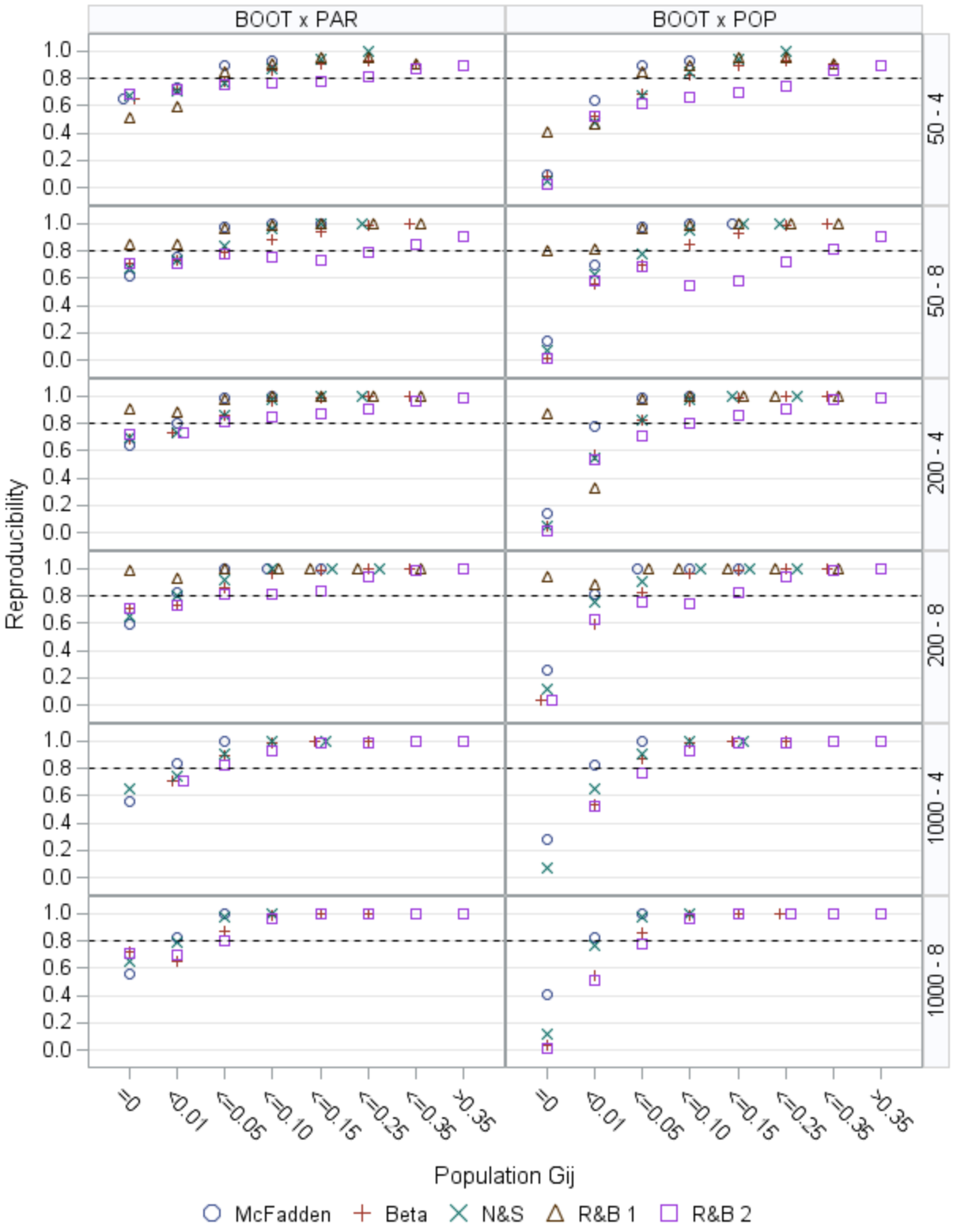

Figure 15 Reproducibility rate of parent sample (left) and population (right) qualitative general dominance relationship (D_{ij}) in the bootstrap samples according to population quantitative dominance effect (G_{ij}).

In practice, applied researchers will only be able to obtain the reproducibility rates of the parent sample dominance relationships; that is, the proportion of bootstrap samples that agreed with the parent sample qualitative dominance values for each pairwise qualitative dominance measure D_{ij}. Figure 16 shows a comparison of the average reproducibility rates by model, measure of fit, and (level-2) sample sizes. Figure 16 also shows the reproducibility rates of the population qualitative dominance relationships D_{ij} in the simple random samples, which is used as a check on the performance of the bootstrap reproducibility results.

The pattern and magnitude of reproducibility rates achieved by the bootstrap and SRS samples are very similar, indicating that the bootstrapping procedure worked appropriately to replicate the sampling distribution of these dominance measures. The pattern of reproducibility of the population values by the bootstrap samples was very similar to the reproducibility of the parent sample values, indicating that reproducibility of the parent sample computed by the bootstrap procedure can be used as an approximation of the reproducibility of the population values after adjusting the magnitude of the rate. Results from the bootstrap procedure indicate that when number of subjects is 50 the reported reproducibility (of the parent sample result) overestimates the reproducibility of the population result by about .15 points, when number of subjects is 200 the overestimation is of about .10 points, and with 1000 subjects it is less than .05.

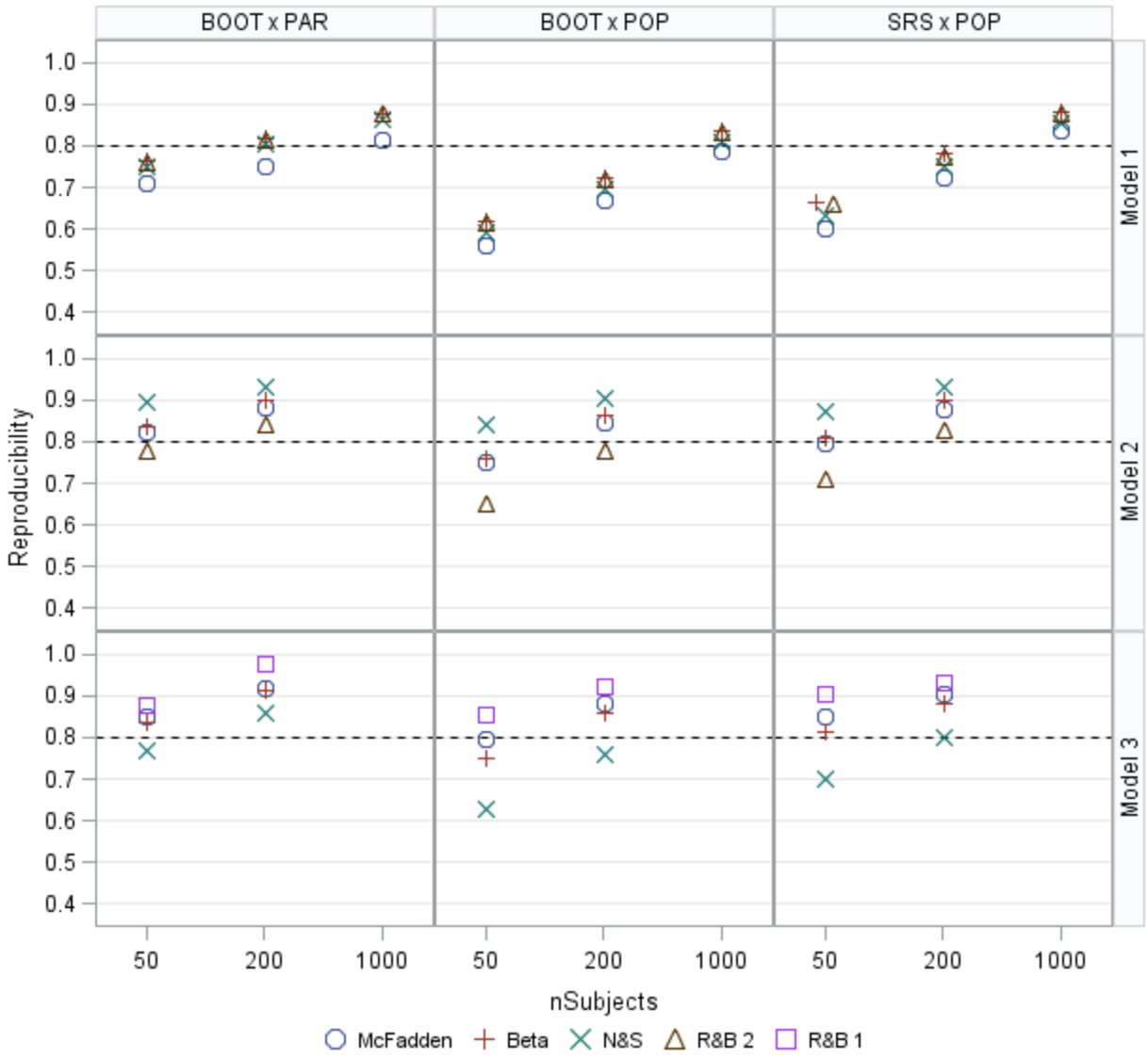

Figure 16 Average reproducibility rates of the parent sample (left) and population (middle) dominance relationships in the bootstrap and simple random samples (right) across models and sample sizes.

Summary

Table 22 Key study results per outcome measure.

Outcome Measure	Key Results
Agreement in most important predictor	- Rates of agreement increased with an increase in level-2 sample size. - McFadden R^2 measure produced the highest rates of agreement followed by R^2 Beta. - R^2 Beta performed better in models with a cross-level interaction. - N&S R^2 performed poorly in models with time-varying predictors (model 3) under the baseline and large effect conditions. - True rate of agreement between bootstrap and population was about 20% lower than the observed agreement between bootstrap and parent sample.
Agreement in least important predictor	- Rates of agreement increased with an increase in level-2 sample size. - Conditions of small effect sizes produced lower agreement rate than baseline effects, probably due to sampling error since population G_{ij} values are very close to zero in this condition (see Population DA section). - McFadden R^2 measure produced the highest rates of agreement on average for models 1 and 2, followed by the N&S R^2. - For the high model complexity condition (model 3), R&B1 R^2 (proportion change in variance of the level-1 residual) performed better than all the other measures and R^2 Beta was the worst performing measure. - True rate of agreement between bootstrap and population was between 5-15% lower than the observed agreement between bootstrap and parent sample.

Overall rank order correlation (Kendall's tau)	- As expected, Kendall tau correlations increased with level-2 sample size and model complexity and decreased with higher collinearity among predictors. - In the low model complexity condition (model 1), correlations were too low when level-2 sample size was only 50. - All R^2 measure performed comparably well when ranking level-2 predictors only (model 1), but the N&S R^2 produced the highest average correlations for models with cross-level interactions (model 2), and the R&B1 R^2 performed better in models with both level-1 and level-2 predictors (model 3). - Against expectations, higher predictor fixed effect was not consistently associated with higher tau correlations. - Patterns of correlation (agreement) between bootstrap and population matched the patterns between bootstrap and parent samples. The true correlation between population and bootstrap was about .2 points lower than the observed correlation between bootstrap and parent sample.
Bias	- Standardized bias values were low (within 0.2 standard deviations from the population values) for all but the R&B measures. - Pattern of bias between the bootstrap and parent samples matched the pattern between bootstrap samples and population as well as between SRS and population.
CI Coverage	- Average confidence interval coverage was close to the 95% nominal rate across sample size and model complexity conditions for all but the R&B R^2 measures. - Asymptotic normal and the percentile confidence intervals performed similarly well in terms of coverage.
CI Width	- Average CI width estimates were reasonable given the population range of the G_{ij} measures for each measure of fit. - The R&B 2 R^2 measure produced extreme values of confidence interval width for the more complex models (models 2 and 3). - CI width decreased with an increase in sample size at both the subject and time levels. - CI width results were consistent across the percentile and asymptotic normal methods.

Type I error rate	- Type I error rates were below the nominal 5% level for most conditions, especially under the lower sample size conditions. - Type I error rates approached the nominal level as level-2 sample size increased except for the McFadden measure under the percentile CI method.
Power	- Power was low for making inferences about the relative importance of level-2 predictors (model 1) when sample size at level-2 was 50 subjects. - As expected, power increased with an increase in sample size at both levels. - A puzzling finding was that the relationship between power and population dominance effect size was non-monotonic in certain instances, particularly at lower sample sizes and at the lower range of the population general dominance (G_{ij}) effect. - R^2 Beta was particularly problematic in the sense of resulting in a non-monotonic relationship between power and dominance effect size. - For comparing level-2 predictors (model 1), only McFadden attained 80% power and only when level-2 sample size was 200 and level-1 sample size was 8, or level-2 sample size was 1000. - When comparing level-2 and cross-level interactions, 80% power was attained by McFadden and N&S R^2 at the highest sample size combination (200 at level-2 and 8 at level-1). - When comparing predictors of the random intercept and type-varying predictors, the R&B 1 (model 3) and McFadden measures were able to attain adequate power when sample size was 200 at level-2 or 8 at level-1.
Reproducibility	- Reproducibility increased with an increase in level-2 sample size and with an increase in effect size, and decreased with an increase in collinearity between predictors. - Reproducibility was lower for the baseline fixed effects condition, as expected. - When comparing level-2 predictors only (model 1), R^2 Beta produced the highest level of reproducibility. - When number of subjects is 50, the observed reproducibility rate (of the parent sample result) overestimates the reproducibility rate of the population result by about .15 points; when number of subjects is 200 the overestimation is of about .10 points; and with 1000 subjects it is less than .05.

CHAPTER 5. DISCUSSION

The purpose of this study was to evaluate the use of dominance analysis for determining the relative importance of predictors in multilevel models for longitudinal data with continuous outcomes. A simulation study was conducted to investigate the impact of model complexity, sample size, collinearity, and covariance misspecification on the accuracy of dominance analysis results in terms of the rank-ordering of predictors by relative importance, the performance of bootstrap-based inferential procedures for the quantitative general dominance measure, and the reproducibility rates of the qualitative general dominance measure over many bootstrap samples. The effect of using different measures of model fit for computing general dominance was also investigated.

One issue identified in this study that was not part of the original study design was the prevalence of non-positive definite random-effect covariance matrices under small sample sizes and higher model complexity. Researchers using multilevel models for longitudinal analysis must be mindful of this issue since there is a potentially large number of parameters to be estimated in models with random slopes and unstructured covariance matrices, which, when coupled with a large number of fixed effects parameters and the usually small number of observations at level-1 (time points), can create problems for algorithms searching for a positive-definite solution and might produce unreliable estimates of the random effects components. Simulation results indicated that the use of an unstructured covariance structure with autoregressive residuals was unfeasible for sample sizes with 50 subjects even when there were only four predictors in the model, and might not produce reliable estimates unless the number of subjects is at least 1000. A solution to this problem was to simplify the covariance structure by relaxing the assumption of autoregressive

errors and use what has been referred to in the longitudinal literature as the "standard" multilevel model for change (Singer & Willett, 2003). This covariance structure specification includes a random intercept for the subjects, a random slope for the effect of time, and a covariance between these two components, but assumes that the errors at level-1 are *i.i.d*. In terms of the current study, the prevalence of non-positive definite matrices was potentially problematic because it impacted the calculation of measures of model fit that depend on the random-effects estimates. Therefore, the data analysis was conducted assuming a misspecified covariance structure. However, comparative analyses conducted for simulation conditions where the covariance specification was correct indicated that this misspecification of the covariance structure did not have a noticeable impact on the outcome measures analyzed here.

Main findings

Ranking accuracy. Accurate ranking of predictors (compared to the population ranking) was impacted mostly by sample size at level-2 and model complexity. The higher the number of subjects the more accurate the bootstrap results are in reflecting the population predictor rankings. The simulation results indicate that at least 200 subjects might be necessary to adequately reproduce the population rankings. Model complexity also impacted the rankings such that the more predictors in the model, the higher the agreement rates. Having a larger number of predictors increases the sample size of the correlation coefficient for the agreement between the rank orderings of predictors.

As expected, predictor effect sizes impacted ranking accuracy through their effect on the population general dominance effect. As shown in Figure 1, in model 1 the values of the general dominance effects followed a pattern similar to the predictor coefficients, even though collinearity

attenuated the effects in the large effect condition. In model 2 (Figure 2), the effects differed by measure of fit. The McFadden and N&S R^2 tended to favor (i.e., assign higher G_i values to) the cross-level interaction terms and the R^2 Beta and R&B2 R^2 tended to assign higher dominance weights to the level-2 predictors. This is likely related to how these measures are computed since, for example, R&B2 R^2 measures proportional change in variability in the random intercept, which is more strongly impacted by level-2 variables. However, the R&B2 R^2 measure did not work well for this model as it produced negative dominance weights for some conditions. For model 3 (Figure 3), which compared time-varying and person-level predictors, all measures assigned higher dominance weights to the level-1 predictors. The R&B1 R^2 measure, which is the proportion change in variance of the level-1 residuals, was not able to differentiate between level-2 predictors but worked well for the level-1 predictor comparisons. Therefore, this measure should only be used when the goal is to compare pure level-1 predictors.

When general dominance effects are small, the DA procedure might not be able to differentiate among predictors very well, and collinearity pulled the dominance effects closer to each other, making ranking estimates less clear-cut. Measures of fit that assigned similar dominance weights to predictors, such as the N&S and Beta R^2 measures in certain conditions of models 2 and 3, produced lower agreement values because sampling error produced samples that did not replicate these small differences. The effects of collinearity on ranking accuracy was not as problematic with the moderate collinearity of 0.5, but ranking accuracy did suffer when collinearity was as high as 0.8. Therefore, predictor importance rankings produced by DA seem to be robust to the presence of moderate collinearity.

Overall, the agreement results between bootstrap and population rankings were consistent with what would be observed between the bootstrap and parent samples, but at an order of

magnitude lower. Therefore, in practice, measures of agreement (Kendall tau correlations) obtained with a single sample using the bootstrapping procedure should be attenuated by about .2 points if the number of subjects (sample size at level-2) is 50, .1 point for 200 subjects, and .05 points for 1000 subjects.

Bias. Standardized bias between the population general dominance measures and their bootstrap estimates were generally low, not surpassing .2 standard deviations on average for most conditions and measures of fit. The only factors identified as having a moderate effect on bias were measure of fit and model complexity. Patterns of bias did not vary between the SRS and bootstrapping sampling procedures, indicating that the bootstrapping procedure did not introduce a large amount of bias in the estimation of general dominance measures. The high standardized biases found in dominance measures using the R&B R^2 may be explained by three different factors. The first is the fact that the models were estimated using full maximum likelihood, which is known to underestimate the variance of the random effects (Gurka, Kelley & Edwards, 2012). The second issue is the fact that these measures, in particular the R&B 2 measure, were more impacted by the npd issue and therefore their estimates are less reliable. The third possible explanation is the fact that the standard deviation of the dominance measures decreased more rapidly than the decrease in bias as sample size increased, therefore causing the standardized values to decrease. Indeed, the average raw bias values (not shown) showed a decrease with sample size.

Inference. Two bootstrap-based methods, namely the percentile and the asymptotic normal methods, were evaluated for making inferences about the significance of the difference between the general dominance measures. These methods were evaluated by calculating confidence interval coverage (the proportion of replications where the constructed CI included the population parameter), average width, type I error rates (proportion of replications in which the confidence

interval did not contain zero even though the population general dominance measure was zero), and power (proportion of samples where the population G_{ij} parameter was non-zero and the confidence interval did not contain zero). Overall, the percentile confidence interval produced similar results to the asymptotic normal CI. Coverage rates were close to the nominal 95% rate across simulation conditions and measures of fit, indicating that the confidence intervals worked as expected. However, the data were generated from a multivariate normal distribution; therefore, these results are not surprising. In terms of confidence interval width, CI width shrunk with an increase in sample size and was generally not problematic.

However, the confidence intervals may not work properly for making inferential claims about the dominance measures. Type I error rates were well below the 5% nominal levels for most of the dominance measures in this study. In general, type I error tended to increase with an increase in level-2 sample size but remained below the nominal 5% rate, indicating that the inferential procedure might be overly conservative. Type I error rates approached the nominal level only when level-2 sample size was 1000. The results from analyzing the power of the inferential procedure corroborates the hypothesis of an overly conservative procedure, with few conditions being able to achieve the usual 80% power rate. One design factor that may have impacted the type I and power rates was the fact that some G_{ij} measures, while not exactly zero in the population, were very small (near zero). Therefore, the near-zero differences in the population (i.e., presence of very small effects) might have lowered the average power rate. Overall, the results suggest that to use either the percentile or the asymptotic normal confidence interval procedures, either a large sample size or a large effect size is needed. The values in Table 19 could provide some guidelines for the minimum effect sizes and sample sizes needed to achieve appropriate power with these procedures for different predictor types and measures of fit.

Reproducibility. The proportion of bootstrap samples where the sample D_{ij} measure matched the population D_{ij} measure was used to evaluate whether the sample dominance relationship was likely to reflect the population dominance relationship. The factors that seemed to most strongly impact the rates of reproducibility were the magnitude of the fixed effects, the model complexity, and sample size at level-2. Reproducibility increased with an increase in level-2 sample size and dominance effect size and decreased with the increase in collinearity between predictors. As a guideline for applied researchers, the simulation results indicate that when sample reproducibility was lower than 0.8 the magnitude of the dominance relationship in the population is likely to be zero or very small. Because reproducibility is calculated based on the qualitative dominance measure D_{ij}, which only indicates whether the general dominance measure of one predictor is greater than that of another and does not take into account the magnitude of the dominance difference, it can detect small differences in the population which may not hold in the sample (and cause mismatches between the sample and the population). In the same vein, if the true population D_{ij} value is zero, DA using sample data might detect a small dominance difference due to sampling error. Both situations would produce poor reproducibility rates. One potential way to remedy this issue is to round the dominance values appropriately according to the measure of fit. In this study it was found that the McFadden R^2 produced small dominance values, and therefore rounding to three decimal points should be used for this measure. For all other measures, rounding to two decimal points is recommended to avoid detecting spurious dominance differences.

Summary. Overall, results from this study indicate that dominance analysis can be extended to longitudinal multilevel models as an additional tool for researchers wishing to determine the relative contribution of person and time-level predictors of outcomes that change

over time. As expected, among the conditions studied here, sample size at level-2, usually the number of subjects in a longitudinal study, was consistently one of the factors identified as having the largest effect on the outcome measures. A sample size of 50 subjects was found to be insufficient to produce adequate agreement, power, and reproducibility rates. In this study, a minimum sample size of 200 subjects was needed to avoid issues with non-positive definite matrices, especially when the model had more than 4 predictors. General sample size recommendation for longitudinal analyses was not in the scope of this paper, but the results indicate that models with 4 predictors and 50 subjects can be estimated using the SGR covariance structure. However, if more predictors are added to the models, a larger number of subjects might be needed to avoid issues with non-positive definite matrices.

In terms of measures of model fit, DA computed using the McFadden R^2 measure produced the most consistent results across outcome measures unless the dominance effect was very small in the population. The McFadden measure, which is calculated from the deviance readily provided in software output, performs reasonably well and might be used as the standard measure for DA in multilevel models. However, due to the small magnitude of this measure, the power to detect true population relationships can be low unless dominance effects and/or sample size are sufficiently large. For inference purposes, at least 200 subjects and 8 time points may be needed to detect a true (non-zero) relationship using this measure. In general, the adequacy of the measures of fit evaluated in this study varied according to the types of predictors being compared by the DA procedure; therefore, recommendations are made accordingly. If interest is in comparing only predictors of the random intercept (i.e., person-level or level-2), all measures of fit studied here would be adequate as they produced consistent rankings. If the goal is to compare the relative importance of cross-level interactions, such as predictors of the linear effect of time, or to compare

predictors at both the subject (level-2) and measurement (level-1, or time-varying predictors) levels, the McFadden and N&S R^2 tend to perform well as these measures are sensitive to variation at both level-1 and level-2. If interest is in comparing only level-1 predictors, both the R&B1 and McFadden's R^2 produced adequate results. The R^2 Beta performed well in some conditions but is deemed inadequate due to the lack of monotonicity.

It was also demonstrated that bootstrapping can be utilized to construct percentile and asymptotic confidence intervals and to measure the reproducibility of sample results, and that these procedures can provide reliable information regarding the generalizability of the dominance relationships found in the sample. The bootstrap and asymptotic normal confidence intervals seemed to perform similarly; therefore, the simpler asymptotic normal CI is recommended for inference and reproducibility calculations. A reproducibility rate of 80% between bootstrap and parent samples seems to be considered a minimum threshold to deem a sample dominance relationship likely to reflect its corresponding population dominance relationship.

The main contribution of this study was to provide evidence that dominance analysis can be successfully conducted on longitudinal data using multilevel models under some specific conditions. This research provided guidance to applied researchers on appropriate sample sizes to target when wishing to conduct longitudinal analysis in general and dominance analysis in particular. The main considerations for using the DA method in applied settings are the number of subjects and the number and types of predictors that one wishes to rank order in terms of relative importance. When resources are limited, it might be more beneficial to obtain a larger number of subjects than a larger number of measurement occasions, especially if the focus is on subject-level outcomes and predictors. This study also provided a comparison of pseudo-R^2 measures that have been recently proposed for multilevel models and found that some measures are more appropriate

than others for comparing predictors at the different levels of analysis. A relevant finding was that a simple measure of model fit based on the deviance might be adequate for determining relative importance in longitudinal multilevel models of varied complexities and among different types of predictors, which might facilitate the use of the procedure in applied settings. Finally, this study demonstrated that dominance analysis is robust to minor misspecifications of the covariance structure, a situation that might occur in practice, especially when the number of subjects and or measurement occasions is not large.

Limitations and Future Directions

As with any simulation study, the main limitation of this research is that results are only generalizable to the factors manipulated here. Therefore, researchers wishing to apply these methods must determine how closely the conditions they want to investigate match the conditions evaluated in this study. Indeed, this study only analyzed a small variety of longitudinal multilevel models, focusing on continuous predictors. Additionally, it must be emphasized that dominance analysis is intended for comparing predictors once a model has been selected; it is not intended to inform decisions about model selection and its results are dependent on having a valid set of predictors. Another important caveat is that the measures of model fit included in this study are just a small sample of the different measures available, and the use of different measures might produce different dominance relationship results. The choice of measure must be determined by researchers based on their research question.

Another limitation of this study relates to the fact that the inferential procedures evaluated here are based on the bootstrap method, and, as such, depend on the parent sample being representative of the population of interest in order for results to be meaningful. Finally, it must be noted that the computing time required for performing dominance analysis on longitudinal

multilevel models can be long if the sample size is large and there are a large number of predictors. Since DA requires fitting all subset models, and bootstrapping procedures add even more computing demands, this procedure might not be ideal for models with more than 8 predictors.

Several opportunities exist for further exploration of dominance analysis for longitudinal models as proposed in this dissertation. The current study focused on evaluating DA for multilevel longitudinal models with continuous outcomes, so an obvious extension would be to investigate the procedure for similar models with categorical responses. Several of the measures of fit analyzed in this study are applicable to generalized linear mixed models and could be used in future research focusing on categorical responses. Another area of future exploration would be the evaluation of more sophisticated inferential procedures, such as bias corrected and accelerated bias corrected confidence intervals, which could improve the inferential power for dominance measures. Additionally, while this study incorporated a random predictor in the form of the linear effect of time, this predictor was not evaluated in terms of relative importance. Future research could look into employing dominance analysis for determining the relative importance of predictors with random slopes.

This study focused on the general dominance measure because it is the most straightforward to calculate and provides a quantitative value for inferential analyses, but the results may not fully apply to conditional and complete dominance. It is reasonable to expect that conditional and complete dominance might require higher levels of sample and effect sizes. Future studies could evaluate the conditions under which these two stronger levels of dominance can be achieved for multilevel longitudinal models.

The sample size at level-2, frequently the number of subjects in a longitudinal setting, was found to be one of the most important factors influencing the results of dominance analysis for

multilevel longitudinal models. Since this study looked at only a limited range of sample sizes at the subject level, future research should investigate more values of level-2 sample size to provide more precise recommendations regarding the minimum sample size needed for dominance analysis in these models. This study also employed the cases bootstrap method as it was simple to implement and was deemed to produce appropriate results, but an investigation of the other two commonly used bootstrap methods for multilevel data, namely parametric bootstrap and the residual bootstrap (Carpenter et al., 2003), might be of interest as well, especially if assumptions about the normality of residuals is violated. Finally, this study assumed complete data since multilevel models are known to be able to handle missing at random data. However, in many longitudinal study designs missingness cannot be assumed to occur at random. Therefore, the impact of missing not at random data on DA results and an evaluation of methods to handle it would be a worthy topic for future investigation.

REFERENCES

Allison, P. (2013, February 13). What's the Best R-Squared for Logistic Regression? [Web log post]. Retrieved November 6, 2017, from https://statisticalhorizons.com/r2logistic

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov, & F. Csaki (Eds.), *Proceedings of the 2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Azen, R. (2013). Using Dominance Analysis to estimate predictor importance in multiple regression. In Y. Petscher, C. Schatschneider, & D. L. Compton (Eds.), *Applied Quantitative Analysis in Education and the Social Sciences*, pp. 34-64. New York, NY: Taylor & Francis/Routledge.

Azen, R., & Budescu, D.V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, 8, 129–48.

Azen, R., & Budescu, D. V. (2006). Comparing predictors in multivariate regression models: An extension of dominance analysis. *Journal of Educational and Behavioral Statistics, 31*(2), 157-180.

Azen, R., & Cançado, L. (2017, July). *Using dominance analysis to determine predictor importance in multilevel models.* Poster presented at the International Meeting of the Psychometric Society, Zurich, Switzerland.

Azen, R., & Sass, D. A. (2008). Comparing the squared multiple correlation coefficients of non-nested models: An examination of confidence intervals and hypothesis testing. *British Journal of Mathematical and Statistical Psychology, 61*(1), 163-178.

Azen, R., & Traxel, N. M. (2009). Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics, 34*, 319-347.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Bring, J. (1996). A geometric approach to compare variables in a regression model. *The American Statistician*, *50*(1), 57-62.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin, 114*, 542-551.

Budescu, D. V., & Azen, R. (2004). Beyond global measures of relative importance: Some insights from Dominance Analysis. *Organizational Research Methods, 7*, 341-350.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. 2d ed. New York: Springer-Verlag.

Carpenter, J., Goldstein, H., & Rasbash, J. (1999). A non-parametric bootstrap for multilevel models. *Multilevel modelling newsletter*, *11*, 2-5.

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *52*(4), 431-443.

Chevan, A., & Sutherland, M. (1991). Hierarchical partitioning. *The American Statistician, 45*, 90-96.

Christensen, R. (1992). Comment on "Hierarchical Partitioning," by A. Chevan and M. Sutherland, *The American Statistician, 46*, 74.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, N.J: L. Erlbaum Associates.

Collins, L., Schafer, J., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods, 6*: 330-351.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (Vol. 32). CRC Press.

De Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, *11*(1), 57-85.

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in medicine*, *27*(29), 6137-6157.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1–26.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Ehrenberg, A. S. C. (1990). The unimportance of relative importance. *American Statistician*, *44*(3), 260-260.

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37*(3), 379-403.

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, *48*(2), 241-251.

Goldstein, H. (2010). *Multilevel Statistical Models* (4th ed., Wiley Series in Probability and Statistics). Chicester: Wiley.

Goldstein, H. (2011) Bootstrapping in Multilevel Models. In Joop J. Hox and J. Kyle Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 163-172). New York: Routledge.

Green, P. E., Carroll, J. D., & Desarbo, W. S. (1978). New Measure of Predictor Variable Importance in Multiple-Regression. *Journal of Marketing Research, 15*, 356-360.

Gromping, U. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician, 61*(2), 139-147.

Gromping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician, 63*(4), 308-319.

Gromping, U. (2015), Variable importance in regression models. *WIREs Comput Stat, 7(2):* 137-152.

Gurka, M., Kelley, G., & Edwards, L. (2012). Fixed and random effects models. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 181-190.

Hamaker, E.L., van Hattum, P., Kuiper, R.M. & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. *Handbook of Advanced Multilevel Analysis* (eds J. Hox & J.K. Roberts), pp. 231–255. Routledge, New York.

Hedeker, D., & Gibbons, R. (2006). *Longitudinal data analysis*. Hoboken, N.J.: Wiley-Interscience.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological bulletin*, *57*(2), 116.

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.

Hox, J. J. (2010). *Quantitative methodology series. Multilevel analysis: Techniques and applications, 2nd ed.* New York: Routledge/Taylor & Francis Group.

Hox, J. & van de Schoot, R. (2013). Robust methods for multilevel analysis. In Scott, M. A., Simonoff, J. S. & Marx, B. D. *The SAGE handbook of multilevel modeling* (pp. 387-402). London: SAGE Publications Ltd.

Huo, Y., & Budescu, D. V. (2009). An extension of dominance analysis to canonical correlation analysis. *Multivariate Behavioral Research, 44*(5), 688-709.

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test (Madrid, Spain), 18*(1), 1–43.

Jaeger, B. C. (2017). *Extending R2 to the Generalized Linear Mixed Model for Longitudinal Data*, (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 10265008).

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*(6), 1086-1105.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research, 35*(1), 1-19.

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R^2 glmm to random slopes models. *Methods in Ecology and Evolution, 5*(9), 944-946.

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods, 7*(3), 238-257.

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17*(4), 433-451.

Kendall, M. (1938). A new measure of rank correlation. Biometrika. *30*(1-2): 81–93.

Kendall, M (1945). The treatment of ties in ranking problems. *Biometrika, 33*(3): 239-251.

Kruskal, W. (1984). Concepts of relative importance. *Qüestiió, 8*(1), 39-45.

Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician, 41*(1), 6-10.

Kruskal, W., & Majors, R. (1989). Concepts of relative importance in recent scientific literature. *The American Statistician, 43*(1), 2-6.

Kvalseth, T. O. (1985). Cautionary note about R^2. *The American Statistician, 39*(4), 279-285.

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

Kwok, O. M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing Longitudinal Data with Multilevel Models: An Example with Individuals Living with Lower Extremity Intra-articular Fractures. *Rehabilitation Psychology, 53*(3), 370–386.

Laird, N. (1988). Missing data in longitudinal studies. *Statistics in Medicine, 7*, 305-315.

Laird, N., & Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963-974.

LeBreton, J. M., Ployhart, R. E., & Ladd, R. T. (2004). A Monte Carlo comparison of relative importance methodologies. *Organizational Research Methods, 7*, 258-282.

Lindeman, R. H., Merenda, P.F., Gold, R.Z. (1980). *Introduction to Bivariate and Multivariate Analysis.* Glenview, IL.

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry, 17*, 319-330.

Little, R. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*, 1112-1121.

Littell, R., Milliken, George A., Stroup, Walter W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute, Inc.

Luchman, J. N. (2014). Relative importance analysis with multicategory dependent variables: an extension and review of best practices. *Organizational Research Methods, 17*(4), 452-471.

Liu, Y., Zumbo, B. D., & Wu, A. D. (2014). Relative importance of predictors in multilevel modeling. *Journal of Modern Applied Statistical Methods, 13*(1), 2.

Luo, W., & Azen, R. (2013). Determining predictor importance in Hierarchical Linear Models using Dominance Analysis. *Journal of Educational and Behavioral Statistics, 38,* 3-31.

Matuszewski, J. (2011). *Properties of an R Square Statistic for Fixed Effects in the Linear Mixed Model for Longitudinal Data,* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3495513).

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105-142). New York, NY: Academic Press.

Modugno, L., & Giannerini, S. (2015). The Wild Bootstrap for Multilevel Models. *Communications in Statistics - Theory and Methods, 44*(22), 4812-4825.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference* (Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-095). Newbury Park, CA: Sage.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692.

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews, 85*(4), 935-956.

Nakagawa, S., & Schielzeth, H. (2013) A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4, 133–142.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface, 14*(134), 20170213.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* (3rd ed.). Orlando, FL: Harcourt Brace.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international Tampere conference in statistics, 1987* (pp. 245-260). Department of Mathematical Sciences, University of Tampere.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of education*, 1-17.

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

Retzer, J. J., Soofi, E. S., & Soyer, R. (2009). Information importance of predictors: Concept, measures, Bayesian inference, and applications. *Computational Statistics & Data Analysis*, *53*(6), 2363-2377.

Roberts, J. K., & Fan, X. (2004). Bootstrapping within the multilevel/hierarchical linear modeling framework: A primer for use with SAS and SPLUS. *Multiple Linear Regression Viewpoints, 30*(1), 23-34.

Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics*, 103-116.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

SAS Institute Inc. (2017). SAS/STAT® 14.3 User's Guide. Cary, NC: SAS Institute Inc.

Schoeneberger, J. A. (2016). The impact of sample size and other factors when estimating multilevel logistic models. *The Journal of Experimental Education*, *84*(2), 373-397.

Self, S., & Liang, K. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association, 82*(398), 605-610.

Shou, Y., & Smithson, M. (2015). Evaluating predictors of dispersion: A comparison of dominance analysis and Bayesian model averaging. *Psychometrika*, *80*(1), 236-256.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of educational and behavioral statistics*, *23*(4), 323-355.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford.

Snijders, T. A., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological methods & research*, *22*(3), 342-363.

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

Soofi, E. S. (1994). Capturing the intangible concept of information. *Journal of the American Statistical Association, 89*, 1243-1254.

Soofi, E. S., Retzer, J. J., & Yasai-Ardekani, M. (2000). A framework for measuring the importance of variables with applications to management research and decision models. *Decision Sciences, 31*, 595-625.

Steele, R. (2013). Model Selection for Multilevel Models. In Scott, M., & Simonoff, Jeffrey S. Marx, Brian D. (Eds.), *The SAGE Handbook of Multilevel Modeling* (pp. 109-126). London: SAGE Publications.

Stufken, J. (1992). On hierarchical partitioning. *American Statistician, 46*(1), 70-71.

Tang, S. (2014). *Inferential Procedures for Dominance Analysis Measures in Multiple Regression*, (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3685642).

Theil, H., & Chung, C. F. (1988). Information-theoretic measures of fit for univariate and multivariate linear regressions. *The American Statistician, 42*(4), 249-252.

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research, 45*(1), 253-275.

Thomas, D. R., Zumbo, B. D., Kwan, E., & Schweitzer, L. (2014). On Johnson's (2000) relative weights method for assessing variable importance: A reanalysis. *Multivariate behavioral research, 49*(4), 329-338.

Tonidandel, S., & LeBreton, J. (2011). Relative Importance Analysis: A Useful Supplement to Regression Analysis. *Journal of Business and Psychology, 26*(1), 1-9.

UCLA: Statistical Consulting Group. (n.d.). FAQ: What are pseudo R-squareds? Retrieved from https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/ (accessed November 2, 2017).

Van den Burg, W., & Lewis, C. (1988). Some properties of two measures of multivariate association. *Psychometrika, 53*(1), 109-122.

Van der Leeden, R., Meijer, E. & Busing, F. (2008). Resampling multilevel models. In J. de Leeuw and E. Meijer (Eds.), *Handbook of Multilevel Analysis*, Chapter 11, pp. 401–433. New York: Springer.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.

Wang, J., Fisher, J. H., & Xie, H. (2011). *Multilevel Models Applications Using SAS*. Berlin: De Gruyter.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics, 14*(4), 1261-1295.

APPENDIX

Table 23 Model 1 population general dominance effect (G_i) by simulation condition.

			McFadden R^2				Beta R^2				N&S R^2				R&B 2 R^2			
nL1	**Coll**	**Effect**	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4
4	**0**	**Base**	.004	.004	.000	.000	.048	.049	.006	.005	.049	.049	.026	.026	.146	.151	.019	.018
		Small	.009	.008	.006	.003	.107	.088	.070	.036	.084	.073	.062	.042	.227	.193	.151	.077
		Large	.021	.012	.005	.001	.221	.126	.054	.015	.160	.099	.053	.027	.394	.227	.097	.027
	.5	**Base**	.006	.006	.003	.003	.069	.068	.034	.032	.060	.060	.041	.040	.173	.171	.088	.083
		Small	.015	.014	.012	.010	.145	.134	.124	.104	.115	.107	.101	.087	.229	.212	.196	.165
		Large	.026	.019	.013	.009	.218	.170	.128	.094	.173	.137	.107	.082	.312	.240	.181	.134
	.8	**Base**	.006	.006	.005	.005	.071	.070	.054	.055	.063	.062	.053	.054	.171	.164	.136	.135
		Small	.017	.016	.015	.014	.154	.148	.143	.136	.126	.122	.119	.113	.229	.220	.215	.201
		Large	.025	.021	.018	.015	.197	.175	.159	.143	.163	.146	.133	.121	.262	.232	.210	.189
8	**0**	**Base**	.003	.002	.000	.000	.064	.059	.007	.007	.053	.052	.042	.043	.155	.144	.016	.017
		Small	.006	.005	.004	.002	.129	.103	.080	.045	.068	.062	.056	.048	.231	.183	.142	.081
		Large	.013	.008	.004	.001	.259	.147	.068	.017	.107	.076	.055	.041	.396	.226	.104	.026
	.5	**Base**	.004	.004	.002	.002	.084	.081	.041	.041	.057	.057	.049	.049	.170	.165	.084	.085
		Small	.010	.009	.008	.006	.168	.153	.139	.115	.087	.082	.078	.070	.237	.217	.197	.163
		Large	.016	.012	.008	.005	.245	.186	.140	.103	.123	.101	.084	.070	.315	.239	.180	.132
	.8	**Base**	.004	.004	.003	.003	.084	.085	.067	.065	.058	.058	.054	.054	.162	.162	.128	.125
		Small	.011	.010	.009	.009	.171	.168	.159	.151	.093	.092	.089	.086	.225	.221	.210	.199
		Large	.015	.013	.011	.009	.212	.190	.171	.154	.117	.108	.100	.093	.260	.233	.209	.189

Table 24 Model 1 population general dominance difference measures (G_{ij}) by simulation condition.

			McFadden R^2						Beta R^2						N&S R^2						R&B 2 R^2					
nL1	**Coll**	**Effect**	G_{12}	G_{13}	G_{14}	G_{23}	G_{24}	G_{34}	G_{12}	G_{13}	G_{14}	G_{23}	G_{24}	G_{34}	G_{12}	G_{13}	G_{14}	G_{23}	G_{24}	G_{34}	G_{12}	G_{13}	G_{14}	G_{23}	G_{24}	G_{34}
4	**0**	**Base**	**0**	.004	.004	.004	.004	**0**	-.001	.042	.043	.043	.044	.001	**0**	.023	.023	.023	.023	**0**	-.005	.127	.128	.132	.133	.001
		Small	.001	.003	.006	.002	.005	.003	.019	.037	.071	.018	.052	.034	.011	.022	.042	.011	.031	.020	.034	.076	.150	.042	.116	.074
		Large	.009	.016	.020	.007	.011	.004	.095	.167	.206	.072	.111	.039	.061	.107	.133	.046	.072	.026	.167	.297	.367	.130	.200	.070
	.5	**Base**	**0**	.003	.003	.003	.003	**0**	.001	.035	.037	.034	.036	.002	**0**	.019	.020	.019	.020	.001	.002	.085	.090	.083	.088	.005
		Small	.001	.003	.005	.002	.004	.002	.011	.021	.041	.010	.030	.020	.008	.014	.028	.006	.020	.014	.017	.033	.064	.016	.047	.031
		Large	.007	.013	.017	.006	.010	.004	.048	.090	.124	.042	.076	.034	.036	.066	.091	.030	.055	.025	.072	.131	.178	.059	.106	.047
	.8	**Base**	**0**	.001	.001	.001	.001	**0**	.001	.017	.016	.016	.015	-.001	.001	.010	.009	.009	.008	-.001	.007	.035	.036	.028	.029	.001
		Small	.001	.002	.003	.001	.002	.001	.006	.011	.018	.005	.012	.007	.004	.007	.013	.003	.009	.006	.009	.014	.028	.005	.019	.014
		Large	.004	.007	.010	.003	.006	.003	.022	.038	.054	.016	.032	.016	.017	.030	.042	.013	.025	.012	.030	.052	.073	.022	.043	.021
8	**0**	**Base**	.001	.003	.003	.002	.002	**0**	.005	.057	.057	.052	.052	**0**	.001	.011	.010	.010	.009	-.001	.011	.139	.138	.128	.127	-.001
		Small	.001	.002	.004	.001	.003	.002	.026	.049	.084	.023	.058	.035	.006	.012	.020	.006	.014	.008	.048	.089	.150	.041	.102	.061
		Large	.005	.009	.012	.004	.007	.003	.112	.191	.242	.079	.130	.051	.031	.052	.066	.021	.035	.014	.170	.292	.370	.122	.200	.078
	.5	**Base**	**0**	.002	.002	.002	.002	**0**	.003	.043	.043	.040	.040	**0**	**0**	.008	.008	.008	.008	**0**	.005	.086	.085	.081	.080	-.001
		Small	.001	.002	.004	.001	.003	.002	.015	.029	.053	.014	.038	.024	.005	.009	.017	.004	.012	.008	.020	.040	.074	.020	.054	.034
		Large	.004	.008	.011	.004	.007	.003	.059	.105	.142	.046	.083	.037	.022	.039	.053	.017	.031	.014	.076	.135	.183	.059	.107	.048
	.8	**Base**	**0**	.001	.001	.001	.001	**0**	-.001	.017	.019	.018	.020	.002	**0**	.004	.004	.004	.004	**0**	**0**	.034	.037	.034	.037	.003
		Small	.001	.002	.002	.001	.001	**0**	.003	.012	.020	.009	.017	.008	.001	.004	.007	.003	.006	.003	.004	.015	.026	.011	.022	.011
		Large	.002	.004	.006	.002	.004	.002	.022	.041	.058	.019	.036	.017	.009	.017	.024	.008	.015	.007	.027	.051	.071	.024	.044	.020

Table 25 Model 1 population rank ordering of predictors by relative importance G_i.

			McFadden R^2				Beta R^2				N&S R^2				R&B 2 R^2			
nL1	**Coll**	**Effect**	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4	w_1	w_2	w_3	w_4
4	**0**	**Base**	1	1	2	2	2	1	3	4	1	1	2	2	2	1	3	4
		Small	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
		Large	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	.5	**Base**	1	1	2	2	1	2	3	4	1	1	2	3	1	2	3	4
		Small	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
		Large	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	.8	**Base**	1	1	2	2	1	2	4	3	1	2	4	3	1	2	3	4
		Small	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
		Large	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
8	**0**	**Base**	1	2	3	3	1	2	3	3	1	2	4	3	1	2	4	3
		Small	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
		Large	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	.5	**Base**	1	1	2	2	1	2	3	3	1	1	2	2	1	2	4	3
		Small	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
		Large	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	.8	**Base**	1	1	2	2	2	1	3	4	1	1	2	2	1	1	2	3
		Small	1	2	3	3	1	2	3	4	1	2	3	4	1	2	3	4
		Large	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4

Table 26 Model 2 population general dominance effect (G_i) by simulation condition.

			McFadden R^2								Beta R^2							
nL1	**Coll**	**Effect**	w_1	w_2	w_3	w_4	w_1T	w_2T	w_3T	w_4T	w_1	w_2	w_3	w_4	w_1T	w_2T	w_3T	w_4T
4	**0**	B-B	.010	.010	.001	.001	.008	.009	.003	.003	.097	.097	.011	.012	.028	.023	-.003	-.005
		B-L	.011	.010	.004	.002	.026	.020	.013	.008	.192	.150	.056	.022	.102	.068	.037	.015
		L-L	.020	.015	.009	.003	.027	.020	.014	.008	.294	.193	.099	.027	.093	.056	.028	.006
	.5	B-B	.010	.010	.003	.003	.011	.014	.010	.010	.124	.124	.042	.043	.046	.038	.013	.008
		B-L	.011	.009	.004	.002	.028	.027	.023	.019	.223	.161	.070	.043	.098	.079	.060	.046
		L-L	.019	.014	.008	.003	.030	.027	.024	.019	.299	.191	.107	.052	.091	.067	.048	.031
8	**0**	B-B	.002	.002	.000	.000	.006	.007	.003	.002	.037	.034	.004	.004	.020	.019	.001	-.001
		B-L	.001	.001	.000	.000	.017	.013	.009	.005	.035	.037	.011	.007	.097	.073	.047	.024
		L-L	.004	.003	.002	.001	.017	.013	.009	.005	.116	.088	.050	.016	.088	.059	.034	.012
	.5	B-B	.002	.002	.001	.001	.007	.009	.007	.007	.063	.063	.019	.019	.032	.029	.008	.005
		B-L	.001	.001	.000	.000	.018	.017	.015	.012	.061	.062	.017	.014	.120	.100	.079	.058
		L-L	.004	.003	.002	.001	.019	.018	.015	.012	.178	.122	.069	.030	.100	.075	.052	.030
			N&S R^2								**R&B 2 R^2**							
nL1	**Coll**	**Effect**	w_1	w_2	w_3	w_4	w_1T	w_2T	w_3T	w_4T	w_1	w_2	w_3	w_4	w_1T	w_2T	w_3T	w_4T
4	**0**	B-B	.062	.062	.015	.015	.092	.108	.051	.044	.023	.023	.001	.001	.069	.079	.031	.025
		B-L	.048	.035	.011	.007	.193	.162	.113	.074	-.604	-.392	-.265	-.078	.236	.183	.109	.063
		L-L	.091	.053	.024	.008	.192	.163	.119	.077	.143	.016	-.030	-.023	.140	.094	.051	.027
	.5	B-B	.048	.048	.021	.021	.094	.124	.104	.107	-.007	.001	.025	.031	.093	.113	.085	.078
		B-L	.025	.019	.009	.007	.143	.168	.170	.162	-.583	-.417	-.198	-.050	.240	.259	.215	.174
		L-L	.052	.034	.019	.010	.134	.163	.170	.165	.103	.061	.047	.052	.129	.115	.096	.079
8	**0**	B-B	.023	.022	.015	.015	.116	.138	.067	.058	.149	.136	.014	.014	.029	.022	.005	.004
		B-L	.008	.008	.006	.005	.218	.182	.126	.082	.120	.140	-.006	.007	.039	.033	.013	.010
		L-L	.018	.014	.009	.006	.222	.184	.133	.086	.371	.214	.084	.018	.086	.061	.038	.022
	.5	B-B	.017	.017	.012	.012	.108	.150	.126	.131	.135	.135	.052	.052	.053	.048	.036	.034
		B-L	.004	.004	.003	.003	.155	.183	.183	.173	.118	.133	.019	.038	.054	.054	.045	.045
		L-L	.011	.008	.006	.004	.151	.183	.187	.178	.258	.162	.091	.052	.083	.083	.080	.075

Table 27 Model 2 population rank ordering of predictors by relative importance G_i.

			McFadden R^2								Beta R^2								N&S R^2								R&B 2 R^2							
nL1	**Coll**	**Effect**	w_1	w_2	w_3	w_4	w_1T	w_2T	w_3T	w_4T	w_1	w_2	w_3	w_4	w_1T	w_2T	w_3T	w_4T	w_1	w_2	w_3	w_4	w_1T	w_2T	w_3T	w_4T	w_1	w_2	w_3	w_4	w_1T	w_2T	w_3T	w_4T
4	**0**	B-B	1	1	5	5	3	2	4	4	1	1	5	4	2	3	6	7	3	3	6	6	2	1	4	5	5	5	6	6	2	1	3	4
		B-L	4	5	7	8	1	2	3	6	1	2	5	7	3	4	6	8	5	6	7	8	1	2	3	4	8	7	6	5	1	2	3	4
		L-L	2	3	5	7	1	2	4	6	1	2	3	7	4	5	6	8	4	6	7	8	1	2	3	5	1	6	8	7	2	3	4	5
	.5	B-B	3	3	4	4	2	1	3	3	1	1	4	3	2	5	6	7	5	5	6	6	4	1	3	2	8	7	6	5	2	1	3	4
		B-L	5	6	7	8	1	2	3	4	1	2	5	8	3	4	6	7	5	6	7	8	4	2	1	3	8	7	6	5	2	1	3	4
		L-L	4	5	6	7	1	2	3	4	1	2	3	6	4	5	7	8	5	6	7	8	4	3	1	2	3	6	8	7	1	2	4	5
8	**0**	B-B	4	4	5	5	2	1	3	4	1	2	5	5	3	4	6	7	5	6	7	7	2	1	3	4	1	2	5	5	3	4	6	7
		B-L	5	5	6	6	1	2	3	4	5	4	7	8	1	2	3	6	5	5	6	7	1	2	3	4	2	1	8	7	3	4	5	6
		L-L	5	6	7	8	1	2	3	4	1	2	4	6	2	3	5	7	5	6	7	8	1	2	3	4	1	2	4	8	3	5	6	7
	.5	B-B	3	3	4	4	2	1	2	2	1	1	4	4	2	3	5	6	5	5	6	6	4	1	3	2	1	1	3	3	2	4	5	6
		B-L	5	5	6	6	1	2	3	4	5	4	7	8	1	2	3	6	4	4	5	5	3	1	1	2	2	1	6	5	3	3	4	4
		L-L	5	6	7	8	1	2	3	4	1	2	5	7	3	4	6	7	5	6	7	8	4	2	1	3	1	2	3	7	4	4	5	6

Table 28 Model 3 population general dominance effect (G_i) by simulation condition.

			McFadden R^2								Beta R^2							
nL1	**Coll**	**Effect**	w_1	w_2	w_3	w_4	x_1	x_2	x_3	x_4	w_1	w_2	w_3	w_4	x_1	x_2	x_3	x_4
4	**0**	B-B	.003	.004	.000	.000	.022	.022	.002	.002	.019	.020	.002	.002	.085	.086	.007	.007
		B-L	.003	.003	.000	.000	.091	.053	.024	.006	.013	.013	.002	.002	.313	.177	.077	.017
		L-L	.015	.009	.004	.001	.088	.051	.023	.006	.072	.041	.018	.005	.282	.154	.064	.008
	.5	B-B	.005	.005	.002	.002	.030	.030	.013	.014	.031	.030	.016	.017	.102	.101	.044	.045
		B-L	.004	.003	.002	.002	.099	.071	.048	.031	.018	.018	.011	.011	.261	.200	.148	.107
		L-L	.017	.012	.009	.006	.094	.067	.046	.030	.075	.061	.048	.037	.217	.161	.117	.080
8	**0**	B-B	.002	.002	.000	.000	.025	.025	.003	.003	.018	.018	.002	.002	.081	.081	.004	.004
		B-L	.002	.002	.000	.000	.100	.058	.027	.007	.013	.011	.001	.002	.308	.173	.075	.016
		L-L	.009	.005	.003	.001	.098	.057	.026	.007	.062	.035	.016	.004	.285	.152	.060	.003
	.5	B-B	.003	.003	.001	.001	.034	.034	.015	.015	.030	.030	.017	.017	.095	.095	.037	.038
		B-L	.002	.002	.001	.001	.108	.077	.052	.033	.018	.018	.011	.011	.262	.198	.144	.102
		L-L	.010	.007	.005	.003	.105	.075	.051	.032	.072	.059	.047	.037	.221	.160	.110	.070
			N&S R^2								R&B 1 R^2							
nL1	**Coll**	**Effect**	w_1	w_2	w_3	w_4	x_1	x_2	x_3	x_4	w_1	w_2	w_3	w_4	x_1	x_2	x_3	x_4
4	**0**	B-B	.035	.036	.013	.013	.037	.036	.014	.014	.000	.000	.000	.000	.060	.064	.007	.007
		B-L	.026	.027	.010	.010	.148	.086	.042	.017	.000	.000	.000	.000	.277	.163	.071	.016
		L-L	.119	.071	.035	.014	.124	.072	.037	.014	.000	-.002	.000	-.001	.282	.157	.067	.016
	.5	B-B	.044	.043	.026	.026	.047	.046	.026	.027	.000	-.001	.000	.001	.082	.090	.047	.042
		B-L	.029	.028	.017	.017	.160	.122	.093	.069	-.001	.000	.000	.000	.252	.197	.148	.111
		L-L	.113	.088	.068	.050	.126	.094	.071	.051	.000	-.001	-.001	.000	.249	.196	.151	.114
8	**0**	B-B	.031	.031	.021	.021	.031	.031	.021	.021	.000	.000	.000	.000	.068	.068	.008	.007
		B-L	.028	.027	.019	.019	.088	.057	.035	.022	.000	.000	.000	.000	.285	.161	.071	.018
		L-L	.078	.051	.032	.020	.080	.052	.032	.020	.000	.000	.000	.000	.288	.160	.071	.018
	.5	B-B	.035	.035	.027	.027	.036	.036	.027	.027	.000	.000	.000	.000	.094	.092	.047	.047
		B-L	.028	.028	.021	.021	.103	.082	.065	.051	.000	.000	.000	.000	.258	.199	.150	.110
		L-L	.084	.068	.054	.043	.086	.069	.055	.043	.000	.000	.000	.000	.259	.201	.149	.110

Table 29 Model 3 population rank ordering of predictors by relative importance G_i.

			McFadden R^2								Beta R^2								N&S R^2								R&B 1 R^2							
nL1	**Coll**	**Effect**	w_1	w_2	w_3	w_4	x_1	x_2	x_3	x_4	w_1	w_2	w_3	w_4	x_1	x_2	x_3	x_4	w_1	w_2	w_3	w_4	x_1	x_2	x_3	x_4	w_1	w_2	w_3	w_4	x_1	x_2	x_3	x_4
4	**0**	B-B	3	2	5	5	1	1	4	4	4	3	6	6	2	1	5	5	3	2	5	5	1	2	4	4	4	4	4	4	2	1	3	3
		B-L	5	5	6	6	1	2	3	4	5	5	6	6	1	2	3	4	5	4	7	7	1	2	3	6	5	5	5	5	1	2	3	4
		L-L	4	5	7	8	1	2	3	6	3	5	6	8	1	2	4	7	2	4	6	7	1	3	5	7	5	7	5	6	1	2	3	4
	.5	B-B	4	4	5	5	1	1	3	2	5	6	8	7	1	2	4	3	3	4	6	6	1	2	6	5	6	7	6	5	2	1	3	4
		B-L	5	6	7	7	1	2	3	4	5	5	6	6	1	2	3	4	5	6	7	7	1	2	3	4	6	5	5	5	1	2	3	4
		L-L	5	6	7	8	1	2	3	4	5	6	7	8	1	2	3	4	2	4	6	8	1	3	5	7	5	6	6	5	1	2	3	4
8	**0**	B-B	3	3	4	4	1	1	2	2	2	2	4	4	1	1	3	3	1	1	2	2	1	1	2	2	4	4	4	4	1	1	2	3
		B-L	5	5	6	6	1	2	3	4	5	6	8	7	1	2	3	4	4	5	7	7	1	2	3	6	5	5	5	5	1	2	3	4
		L-L	4	6	7	8	1	2	3	5	3	5	6	7	1	2	4	8	2	4	5	6	1	3	5	6	5	5	5	5	1	2	3	4
	.5	B-B	3	3	4	4	1	1	2	2	4	4	5	5	1	1	3	2	2	2	3	3	1	1	3	3	4	4	4	4	1	2	3	3
		B-L	5	5	6	6	1	2	3	4	5	5	6	6	1	2	3	4	5	5	6	6	1	2	3	4	5	5	5	5	1	2	3	4
		L-L	5	6	7	8	1	2	3	4	4	6	7	8	1	2	3	5	2	4	6	7	1	3	5	7	5	5	5	5	1	2	3	4

CURRICULUM VITAE

Luciana P. Cançado

EDUCATION

Ph.D. Educational Psychology — Dec 2018
Concentration: Educational Statistics & Measurement
University of Wisconsin-Milwaukee, Milwaukee, WI
Dissertation: Determining Predictor Importance in Multilevel Models for Longitudinal Data: An Extension of Dominance Analysis

M.A. International Studies — Aug 2005
Ohio University, Athens, OH
Thesis: Economic Growth: Panel Data Evidence from Latin America

B.A. Data Processing Technology — Apr 2002
Centro Universitario UNA, Belo Horizonte, Brazil

RESEARCH EXPERIENCE

Associate Data Scientist, Research Team — Jul 2018 – present
Curriculum Associates, North Billerica, MA

- Optimize data management, analyze complex datasets, provide strategy for "big data", carry out linking, efficacy and evaluation studies, and present research findings at conferences.

Research Assistant, Consulting Office for Research & Evaluation (CORE) — Jan 2014 – Jul 2018
University of Wisconsin-Milwaukee, Milwaukee, WI

- Statistical consultant. Provided research design, evaluation and data analysis services to university researchers.

Summer Intern, Educational Assessment and Accountability — Jun 2015 – Jul 2015
National Center for the Improvement of Educational Assessment, Dover, NH

- Worked on a project to create a web-based open-source educational assessment literacy module to help improve understanding of assessment concepts by a broader audience.

TEACHING EXPERIENCE

Teaching Assistant, Educational Statistics and Measurement — Sep 2015 – May 2016
University of Wisconsin-Milwaukee, Milwaukee, WI

- Teaching Assistant for EDPSY 724, a graduate-level class on data analysis techniques.
- Lead lab instruction on using SAS and SPSS to conduct the analyses taught in class, graded homework/exams, and provided student support.

PUBLICATIONS

- Cançado, L., Reisel, J. R., & Walker, C. M. (2018). Impacts of a Summer Bridge Program in Engineering on Student Retention and Graduation. *Journal of STEM Education: Innovations and Research, 19*(2), 26-31.
- Cançado, L., Reisel, J. R., & Walker, C. M. (2018). Impact of first-year mathematics study groups on the retention and graduation of engineering students. *International Journal of Mathematical Education in Science and Technology, 49*(6), 856-866.
- Nix, T., Porterfield, L., & Cançado, L. (accepted book chapter). Reshaping Perceptions through Experiences: Recruiting, Promoting and Retaining High Quality Educators for Urban Districts. In: C. R. Rinke & L. Mawhinney (Eds.), *Opportunities and Challenges in Teacher Recruitment and Retention.*

CONFERENCE PRESENTATIONS

- Rome, L., Swerdzewski, P., Hu, J., & Cançado, L. (2018, October). *Where's My Crystal Ball?: Statistical Methods for Relating Formative and Summative Assessments*. Symposium presented at the 2018 Annual Meeting of the Northeastern Educational Research Association, Trumbull, CT.
- Cançado, L. & Azen, R. (2018, July). *Dominance analysis for determining predictor importance in longitudinal multilevel models*. Paper presented at the 2018 International Meeting of the Psychometric Society (IMPS), New York, NY.
- Azen, R. & Cançado, L. (2018, July). *Measures of model fit for longitudinal multilevel models*. Poster presented at the 2018 IMPS, New York, NY.
- Cançado, L. & Azen, R. (2018, April). *Determining predictor relative importance in explanatory multilevel IRT models*. Paper presented at the 2018 Annual Meeting of the National Council on Measurement in Education (NCME), New York, NY.
- Azen, R., & Cançado, L. (201**7,** July). *Using dominance analysis to determine predictor importance in multilevel models.* Poster presented at the 2017 IMPS, Zürich, Switzerland.
- Cançado, L. & Zhang, B. (2017, April). *Using small area estimation to rank subpopulations on international large-scale assessments*. Poster presented at the 2017 NCME Annual Meeting, San Antonio, TX.
- Betebenner, D., DePascale, C., Cançado, L., Sharpe, A., & Ryan, K. (2017, April). *A Framework and Platform for the Development of Assessment Literacy*. Pre-conference training session, 2017 NCME, San Antonio, TX.
- Azen, R., & Cançado, L. (2016, July). *Criticality analysis for multilevel model selection and predictor rankings.* Poster presented at the 2016 IMPS, Asheville, NC.
- Rome, L. A., Cançado, L., Azen, R., & Zhang, B. (2015, April). *Ability estimation and DIF detection in large-scale* a*ssessments*. Poster presented at the 2015 NCME Annual Meeting, Chicago, IL.

RESEARCH IN PROGRESS

- Nix, T., Porterfield, L., & Cançado, L. (submitted). Initial Validation: The Teacher-Candidate Cultural Awareness, Relevancy and Efficacy Scale (TC-CARES).

FELLOWSHIPS AND AWARDS

- University of Wisconsin-Milwaukee — Graduate Assistantship — Jan 2014 - Jul 2018
- University of Wisconsin-Milwaukee — Graduate Travel Award — 2015 - 2018
- Ohio University — Graduate Fellowship — Sep 2002 - Dec 2003
- Centro Universitario UNA — Golden Medal "Portal de Ouro" (highest GPA of the graduating class) — Apr 2002

PROFESSIONAL ORGANIZATIONS

- NCME, National Council on Measurement in Education
- Psychometric Society
- Northeastern Educational Research Association

CERTIFICATIONS

- Certificate in Machine Learning by Stanford University on Coursera. Nov 2016.
- Graduate Certificate in Applied Data Analysis Using SAS. University of Wisconsin-Milwaukee. May 2016.

SOFTWARE/PROGRAMMING SKILLS

- Statistical and measurement software: SAS (advanced); SPSS, R, Weka, WinBUGS, Matlab/Octave, IRTPRO, SCORIGHT
- Survey development: Qualtrics, Google Forms
- Programming: SQL, PL/SQL, Unix shell scripting, Perl, JavaScript, HTML, ASP, Delphi
- Databases: Oracle, SQL Server, MS Access

OTHER PROFESSIONAL EXPERIENCE

IT Service Delivery Coordinator, UITS — Aug 2009 - Jun 2013
University of Wisconsin-Milwaukee, Milwaukee, WI

Sr. Production Support Analyst, Data Center Operations — Sep 2004 - Jul 2009
Amdocs Inc., Champaign, IL

Programmer, Information Technology Department — Apr 1997 - Jul 2002
Cedro Textil, Belo Horizonte, Brazil