ANALYZING CLUSTERED LONGITUDINAL DATA USING LATENT CURVE MODEL WITH STRUCTURED RESIDUALS (LCM-SR)

DONGHO CHOI
University of Nebraska-Lincoln, tom.choi@huskers.unl.edu

ANALYZING CLUSTERED LONGITUDINAL DATA USING LATENT

CURVE MODEL WITH STRUCTURED RESIDUALS (LCM-SR)

by

Dongho Choi

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of the Requirements

For the Degree of Master of Arts

Major: Educational Psychology

Under the Supervision of Professor James Bovaird

Lincoln, Nebraska

February, 2022

ANALYZING CLUSTERED LONGITUDINAL DATA USING LATENT CURVE MODEL WITH STRUCTURED RESIDUALS (LCM-SR)

Dongho Choi, M. A.

University of Nebraska, 2022

Advisor: James Bovaird

The LCM-SR can provide an inferential basis for understanding reciprocal relations while controlling for individual differences in the trajectories of young children's psychological development. Yet, a hierarchical structure in the data has not been often adequately addressed even though that is common in social and educational research. The purpose of this study is to explore the impact of dependency among observations on the results when using the LCM-SR, and how to appropriately analyze the clustered longitudinal data for more accurate inference. To do this, the MLCM-SR (disaggregated approach; the "two-level" model) was introduced and compared with the single level LCM-SR considering nesting effects (aggregated approach; the "complex" model), and the single level LCM-SR ignoring nesting effects (conventional approach; the "default" model). This study used both simulated data and actual data to compare the performances of the models.

The simulation study results showed that all the models showed high rates of non-convergence or improper solutions in certain conditions, especially in low sample size conditions. The total number of proper solutions was higher for the complex/default model than for the two-level model in general. Also, bad model fit, severe bias, low coverage rate, and low power were found in conditions with a large percentage of variance as well as a large residual variance at the between-group level. The severity of bias increased as the sample size decreased. The two-level model showed little or no bias in general, thus showing a decent level of power and a nominal

level of type 1 error rate. The actual data analysis results showed that even though there was a difference in the standard errors found between the models, using different modeling strategies did not lead to different conclusions.

DEDICATION

To my grandmother

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

In developmental and educational science, understanding children's psychological development – how it occurs, what the outcomes are, whether children differ in terms of process and outcome, and what makes those differences – is one of the major goals of researchers. In the absence of experimental evidence, researchers often rely on longitudinal data analysis using publicly available datasets. In the last several decades, there has been a national emphasis on gathering longitudinal data on child development and early learning experience in response to growing awareness of the importance of early childhood experiences and their relationship(s) with future social and academic success (West, 2017). Many states across the U.S. have developed their own early childhood integrated data system (ECIDS) to connect information from multiple early care and education programs and agencies within the state and provide policymakers, early childhood practitioners, and parents and caregivers with the information to improve education policy and practice (Early Childhood Data Collaborative, 2010; U.S. Department of Education, 2019). At the national level, longitudinal studies collecting nationally representative samples of children in their early years have been developed and conducted. Some examples are the National Longitudinal Survey of Children and Youth (NLSCY; Bureau of Labor Statistics, 2019), the Early Childhood Longitudinal Study-Kindergarten Cohort, 2011 (ECLS-K:2011; Tourangeau et al., 2015), and Head Start Family and Child Experiences Survey (FACES; U.S. Department of Health and Human Services, 2020).

When analyzing these longitudinal data in accordance with developmental theories, researchers must consider that because cluster sampling is often used in data collection procedures, many longitudinal data systems have a hierarchical structure with different levels of clustering or nesting (Bovaird, 2007). For instance, when studying children in a school context, a

group of students in the same classroom (i.e., thus in the same school) may be selected as a sampling unit. In that case, students (Level 1) are described as nested within classrooms (Level 2), and classrooms can be further nested within schools (Level 3). Then, repeated measures (Level 1) are also inherently nested within each individual (Level 2). These data are often considered multilevel, hierarchically structured, or clustered.

One distinctive characteristic of these types of data is that observations within the same cluster tend to show a more homogenous pattern than those from different clusters. Hence, this violates the assumption of independence of observations. When this happens, conventional statistical methods that assume independent observations should not be used since incorrect parameter estimates, standard errors, and inappropriate fit statistics may be obtained, which might result in incorrect inferences regarding the proposed hypotheses (Kish, 1965; Maas & Hox, 2005; McNeish & Stapleton, 2016; Ryu, 2014; Stapleton & Kang, 2018). In such cases, the recommended procedure to analyze data is called multilevel modeling (MLM). MLM permits the explicit modeling of the relationships between variables that might be measured at different levels and allows researchers to make simultaneous inferences at all levels of the hierarchy (Bovaird, 2007).

On the other hand, when multivariate analysis with longitudinal data should take place, as when developmental researchers investigate the relationships between two or more constructs that may be intertwined in their change over time, structural equation modeling (SEM) has been adopted as a general analytic framework. SEM can be applied to a wide range of applications where, for example, latent variables or more than one dependent variable (multiple constructs and/or multiple measures of multiple constructs) are involved. Traditional SEM can be combined with MLM when the data system reflects a clustered or hierarchical nature, resulting in

multilevel structural equation modeling (MSEM). In MSEM, a single-level model with repeated measures is extended to consider variance attributable to a higher level of nesting by dividing the total variance-covariance matrix into separate within-groups (repeated measures) and between-groups (persons) covariance matrices.

Unfortunately, applied researchers working with clustered longitudinal data often do not adequately explore the multilevel nature of the data. The purpose of this study is to investigate the consequences of ignoring the effect(s) of clustered data when analyzing longitudinally clustered data and on the merits of incorporating hierarchically clustered data through MSEM with a specific focus on the multivariate examination of the relationship between two constructs over time.

Longitudinal modeling questions

In studies to understand child development, typical model-related research questions would include but are not limited to: 1) how a developmental construct changes over time, 2) whether there is a difference in that change among individuals and, if there is, which factor(s) predict those differences, and 3) whether any relationship between the changes in different domains exists (Scott, 2017). The first question relates to intra-individual variability – when repeated observations are made on the same individual over time, there could be a difference in the level of an individual's characteristics or performance as time passes. This change could show some systematic pattern (e.g., linear or curvilinear trajectory) or non-systematic pattern (e.g., short-term fluctuations in behavior that do not represent durable or systematic change). Young children's pattern of cognitive development could follow a linear growth trajectory, non-linear (accumulative) growth, or a non-continuous growth at all (Fischer & Bullock, 1984). Intra-

individual variability cannot be examined with cross-sectional data where observations about individuals are only made once.

The second question relates to inter-individual variability – individuals can show variability in their overall levels of characteristics (the intercepts) or their rates of change over time (the slopes). These between-person differences (i.e., how this change varies between individuals) over time are often of major interest to researchers exploring what factors make or predict such differences between individuals. For example, gender, ethnicity, school, or classroom membership could be utilized as predictors of systematic differences in the growth trajectories among children.

The third question involves both intra-individual and inter-individual variability. First, the relationship between developmental processes can be described inter-individually or at a between-person level – by comparing individual differences in developmental processes in terms of overall levels of developmental processes across time and their rates of change over time. For example, on average, children experiencing higher levels of emotional support from a parent or teacher tend to show higher levels of social skills (Roy & Giraldo-García, 2018). Similarly, children reporting systematic increases in parental involvement in education over time are also more likely to report systematic increases in academic achievement at school (Wilder, 2014). In addition, the relationship between developmental processes can be described intra-individually or at a within-person level – by relating time-specific change in a developmental process to the one in another process. For instance, if a child experiences higher levels of negative parenting relative to a previous time point, s/he is more likely to exhibit higher levels of externalizing behaviors at later time points than s/he did before (McKee et al., 2008). The time-specific relations among developmental processes are distinctly different from the between-person level

relationship among processes in that the latter cannot serve as a basis of causal inference (Zyphur, Allison, et al., 2020).

Modeling approaches to address the questions

To answer the first and second questions, latent curve modeling (LCM) is widely employed. The LCM is a flexible and powerful analytic process that can model any systematic change in a variable over time, often called a trajectory, where the variable could be linearly increasing across time, linearly decreasing, or changing in some other way (e.g., quadratic, exponential, etc.). The LCM draws on the strengths of structural equation modeling (SEM), where repeated measures of the same variable are incorporated as multiple indicators of one or more latent factors which jointly describe the shape of person-specific growth trajectories. The mean values of growth factors describe an average linear (or non-linear) form of change over time, and the variances of growth factors then capture individual differences in the trajectories. The growth factors could then be regressed on the covariates which may be attributable to such differential growth between individuals.

Yet, the LCM is not without its limitations. The LCM gives limited information about the third question since the covariance between latent factors can provide information about the between-person relationship only, which omits information on the within-person effect such as the directionality of the effect. For example, the LCM results can say that children whose parents show negative parenting behaviors tend to exhibit more antisocial behaviors, but it does not say whether negative parenting behaviors of parents affect antisocial behaviors or vice versa, or whether there is a reciprocal relationship between them as Patterson's Coercion Model suggested (Patterson & Yoerger, 2002). Also, there is no temporal order among the latent variables, which makes causal inference not feasible (Zyphur et al., 2020).

To answer the other part of the third question, the cross-lagged panel model (CLPM) is widely used. The CLPM originated from autoregressive (AR) models in time-series analysis and was later incorporated into the SEM framework. The key feature of AR models that distinguishes them from simple univariate regression models is that the prior values of a variable are used as predictors of current or future values of the same variable. For example, a researcher can examine how children's academic achievement scores at time $t - 1$ carry over in the prediction of the score at time t.

The CLPM is an extension of autoregressive models where two or more variables are involved. In addition to autoregressive parameters, repeated measures of a variable at previous time points are used in the CLPM to predict the measures on the other variable(s) at later time points through cross-lagged parameters.

The CLPM allows researchers to model temporal dependence among developmental processes and their directionality - if one's influence is dominating another, or if the relationship is reciprocal (Usami et al., 2019; Zyphur et al., 2020). For example, Ross and Broh (2000) found that earlier academic achievement is associated with locus of control and academic achievement at later time points, they also found that the previous locus of control predicts academic achievement at later time points. However, the CLPM does not assume trait-like individual differences in the relationship among the variables. Here, trait-like differences refer to overall differences in developmental trajectories between individuals that persist over time.

As a result, in the presence of some extent of a trait-like between-person difference in the variables, which are likely due to individual differences in the underlying developmental trajectory, the CLPM may fail to adequately account for it. Then, cross-lagged parameters may not represent the actual within-person relationship over time which in turn leads to false

interpretations of the results (Hamaker et al., 2015). Thus, the CLPM is not an ideal approach when such between-person differences in the trajectory are assumed.

Alternative model and the study purpose

There have been several extensions of the CLPM which account for both between-person differences in relationships among the variables across time and within-person level relationships over time (Mund & Nestler, 2019; Orth et al., 2020; Usami et al., 2019). In 2014, Curran and his colleagues proposed the latent curve model with structured residuals (LCM-SR; originally termed latent curve model with structural residuals) which combines both the LCM and the CLPM (Curran et al., 2014). The LCM-SR is the multivariate LCM where the CLPM is simultaneously modeled using time-specific residuals from the LCM. In the LCM-SR, the LCM part describes person-specific developmental trajectories over time and also individual differences in these trajectories, whereas the CLPM part describes whether there are autoregressive and cross-lagged effects within- and across the variables above and beyond the influence of the underlying trajectories. The LCM-SR can provide unbiased estimates and an inferential basis for reciprocal relations while estimating individual differences in developmental trajectories, which is deemed essential in the developmental context of young children. Consequently, the LCM-SR can adequately answer all of the questions given above.

The LCM-SR has been utilized in various fields including education (Clark et al., 2020; Berry & Willoughby, 2017; Willoughby et al., 2019). Still, a hierarchical structure in the data has not been often adequately addressed even though that is common in social and educational research due to complex sampling design (Berry & Willoughby, 2017; Feldon & Litson, 2021; Scott, 2017). Although it is straightforward to extend the LCM-SR models to allow for some forms of nesting effects, the identification status of such models, their related performance in

comparison with alternative models, and the interpretation of parameters have not been systematically assessed.

Thus, the main purpose of this study is to explore the impact of dependency among observations on the results when using the LCM-SR, and how to appropriately analyze the clustered longitudinal data for more accurate inference. To do this, the multilevel LCM-SR (MLCM-SR), a parametrization of the LCM-SR in an MSEM framework, was introduced and compared with the single level LCM-SR considering nesting effects (aggregated approach), and the single level LCM-SR ignoring nesting effects (conventional approach). The concepts of LCM-SR and methods needed to conduct MLCM-SR are briefly reviewed, followed by the design and analysis for the Monte Carlo simulation in Study 1 and an empirical data illustration in Study 2. Lastly, the implication of the findings with suggestions for applications is discussed.

CHAPTER 2. LITERATURE REVIEW

Since the proposed model (the MLCM-SR) is an application of structural equation modeling (SEM), it is recommended that the reader have a basic knowledge of SEM. The proposed model is also based on an extension of the LCM and the CLPM. Finally, our proposed model incorporates the principles of MSEM to account for dependency between observations in the clustered data. Therefore, information on SEM, the LCM, the CLPM, the LCM-SR, and approaches to analyze clustered data in the SEM framework are briefly reviewed before introducing the MLCM-SR.

SEM

Structural equation modeling (SEM) refers to a modeling framework that incorporates many types of statistical models as well as accommodates a variety of estimation and testing methods. These statistical models include but are not limited to, models of analysis of variance (ANOVA), multiple regression analysis, and factor analysis. In fact, all statistical procedures based on general linear modeling, whether univariate or multivariate, are special cases of structural equation models (Bovaird, 2007; Graham, 2008). The primary data for use in SEM are covariances of observed variables, not actual scores themselves as done in ordinary least squares regression models, which explains why SEM has also been referred to as covariance structure modeling.

SEM has several distinctive analytic features. One of them is that it enables a set of statistical procedures to be run and evaluated simultaneously. For example, a series of factor analyses and multiple regression analyses can be run in a single model. This allows created factors to be directly used as variables for analysis in the same model. In this regard, factors are referred to as latent variables in SEM. In addition, since a variable in SEM can serve both as an

independent and dependent variable in a single model, the terms endogenous and exogenous are introduced accordingly (Joreskog & Sorbom, 1993). Exogenous variables serve exclusively as predictors in the model, whereas endogenous variables serve as outcome variables as well as predictors of some other endogenous variables in the same model.

Another analytic feature of SEM is that researchers can build their research model and further determine which parameters are to be freely estimated, fixed, or constrained to be equal to another parameter. This process is called model specification. Then, the specified model's performance is evaluated in terms of the degree of discrepancy between the observed covariance matrix and covariance matrix estimated or reproduced by the model (i.e., model-implied covariance matrix). This process is called model evaluation and the degree of discrepancy between the observed and reproduced covariance matrices is gauged by measures of model fit. There are various model fit indices with different features and penalties (Hu & Bentler, 1999; Fan & Wang, 1998; Yuan, 2005). These fit indices serve as evidence, along with theoretical justification, for direct statistical comparison among alternative models with different complexity to select the best model. For such reasons, SEM is often used in confirmatory research that aims to determine whether a hypothetical model is valid, and test theories of causal relationships accordingly (Bollen, 1989). All these analytic features make SEM useful for testing complex relationships among variables. Readers interested in a more thorough overview are referred to SEM texts including Bollen (1989), Hoyle (2012), Kline (2015), and Raykov and Marcoulides (2000).

Full structural equation models (FSEM) consist of a system of linear equations which could be then divided into two parts: the portion related to factor analysis and the portion related to regression analysis. The first part of the equations relates to factor analysis, which describes

how factors are extracted or measured by observed variables. This part is referred to as the measurement model. The second part of the equations is related to a series of multiple regression equations, which describe the relationship between the latent variables and other covariates. This part is referred to as the structural model.

Consider a simple example of full structural equation models with three latent variables - one exogenous and two endogenous where each latent variable is measured by three observed variables (i.e., 9 item responses are gathered from N participants). This model is depicted in Figure 1.

Figure 1.. *A full structural equation model with three latent variables each measured by three observed variables*
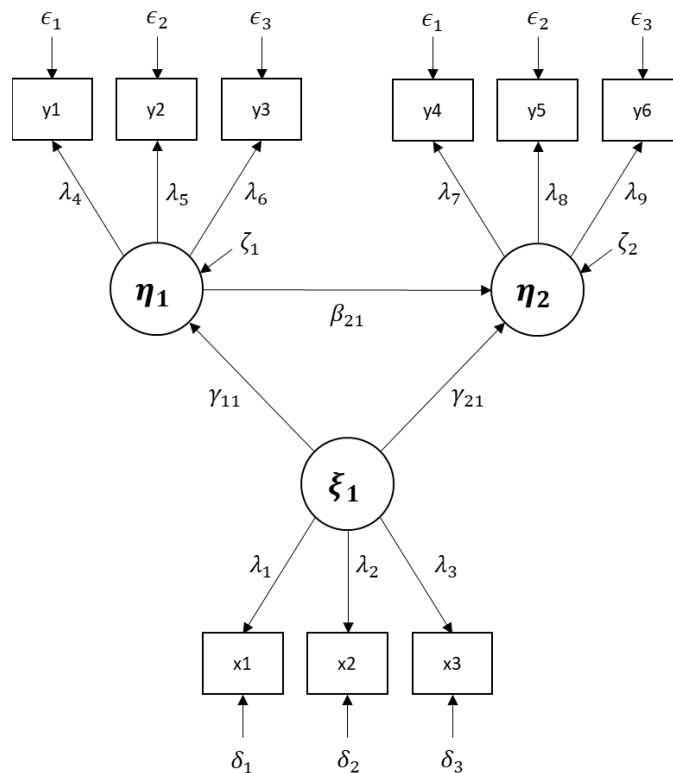
Following LISREL notation (Bollen, 1989), the measurement model can be represented as:

$$x_1 = v_{x1} + \lambda_{x1}\xi_1 + \delta_1$$

$$x_2 = v_{x2} + \lambda_{x2}\xi_1 + \delta_2 \tag{1}$$

$$x_3 = v_{x3} + \lambda_{x3}\xi_1 + \delta_3$$

$$y_1 = v_{y1} + \lambda_{y1}\eta_1 + \epsilon_1$$

$$y_2 = v_{y2} + \lambda_{y2}\eta_1 + \epsilon_2$$

$$y_3 = v_{y3} + \lambda_{y3}\eta_1 + \epsilon_3$$

$$y_4 = v_{y4} + \lambda_{y4}\eta_2 + \epsilon_4 \tag{2}$$

$$y_5 = v_{y5} + \lambda_{y5}\eta_2 + \epsilon_5$$

$$y_6 = v_{y6} + \lambda_{y6}\eta_2 + \epsilon_6$$

Here, x_i (i = 1, 2, 3) represents the three indicators of an exogenous latent variable ξ_1 (each of x_i is an N*1 vector of observations from N participant), y_j (j = 1, 2, 3, 4, 5, 6) represents the indicators of endogenous variables η_1 and η_2, λ_{xi} and λ_{yj} represent factor loadings corresponding to each latent variable, respectively, v_{xi} and v_{yj} are measurement intercepts for each item, respectively – these are often fixed to zero since mean-centered scores are often used or the mean structure is given to latent variables -, and δ_i and ϵ_j represent the residuals corresponding to each observed variable, respectively - in factor analysis δ_i and ϵ_j are called unique factors but in SEM they are referred to as measurement errors. They have an expected value of zero and variance-covariance matrix $\boldsymbol{\Theta}_\delta$ and $\boldsymbol{\Theta}_\epsilon$ (i.e., $\delta_i \sim MVN(\mathbf{0}, \boldsymbol{\Theta}_\delta)$;

$\epsilon_j \sim MVN(\mathbf{0}, \boldsymbol{\Theta}_\epsilon))$. In general, δ_i and ϵ_j are assumed to be uncorrelated with all ξs, and ηs, and

that δ_i and ϵ_j are uncorrelated with each other for all i and j (i.e., $\delta_i \sim N(0, \sigma_{xi}; \epsilon_j \sim N(0, \sigma_{yj})$. We

also assume that δ_i and ϵ_j are homoscedastic and non-autocorrelated across observations (i.e., for

different individual k and l, $\sigma_{xik} = \sigma_{xil} = \sigma_{xi}$ & $cov(\delta_{ik}, \delta_{il}) = 0$; see Bollen, 1989). This

assumption parallels that of ζs, the latent disturbances in the structural model.

Equations (1) and (2) can be re-written in matrix form as:

$$x = v_x + \Lambda_x \xi + \delta$$

$$y = v_y + \Lambda_y \eta + \epsilon$$

(3)

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \qquad v_x = \begin{bmatrix} v_{x1} \\ v_{x2} \\ v_{x3} \end{bmatrix}, \qquad \Lambda_x = \begin{bmatrix} \lambda_{x1} \\ \lambda_{x2} \\ \lambda_{x3} \end{bmatrix}, \qquad \xi = [\xi_1], \qquad \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$$

(3a)

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix}, \qquad v_y = \begin{bmatrix} v_{y1} \\ v_{y2} \\ v_{y3} \\ v_{y4} \\ v_{y5} \\ v_{y6} \end{bmatrix}, \qquad \Lambda_y = \begin{bmatrix} \lambda_{y1} & 0 \\ \lambda_{y2} & 0 \\ \lambda_{y3} & 0 \\ 0 & \lambda_{y1} \\ 0 & \lambda_{y1} \\ 0 & \lambda_{y1} \end{bmatrix}, \qquad \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}, \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

(3b)

Here, x is a p * 1 vector of indicators of the latent exogenous variable ξs (where p is the

number of indicators of ξs), is a q * 1 vector of indicators of the latent endogenous variable ηs

(where q is the number of indicators of ηs), v_x is a p * 1 matrix of measurement intercepts for y,

v_y is a q * 1 matrix of measurement intercepts for x. Λ_x and Λ_y are the factor loading matrices

containing the λ_{xi} and λ_{yj} parameters. $\mathbf{\Lambda_x}$ is a q * n matrix (where n is the length of ξ) and $\mathbf{\Lambda_y}$ is

a p * m matrix (where m is the length of $\boldsymbol{\eta}$).

Then the structural model can be represented as:

$$\eta_1 = \alpha_1 + \gamma_{11}\xi_1 + \zeta_1$$

$$\eta_2 = \alpha_2 + \beta_{21}\eta_1 + \gamma_{21}\xi_1 + \zeta_2$$

(4)

Where η_k (k=1,2) is a latent endogenous variable, ξ_l (l=1) is a latent exogenous variable

with an expected value of κ_l and variance-covariance matrix $\boldsymbol{\Phi}$ (i.e., $\xi_l \sim MVN(\mathbf{0}, \boldsymbol{\Phi})$). α_k is a

latent variable intercept. β_{kk} is a regression coefficient interrelating endogenous variables where

k refers to row and column positions and γ_{kl} is a regression coefficient relating exogenous

variables to endogenous variables where k and l refer to row and column positions. ζ_k is a

residual term or latent disturbance with an expected value of zero and variance-covariance

matrix psi (i.e., $\zeta_k \sim MVN(\mathbf{0}, \boldsymbol{\Psi})$). ζs are assumed to be homoscedastic and non-autocorrelated.

Equations (4) can be re-written in matrix form as:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

(5)

Where

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}, \qquad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix}, \qquad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix},$$

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} \\ \gamma_{21} \end{bmatrix}, \qquad \boldsymbol{\xi} = [\xi_1], \qquad \boldsymbol{\zeta} = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

(5a)

Here, $\boldsymbol{\eta}$ is m * 1 vector of the latent endogenous variables (where m is the number of endogenous latent variables). $\boldsymbol{\xi}$ is n * 1 vector of the latent exogenous variables (where n is the number of exogenous latent variables). $\boldsymbol{\alpha}$ is an m * 1 vector of latent variable intercepts. The **B** matrix is an m * m coefficient matrix for the latent endogenous variables with its elements β_{kk}. The **Γ** matrix is an m * n matrix for the latent exogenous variables with its elements γ_{kl}. $\boldsymbol{\zeta}$ is an n * 1 vector disturbances of latent endogenous variables with its element ζ_k corresponding to each η_k.

FSEM can be reduced to simpler models (Bauer, 2003). For example, if there are no latent variables involved in the model (i.e., only observed variables involved), then the measurement equation part can be omitted, and both x and y can be substituted for η and xi in the structural model. This model reduces to a path analysis model. Alternatively, if there is no specific causal structure for the latent variables assumed in the model (i.e, correlations between latent variables are assumed), the structural equation part can be omitted. This model reduces to confirmatory factor analysis (CFA). These models are depicted in Figure 2.

Figure 1. *A path analysis model (top) and a confirmatory factor analysis model (bottom)*
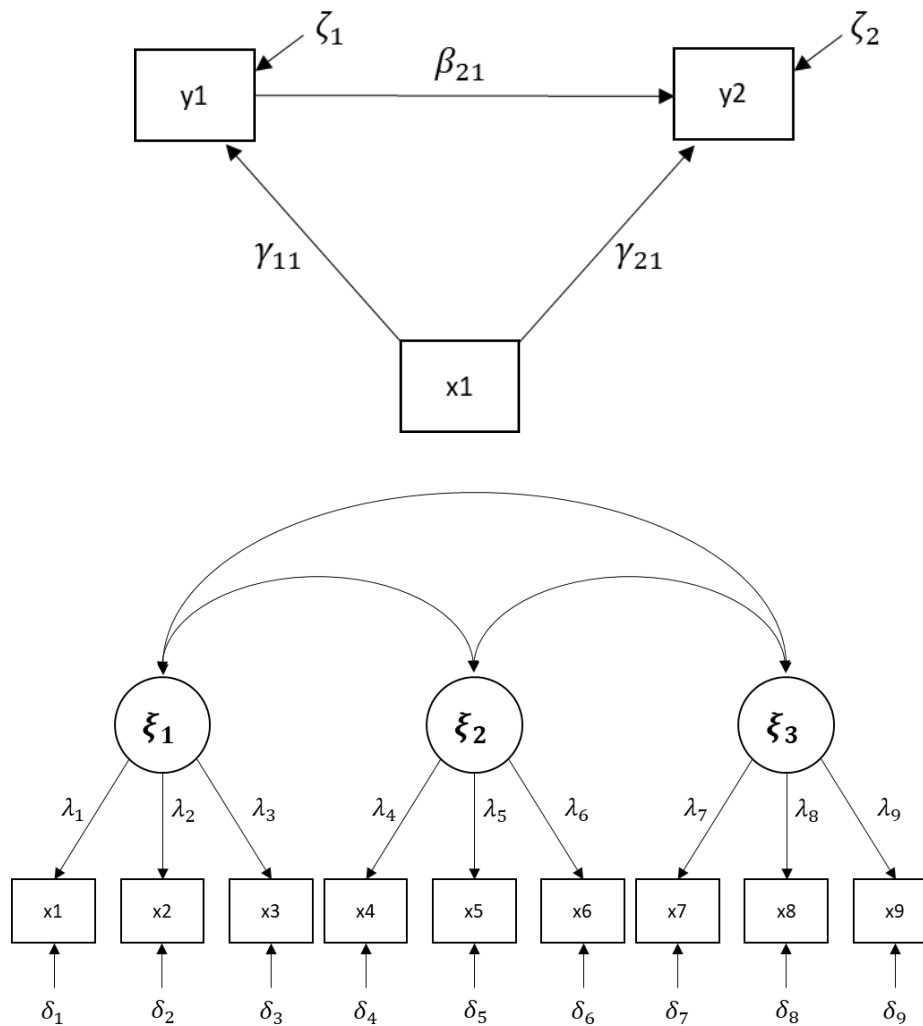
Given the models and data, optimal estimates of the parameters are derived using one of

the estimation methods in SEM - this process is called model estimation. Among the multiple

estimation method options available - maximum likelihood estimation (ML), unweighted least-

squares estimation (ULS), generalized least squares estimation (GLS), weighted least squares

estimation (WLS), asymptotically distribution-free estimation (ADF), and Bayesian estimation.

The most commonly used estimation method is the ML method (Hoyle, 2012). The ML method

tends to produce relatively unbiased parameter estimates. However, the ML method tends to

overestimate model fit statistics and underestimate the standard errors of the parameter estimates when data are not normally distributed or dependency among observations exists. In such cases, several variants of ML methods providing robust model fit statistics and standard errors are available (Asparouhov & Muthén, 2005; Asparouhov & Muthén, 2010; Satorra & Bentler, 1994). With model estimation, model-implied variance-covariance matrices of observed variables can be computed based on the obtained parameter estimates and the given formula (Bollen, 1989):

$$\widehat{\boldsymbol{\Sigma}}_{yy} = \boldsymbol{COV}(y,y) = \boldsymbol{VAR}(y) = \boldsymbol{\Lambda}_y(\boldsymbol{I} - \boldsymbol{B})^{-1}(\boldsymbol{\Gamma\Phi\Gamma'} + \boldsymbol{\Psi})(\boldsymbol{I} - \boldsymbol{B'})^{-1}\boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\epsilon$$
$$\widehat{\boldsymbol{\Sigma}}_{yx} = \boldsymbol{COV}(y,x) = \boldsymbol{\Lambda}_y(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Gamma\Phi\Lambda}_x'$$
$$\widehat{\boldsymbol{\Sigma}}_{xy} = \boldsymbol{COV}(x,y) = \boldsymbol{\Lambda}_x\boldsymbol{\Phi\Gamma'}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Lambda}_y'$$
$$\widehat{\boldsymbol{\Sigma}}_{xx} = \boldsymbol{COV}(x,x) = \boldsymbol{VAR}(x) = \boldsymbol{\Lambda}_x\boldsymbol{\Phi\Lambda}_x' + \boldsymbol{\Theta}_\delta$$

$$(6)$$

With

$$\boldsymbol{\Phi} = covariance\ matrix\ of\ \xi$$

$$\boldsymbol{\Psi} = covariance\ matrix\ of\ \zeta$$

$$\boldsymbol{\Theta}_\epsilon = covariance\ matrix\ of\ \epsilon$$

$$\boldsymbol{\Theta}_\delta = covariance\ matrix\ of\ \delta$$

Here $\widehat{\boldsymbol{\Sigma}}$ represents model-implied covariance matrix of observed variables where subscripts indicate which variable the covariance relates to. For example, $\widehat{\boldsymbol{\Sigma}}_{yy}$ indicates the model-implied covariance matrix of y and y; thus the model-implied variance-covariance matrix of y, $\widehat{\boldsymbol{\Sigma}}_{yx}$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ both indicate the model-implied covariance matrix of y and x but with different location arrangements. $\widehat{\boldsymbol{\Sigma}}_{xx}$ indicates the model-implied covariance matrix of y and y; thus the model-implied variance-covariance matrix of x.

$\boldsymbol{\Phi}$ is an n * n variance-covariance matrix for latent exogenous variables, where the main diagonals of $\boldsymbol{\Phi}$ contain the variances associated with each latent exogenous variable. The off-diagonal elements of $\boldsymbol{\Phi}$ are the covariances of disturbances for the different latent exogenous variables. $\boldsymbol{\Psi}$ is an m * m variance-covariance matrix of disturbances. The main diagonals of $\boldsymbol{\Psi}$ contain the disturbance variances associated with each latent endogenous variable. The off-diagonal elements are the covariances of disturbances for the different latent endogenous variables. The last two matrices, $\boldsymbol{\Theta}_\epsilon$ and $\boldsymbol{\Theta}_\delta$ are the variance-covariance matrices of the measurement errors. The main diagonals contain the error variances associated with each observed variable. The off-diagonal elements are the covariances of the measurement errors for the different observed variables. The $\boldsymbol{\Theta}_\epsilon$ is a p * p matrix that contains the error variances and their covariances with regards to $\boldsymbol{y}$, and $\boldsymbol{\Theta}_\delta$ is a q * q matrix and has the error variances and their covariances for $\boldsymbol{x}$.

After assembling the aforementioned sub-matrices, the total model-implied variance-covariance matrix of $\boldsymbol{y}$ and $\boldsymbol{x}$ is expressed as follows:

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{yy} & \widehat{\boldsymbol{\Sigma}}_{yx} \\ \widehat{\boldsymbol{\Sigma}}_{xy} & \widehat{\boldsymbol{\Sigma}}_{xx} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_y(\boldsymbol{I}-\boldsymbol{B})^{-1}(\boldsymbol{\Gamma\Phi\Gamma}'+\boldsymbol{\Psi})(\boldsymbol{I}-\boldsymbol{B}')^{-1}\boldsymbol{\Lambda}_y'+\boldsymbol{\Theta}_\epsilon & \boldsymbol{\Lambda}_y(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Gamma\Phi\Lambda}_x' \\ \boldsymbol{\Lambda}_x\boldsymbol{\Phi\Gamma}'(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{\Lambda}_y' & \boldsymbol{\Lambda}_x\boldsymbol{\Phi\Lambda}_x'+\boldsymbol{\Theta}_\delta \end{bmatrix} \qquad (7)$$

LCM

The latent curve modeling (LCM), also termed latent growth curve modeling (LGCM) or latent growth modeling (LGM), is a special case of SEM for modeling changes in a variable over time. More specifically, the LCM not only describes a group-level systematic trajectory of change but also captures individual differences in the trajectories over time. The current approach of the LCM was first suggested by Meredith and Tisak (1990) as a variant of

confirmatory factor analysis (CFA). Unlike typical CFA where a set of items measured at a single occasion is used to construct latent variables, the LCM uses a set of repeated measures of a single item (often a summary score) as indicators of latent variables. In the LCM, repeated observations of a variable (or variables) are assumed to be the result of a systematic underlying trajectory of change over time and a random time-specific error. The shape of the trajectory or growth curve varies by individual and can be summarized by a few parameters. These parameters are then modeled as latent variables. Researchers can specify loadings of these latent variables so that they reflect specific hypothesized trends in repeated-measures data. The univariate LCM requires three or more waves of data to identify the model.

Consider a univariate, unconditional -that is, without time-invariant or time-varying covariates- the LCM with T equally spaced repeated measurements on the variable y (i.e., T waves of repeated measures y are gathered from each of N participants). This model is depicted in Figure 3.

Figure 2. *A linear latent curve model with T repeated measures*
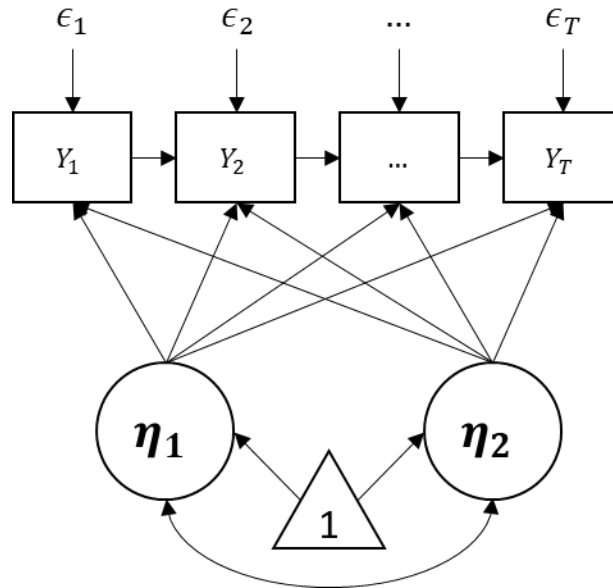
Following LISREL notation (Preacher, Wichman, Maccallum, Briggs, 2011), the measurement and structural models can be expressed as:

$$y_{ti} = \nu_t + \lambda_{1t}\eta_{1i} + \lambda_{2t}\eta_{2i} + \epsilon_{ti}$$

$$\eta_{1i} = \alpha_1 + \zeta_{1i} \tag{8}$$

$$\eta_{2i} = \alpha_2 + \zeta_{2i}$$

Where y_{ti} (t=1, 2, 3, …T; i= 1, 2, 3, …N) represents the individual- and time-specific outcome score for individual i measured at the occasion t, ν_t is measurement intercepts which are in general fixed to zero for model identification reasons ($\nu_1 = 0$; they will not be retained in further equations), λ_{jt} (j=1, 2) represents the factor loading defining the functional form of the latent trajectories over time, η_{ji} represents latent variables for a person i, ϵ_{ti} is individual- and time-specific residual which follows a normal distribution with mean 0 and variance σ_t^2 (i.e., $\epsilon_{ti} \sim N(0, \sigma_t^2)$) where the subscript t indicates that the residuals are given a unique variance at

each time point t, α_k is a latent variable mean across all individuals, and ζ_{ji} is individual

deviations around these mean values with an expected value of zero and a 2 * 2 variance-

covariance matrix $\boldsymbol{\Psi}$ (i.e., $\zeta_{ji} \sim MVN(0, \boldsymbol{\Psi})$). In general, ϵs are assumed to be uncorrelated with

all ηs, and that ϵs are uncorrelated with each other for all t and i. Yet, this assumption can be

relaxed if a correlation between time-adjacent residuals is considered (Curran et al., 2014). For

example, when assuming T=3 (i.e., 3 waves of repeated measures), residuals can follow a

multivariate normal distribution with a mean of zero and a variance-covariance matrix $\boldsymbol{\Theta}_\epsilon$ as

follows:

$$\boldsymbol{\Theta}_\epsilon = var \begin{bmatrix} \epsilon_{t1} \\ \epsilon_{t2} \\ \epsilon_{t3} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & & \\ \sigma_{12} & \sigma_2^2 & \\ 0 & \sigma_{23} & \sigma_3^2 \end{bmatrix} \tag{9}$$

Where σ_t^2 represents the time-specific variance over time and $\sigma_{(t-1)t}$ represents the time-

adjacent covariance. On the other hand, time-specific variance can be constrained to be equal

over time (i.e., $\sigma_t^2 = \sigma^2$) for simplicity of the model (Curran et al., 2014).

Latent variables in the LCM have a different meaning from that in standard SEM. Instead

of being defined as some hypothetical constructs, latent variables in the LCM are called growth

factors and serve as parameters that define the shape of individual growth curves. Assuming

linear growth in this example, η_{1i} is often called the random intercept and serves as the intercept

of the underlying trajectory for individual i – that is, the initial level of outcome scores at

baseline - and η_{2i} is called the random slope as the slope of the underlying trajectory for

individual i – that is, the rate of change in outcome scores over time.

The growth factors are then modeled as a function of mean values and individual-specific deviation from the mean whose variances reflect the variability of these factors across individuals. The means and variances are also called the fixed effects and the random effects, respectively (Curran et al., 2014).

This parameterization is possible by constraining the factor loadings to a certain set of values. For example, when assuming T=4 (i.e., 4 waves of repeated measures), factor loadings can be constrained to:

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \tag{10}$$

Where values at the first column are fixed to 1 in order to represent multipliers for the intercept – since intercepts are time-invariant values uniformly given to repeated measures across time– while the remaining columns – here only column 2 exists – are constrained to certain values related to time intervals to represent functions of time.

As for the latter, when assuming repeated measurements are equally spaced in time, a sequence of linearly increasing values such as $\lambda_{2t} = 0, 1, 2, 3$ would represent the linear function of time in the model. The corresponding factor would represent the slope of linear trajectory - for every one unit increase in time, the predicted value of y will increase by the value of the slope. In addition, the location of the zero-point in the λ_{2t} represents the baseline or the occasion at which the intercept is interpreted. Here, λ_{2t} starts with 0 to show that the baseline is at t=1. When repeated measurements are not equally spaced in time, factor loadings should be adjusted accordingly. For example, when repeated measures are observed at baseline, 1 month post-

baseline, 2 months post-baseline, and 5 months post-baseline, the proper specification for η_2 to represent the linear effect of time would be $\lambda_{2t} = 0, 1, 2, 5$.

There are multiple equivalent ways to model the linear effect of time with different zero-point locations, which serve as the origin or the intercept along with different units of time (Preacher, 2010). Furthermore, the linear LCM can be extended to incorporate non-linear growth and unequally spaced measurement occasions. Researchers should choose their options in a way in which the hypothesized trajectory of change would be most consistent with theory and research context.

Several alternative notations for the LCM are available. Following one alternative notation (Curran et al., 2014), with a few adjustments, the measurement and structural models for the linear LCM can be expressed as:

$$y_{ti} = I_{yi} + (t - 1)S_{yi} + \epsilon_{yti}$$
$$I_{yi} = \mu_{yI} + \zeta_{yIi} \tag{11}$$
$$S_{yi} = \mu_{yS} + \zeta_{ySi}$$

Where y_{ti} is defined previously, I_{yi} and S_{yi} are a combination of factor loadings, the random intercept, and slope parameters (i.e., $\lambda_{1t}\eta_{1i} = I_{yi}$; $\lambda_{2t}\eta_{2i} = (t - 1)S_{yi}$) for individual i, respectively. The (t-1) is a factor loading λ_{2t} attached to S_{yi}, which represent the value of the time trend variable at time t (λ_{1t} is omitted since it is fixed to 1), ϵ_{ti} is also defined previously, μ_{yI} and μ_{yS} are the mean intercept and slope, respectively, and ζ_{yIi} and ζ_{ySi} are individual deviations around these mean values. For modeling convenience in the later sections, we will use this notation.

Figure 4 presents a path diagram of an unconditional linear latent curve model with T waves of repeated measures (e.g., T=4). Here, the intercept alpha gives the value of y implied by the model at the first time point; beta is the linear growth component (giving the growth rate over time).

Figure 3. *A linear latent curve model (T=4)*
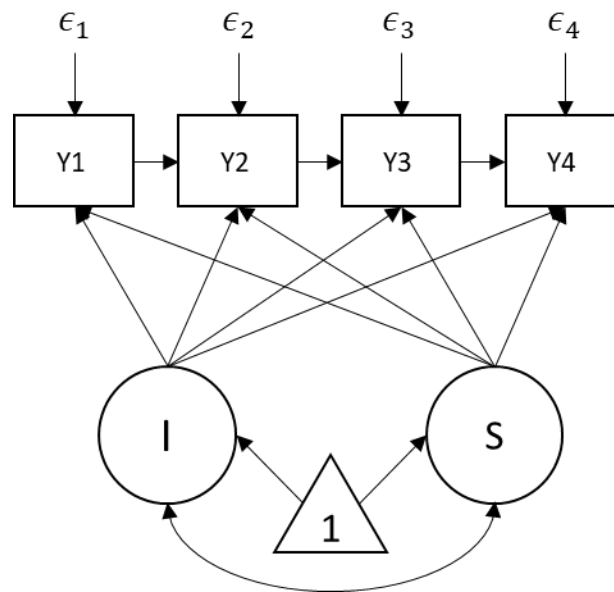
CLPM

The cross-lagged panel model is another SEM extension to analyze longitudinal data which consists of a relatively small number of repeated measurements (at least 2, typically less than 6; Usami, Murayama, & Hamaker, 2019). Yet, unlike the LCM, the focus of the CLPM is not on modeling the underlying trajectories of change in variables. Rather, the CLPM and its extensions attempt to remove such systemic patterns of change from the cross-lagged relations, whether implicitly or explicitly to prevent spurious associations (Grimm, Helm, Rodgers, & O'Rourke, 2021). The primary goal of the CLPM is to examine the predictive or causal relationships between variables. For example, when considering two variables x and y measured

at two different occasions, the CLPM compares the relationship between variable x at Time 1 and variable y at Time 2 with the relationship between variable y at Time 1 and X at Time 2. This in turn would determine if the variables x and y have reciprocal, one-directional, or no relationship at all, controlling for any other confounding variables (Granger, 1969; Hamaker, Kuiper, & Grasman, 2015). Here, the confounding variables include but are not limited to, systematic change in the variable over time, the influence of the past values, correlations between variables within the same time point - contemporaneous effects -, and autoregressive effects or influence of the variable at previous time points (Kearney, 2017). The CLPM only requires two waves of data to identify the model.

Consider the CLPM with 4 equally spaced repeated measurements on the observed variables x and y (i.e., 4 waves of repeated measures x and y are gathered from each of N participants). Here x and y do not indicate exogenous or endogenous variables, respectively. This model is depicted in Figure 5.

Figure 4. *A cross-lagged panel model (T=4).*
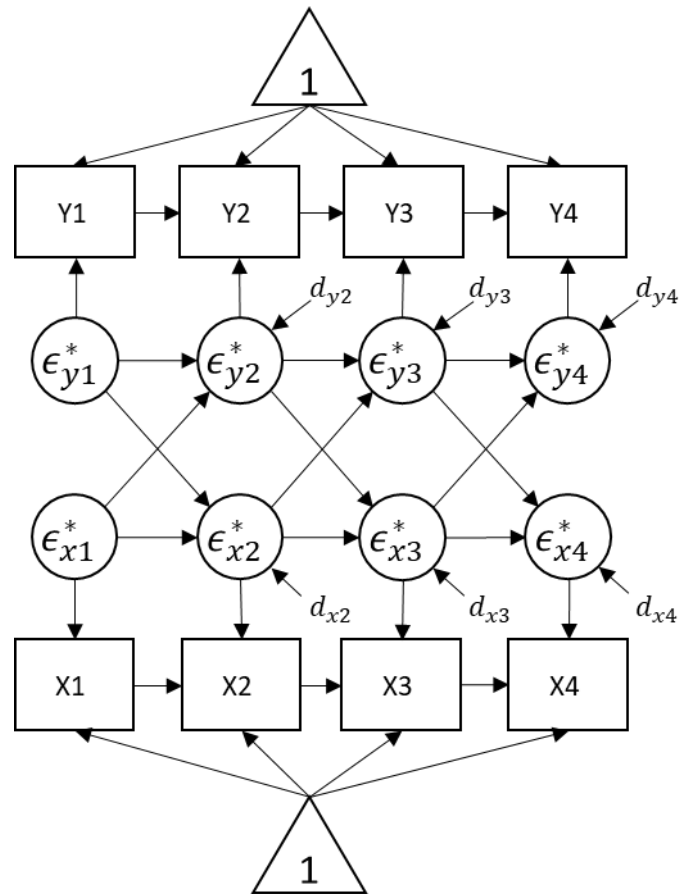
Note. Covariances between d_{xti} and d_{yti} are omitted.

Following Usami, Murayama, & Hamaker's notation (2019), with a few adjustments, the measurement model for the observed variables can be expressed as:

$$x_{ti} = \mu_{xt} + \epsilon^*_{xti}$$

$$y_{ti} = \mu_{yt} + \epsilon^*_{yti}$$

(12)

Where x_{ti} and y_{ti} (t=1, 2, 3, …T; i= 1, 2, 3, …N) are the values of the repeated measures at time t for individual i, μ_{xt} and μ_{yt} are the temporal group means of the entire sample at time t (i.e., $E(x_{ti}) = \mu_{xt}$, $E(y_{ti}) = \mu_{yt}$), and ϵ^*_{xti} and ϵ^*_{yti} are temporal deviation terms of individuals

from these group means. Note that these means are allowed to change over time. Also, deviation

terms ϵ_{xti}^* and ϵ_{yti}^* are modeled as mean-centered, single-indicator latent variables with no

measurement errors (Hamaker, Kuiper, Grasman, 2015; Hoyle, 2012). Given that, the structural

model for deviation terms can be expressed as:

$$\epsilon_{xti}^* = \beta_{xt}\epsilon_{x(t-1)i}^* + \gamma_{xt}\epsilon_{x(t-1)i}^* + d_{xti}$$
$$\epsilon_{yti}^* = \beta_{yt}\epsilon_{x(t-1)i}^* + \gamma_{yt}\epsilon_{x(t-1)i}^* + d_{yti}$$

(13)

Where ϵ_{x1i}^* and ϵ_{y1i}^* (t = 1) are latent exogenous variables with a mean of 0, variance σ_{x1}^2

and σ_{y1}^2, and covariance σ_{xy1}, ϵ_{xti}^* and ϵ_{yti}^* for t ≥ 2 are latent endogenous variables with a mean

of 0, β_{xt} and β_{yt} are autoregressive parameters, γ_{xt} and γ_{yt} are cross-lagged parameters, and d_{xti}

and d_{yti} for t ≥ 2 are latent disturbances which are termed as *innovations* or *dynamic errors* in

the literature (Usami, Murayama, Hamaker, 2019). d_{xti} and d_{yti} are typically assumed to be

normally distributed and correlated with each other only when they are measured at the same

time point (e.g., cov(d_{xti}, d_{yti}) = σ_{dxyt}, cov(d_{xti}, $d_{y(t-1)i}$) = 0). For example, when assuming

T=4 (i.e., 4 waves of repeated measures), innovations d_{xti} and d_{yti} can follow a multivariate

normal distribution with a mean of zero and a variance-covariance matrix $\boldsymbol{\Theta}_d$ as follows:

$$\boldsymbol{\Theta}_d = var\begin{bmatrix} d_{x2i} \\ d_{x3i} \\ d_{x4i} \\ d_{y2i} \\ d_{y3i} \\ d_{y4i} \end{bmatrix} = \begin{bmatrix} \sigma_{dx2}^2 & & & & & \\ & \sigma_{dx3}^2 & & & & \\ & & \sigma_{dx4}^2 & & & \\ \sigma_{dxy2} & & & \sigma_{dy2}^2 & & \\ & \sigma_{dxy3} & & & \sigma_{dy3}^2 & \\ & & \sigma_{dxy4} & & & \sigma_{dy4}^2 \end{bmatrix}$$

(14)

Where σ_{dxt}^2 and σ_{dyt}^2 are variances of time-specific innovations and σ_{dxyt} is the covariance between them.

αs or latent variable intercepts are omitted since ϵ_{xti}^* and ϵ_{yti}^* are mean-centered variables. Note that autoregressive, cross-lagged parameters, residual variances, and covariances can vary over time but these parameters are often set equal across time points (e.g., $\beta_{x2} = \beta_{x3} = \beta_{x4} = \beta_x$). This relates to the stationarity assumption where the mean, variance, and lagged covariance structure of the data are independent of time so that causal effects can be inferred under the Rubin Causal Model (Usami, Murayama, Hamaker, 2019; Grimm, Helm, Rodgers, O'Rourke, 2021; Rubin, 1974). The equal constraints can only hold or make sense when repeated measurements are equally spaced in time (Hoyle, 2012). Thus, if one wants to make an inference about the causal relationship, longitudinal data with equally distant time points would be needed.

Here, autoregressive parameters account for the influence of the past values of a variable on its future values. In the CLPM, it represents the rank-order stability of a variable - the degree to which one's relative standing amongst individuals concerning their scores on x or y is unchanging over time (Bornstein, Putnick, Esposito, 2017). Since the effects dissipate over time, these are more likely to be called temporal stability rather than trait-like stability (Hamaker, Kuiper, Grasman, 2015).

The cross-lagged parameters are the key parameters for inferring the causal relationship between the variables. The cross-lagged parameters represent a simple partial regression coefficient from the predictor (e.g., $x_{(t-1)i}$) on the outcome variable (e.g., y_{ti}), while controlling for influence from the past values (e.g., $y_{(t-1)i}$)). This can be interpreted as the extent to which

the change in y (i.e., $y_{ti} - y_{(t-1)i}$) can be predicted from an individual's prior deviation from the

group mean on x (i.e., $x_{(t-1)i} - \mu_{x(t-1)}$) while controlling for change in temporal group means

(i.e., $\mu_{yt} - \mu_{y(t-1)}$) and one's prior deviation from the group mean on y (i.e., $y_{(t-1)i} - \mu_{y(t-1)}$)

(Hamaker, Kuiper, Grasman, 2015). For a detailed discussion about how cross-lagged

parameters in the CLPM can have a causal interpretation, interested readers are referred to

Usami, Murayama, Hamaker (2019).

In the CLPM, additional lagged relations with different lag orders can be modeled

(Zyphur et al., 2021). In the current example, we only modeled the autoregressive (AR) and

cross-lagged (CL) effects or processes between time-adjacent repeated measures which are 1

timepoint apart (i.e., from t-1 to t). These are the AR process with a lag order of 1 (AR(1)) and

the CL process with a lag order of 1 (CL(1)). In other cases, multiple AR terms and CL terms

with different lags can be added to the model. For example, the AR process with a lag order of 2

(AR(2)) and CL process with a lag order of 2 (CL(2)) (i.e., involving repeated measures which

are 2 timepoints apart; from t-2 to t) can be added. The number of lags in the model is typically

determined by theory and empirical necessity. In this study, we will only model AR(1) and

CL(1) processes for simplicity reasons.

Again, there are several alternative available notations for the CLPM. Following one

alternative notation (Curran et al., 2014), with a few modifications, the models for deviation

terms can be expressed as:

$$\epsilon^*_{xti} = \rho_{xxt}\epsilon^*_{x(t-1)i} + \rho_{xyt}\epsilon^*_{y(t-1)i} + d_{xti}$$
$$\epsilon^*_{yti} = \rho_{yyt}\epsilon^*_{y(t-1)i} + \rho_{yxt}\epsilon^*_{x(t-1)i} + d_{yti}$$

$$(15)$$

Where ϵ^*_{xti} and ϵ^*_{yti} are defined previously, ρ_{xxt} and ρ_{yyt} are time-specific

autoregressive parameters, ρ_{xyt} and ρ_{yxt} are time-specific cross-lagged parameters, and d_{xti} and

d_{xti} are time-specific latent disturbances or innovations. For modeling convenience in later

chapters, this notation was utilized.

LCM-SR

Along with other assumptions for inferring causal effects, the traditional CLPM

implicitly assumes that there are no trait-like individual differences, and only accounts for

temporal stability through the inclusion of autoregressive parameters. This implies that if the

stability of the variables has a time-invariant nature, the inclusion of autoregressive parameters

will not adequately control for it, hence yielding biased estimates (Usami, Murayama, Hamaker,

2019).

Many alternative models to the traditional CLPM have been proposed, providing more

valuable insights about inferring the predictive or causal relationship between constructs (Usami,

Murayama, Hamaker, 2019; Mund & Nestler, 2019; Orth et al., 2020). All these models control

for or de-trend to some extent trait-like individual differences in the relationship between

variables. The latent curve model with structured residuals (LCM-SR) is one such approach

(Curran et al., 2014). The LCM-SR simultaneously considers individual differences in the

underlying trajectory of changes in variables - the LCM part - as well as temporal stability and

cross-lagged relationship between variables within each individual - the CLPM part - in the

model. At the same time, the LCM-SR separates the LCM part from the CLPM part. This is

possible because the LCM part decomposes the observed score into expected scores for each

individual predicted by the growth factors and time-specific residuals, while the CLPM part

models the cross-lagged relations only on the residuals which are uncorrelated with expected

scores. From the LCM tradition, residuals are not considered of substantive interest beyond

finding the optimal covariance structure for a given set of data (Grimm & Widaman, 2010;

Curran et al., 2014). On the contrary, from the CLPM tradition, the underlying trajectory of the

repeated measures would be considered a nuisance influence that should be controlled so that

only random fluctuation around the temporal group means remains (Usami, Murayama,

Hamaker, 2019). In LCM-SR, both parts have substantial meaning in that the LCGM part

provides information about the developmental trajectories of individuals, while the CLPM part

provides inferential bases for predictive or causal relationships between variables. In fact, when

assuming the linear trend among repeated measures, the LCM-SR provides more valid estimates

about cross-lagged parameters than other alternatives (Berry & Willoughby, 2017; Usami,

Murayama, Hamaker, 2019). The univariate LCM-SR requires three waves of data to identify the

model when stationarity of parameters can be assumed, four waves if not. The bivariate LCM-SR

also requires three waves of data when stationarity of parameters can be assumed, four waves if

not. The model identification is examined by looking at whether the SEM program can compute

a proper solution or not (Kenny & Milan, 2012).

Consider LCM-SR with T equally spaced repeated measurements on the observed

variables x and y (e.g., T=4). This model is depicted in Figure 6.

Figure 5. *A cross-lagged panel model (T=4).*
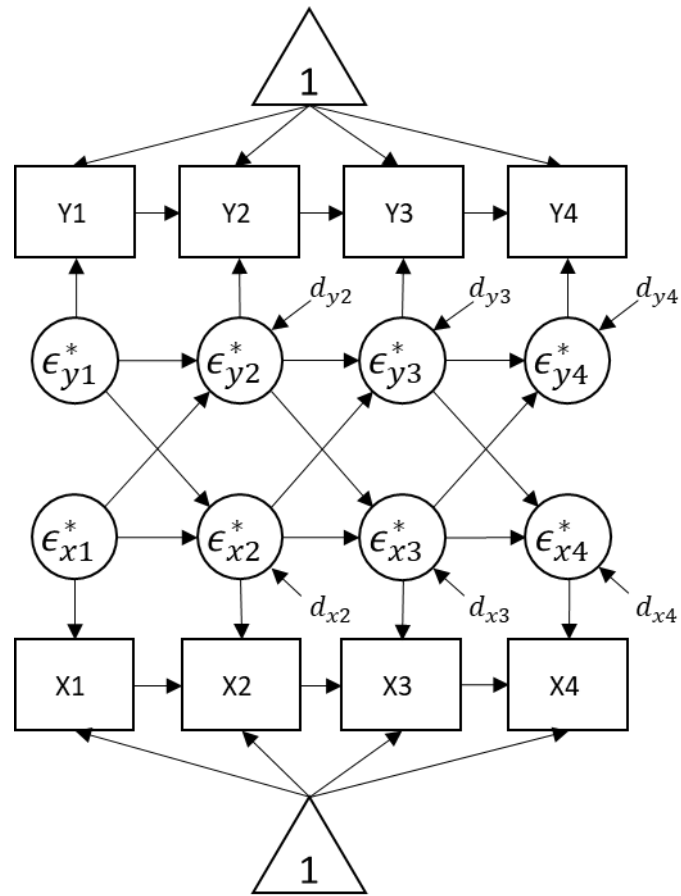
Note. Covariances between d_{xti} and d_{yti} are omitted.

Figure 6. *A linear latent curve model with structured residuals (T=4).*
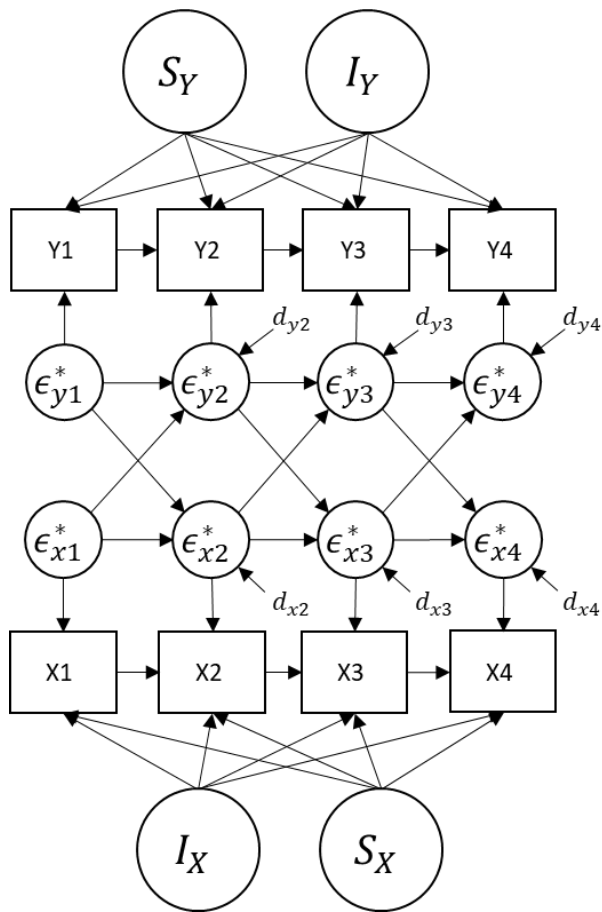
Note. Growth factor covarinace structure and covariances between d_{xti} and d_{yti} are omitted.

Following Curran et al.'s notation (2014), with a few adjustments, the measurement

model for the observed variables can be expressed as:

$$x_{ti} = I_{xi} + (t-1)S_{xi} + \epsilon_{xti}^*$$

$$y_{ti} = I_{yi} + (t-1)S_{yi} + \epsilon_{yti}^*$$

$$I_{xi} = \mu_{xI} + \zeta_{xIi}$$

$$S_{xi} = \mu_{xS} + \zeta_{xSi}$$

(16)

$$I_{yi} = \mu_{yI} + \zeta_{yIi}$$

$$S_{yi} = \mu_{yS} + \zeta_{ySi}$$

$$\epsilon^*_{xti} = \rho_{xxt}\epsilon^*_{x(t-1)i} + \rho_{xyt}\epsilon^*_{y(t-1)i} + d_{xti}$$

$$\epsilon^*_{yti} = \rho_{yyt}\epsilon^*_{y(t-1)i} + \rho_{yxt}\epsilon^*_{x(t-1)i} + d_{yti}$$

$$x_{ti} = I_{xi} + (t-1)S_{xi} + \epsilon^*_{xti}$$

$$y_{ti} = I_{yi} + (t-1)S_{yi} + \epsilon^*_{yti}$$

$$I_{xi} = \mu_{xI} + \zeta_{xIi}$$

$$S_{xi} = \mu_{xS} + \zeta_{xSi}$$

$$I_{yi} = \mu_{yI} + \zeta_{yIi}$$ (16)

$$S_{yi} = \mu_{yS} + \zeta_{ySi}$$

$$\epsilon^*_{xti} = \rho_{xxt}\epsilon^*_{x(t-1)i} + \rho_{xyt}\epsilon^*_{y(t-1)i} + d_{xti}$$

$$\epsilon^*_{yti} = \rho_{yyt}\epsilon^*_{y(t-1)i} + \rho_{yxt}\epsilon^*_{x(t-1)i} + d_{yti}$$

Where ϵ^*_{xti} and ϵ^*_{yti} are individual- and time-specific deviation terms from individual-specific expected scores (i.e., $I_{xi} + (t-1)S_{xi}$ or $I_{yi} + (t-1)S_{yi}$), rather than from the group means (i.e., μ_{xt} or μ_{yt}). All other parameters are defined previously from the LCM and the CLPM.

However, there are differences in how to interpret autoregressive and cross-lagged parameters between the traditional CLPM and LCM-SR (Usami, Murayama, Hamaker, 2019). For example, autoregressive parameters still account for the influence of past values of a variable on its future values. However, the values being used are not deviations from the overall group means for the entire sample as in the traditional CLPM. Rather, they are deviations from individual-specific expected scores which do not provide information of actual rank order

between individuals. Thus, they represent the carry-over effect from one occasion to the later ones (Hamaker, Kuiper, Grasman, 2015). More specifically, in the LCM-SR, autoregressive parameter ρ_{xxt} indicates the degree by which deviations from an individual's expected score on x (i.e., $x_{ti} - I_{xi} - (t-1)S_{xi}$) can be predicted from preceding deviations from one's expected score on x (i.e., $x_{(t-1)i} - I_{xi} - (t-2)S_{xi}$), while controlling for the individual's deviation from the preceding expected score on y (i.e., $\epsilon^{*}_{y(t-1)i}$).

Similarly, cross-lagged parameters account for the influence of past values of a variable on future values of another variable but do not provide information about one's rank-order anymore. In the LCM-SR, cross-lagged parameter ρ_{xyt} indicates the extent to which deviations from an individual's expected score on x (i.e., $x_{ti} - I_{xi} - (t-1)S_{xi}$) can be predicted from the individual's prior deviation from one's expected score on y (i.e., $y_{(t-1)i} - I_{yi} - (t-2)S_{yi}$), after controlling for the prior deviation from one's expected score on x (i.e., $\epsilon^{*}_{x(t-1)i}$). In other words, the LCM-SR provides information about predictive or causal relationships between variables above and beyond the influence of the underlying trajectory. This is favorable when compared to the traditional CLPM because the CLPM does not adequately control for individual differences in the underlying trajectories, and therefore might provide inaccurate estimates for the reciprocal effect (Hamaker, Kuiper, Grasman, 2015; Kearney, 2017).

The LCM-SR is a direct extension of the LCM (i.e., the LCM is nested within the LCM-SR; Curran et al., 2014). This allows for model comparison using likelihood ratio tests to evaluate relative improvement in model fit given increasing model complexity. In addition, the LCM-SR could be transformed into other alternative models, which include the CLPM, according to the unified framework given by Usami, Murayama, & Hamaker (2019). Those can

be compared directly using information criteria. For more recent advancements in fit assessment and model selection, see Merkle, You, & Preacher (2016) and Lai (2020). Finally, the LCM-SR can be specified as equivalent to RI-CLPM, which is another alternative to the CLPM (Grimm, Helm, Rodgers, O'Rourke, 2021).

Clustered Longitudinal Data in SEM

Many educational data, including large-scale surveys with nationally representative samples, are collected using cluster sampling to minimize costs from data collection (Muthén & Satorra, 1995; Stapleton, 2006; Hsu, Lin, & Skidmore, 2018). For a simple example, consider a sampling strategy for a survey on school-aged children where schools are randomly selected and students within the selected school were then randomly sampled. Then, collected data would be two-level data where students (Level 1) are nested within schools (Level 2). Here, schools would be called primary sampling units (PSU), and the students nested within the school are the secondary sampling units (SSU). When repeated measures are observed over time on these sampling units, then these data would be called clustered longitudinal data at three levels where repeated measures (Level 1) are nested within students (Level 2) and the students are then nested within schools (level 3).

Dependency between repeated measures within individuals can be effectively dealt with the single-level LCM where random effects in Level 2 are modeled as latent variables (Preacher, Zhang, Zyphur, 2011; Bovaird, 2007). In fact, SEM and multilevel modeling (MLM) are analytically equivalent methods when accounting for the two-level nesting effect due to repeated observations over time (Bauer, 2003; Curran, 2003). Yet, in a 3 (or more) level nesting structure where individuals are nested within higher-level clusters, there remains similarity or dependency

among individuals within the same clusters that should be controlled to prevent biased parameter estimates and standard errors (Hsu, Lin, & Skidmore, 2018).

Within the SEM framework, there are two ways of accounting for individual dependency in clustered longitudinal data - aggregated approach and disaggregated approach (Muthén & Satorra, 1995; Muthén, 1997). The aggregated approach uses the same models utilized in traditional single-level SEM analysis, but takes the dependency of clustered data into account by adjusting the standard errors of parameter estimates and Chi-square goodness of fit test to be robust against violations of data assumption including complex sampling situations. This has been achieved by using the Huber-White sandwich ML estimator in Mplus (Asparukh & Muthén, 2005; Asparouhov & Muthén, 2010; Satorra & Bentler, 1994).

On the other hand, disaggregated approach adopts new, distinctive design features such as specifying separate, level-specific models– a within-group level model and a between-group level model – with corresponding level-specific parameters and variance components (Muthén & Satorra, 1995; Muthén, 1997). This strand of the modeling approach is also referred to as multilevel SEM (Bovaird, 2007; Wu & Kwok, 2012; Hsu, Lin, & Skidmore, 2018). Although the aggregated approach is simpler regarding model specification, it allows for greater modeling flexibility in that different model structures can be specified at each level. Yet, Muthén and Satorra (1995) showed that these two approaches performed equally well when analyzing complex survey data with the same model structure at all data levels.

Consider a simple example of a two-level CFA (that is, no structural relationship between latent variables exists except correlation) with a two-level dataset, where P item responses were

gathered from each of ng participants nested within G groups (e.g., P=3). Then we have a total of N (i.e., $\sum_{g=1}^{G} n_g = N$) participants. This model is depicted in Figure 7.

Figure 7. *A two-level CFA with one latent variable in each level (P=3)*
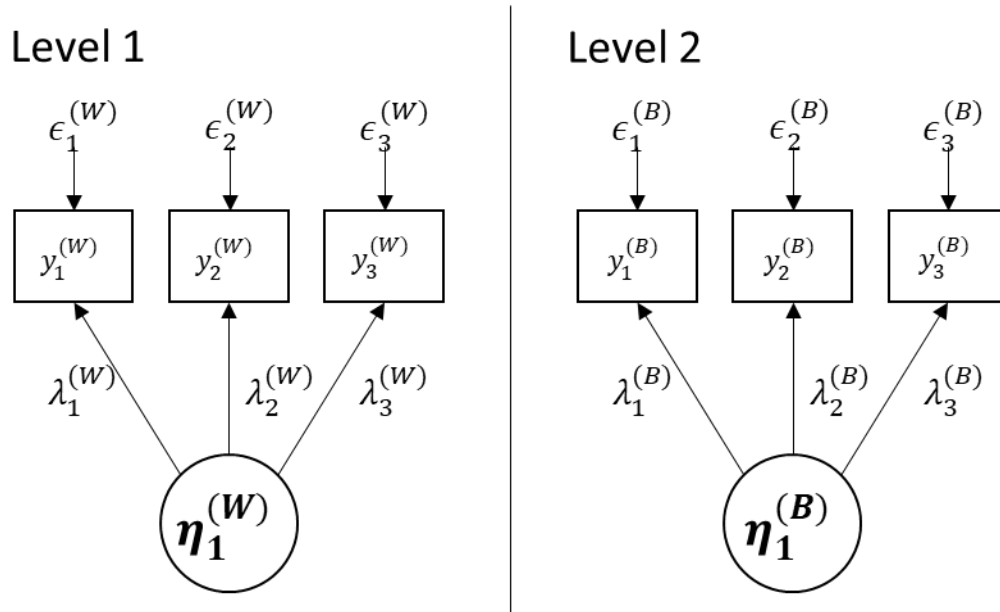
Following Wu et al.'s notation (2017), with a few adjustments, let y_{pig} be the p-th response variable (p=1, 2, 3, …, P) for the i-th participant (i.e., a within-group level unit; i= 1, 2, 3, …, n_g) within the g-th group (i.e., a between-group level unit; g= 1, 2, 3, …, G).

Here y_{pig} can be decomposed into its between-group component and within-group component, that is,

$$y_{pig} = y_{p.g}^{(B)} + y_{pig}^{(W)}$$

$$p = 1, 2, 3, …, P; i = 1, 2, 3, …, n_g; g = 1, 2, 3, …, G$$

(17)

Where the notation with superscript (B) indicates the elements in the between-group

level, and the one with superscript (W) indicates the elements in the within-group level.

Following the analysis of variance (ANOVA) tradition, the dot-notation in the subscripts

indicates over which index the mean is taken (Searle, 1971). $y_{p.g}^{(B)}$ is the between-group

component which represents the aggregated mean over group g and $y_{pig}^{(W)}$ is the within-group

component which represents an individual-specific deviation from the group mean. Here, $y_{p.g}^{(B)}$

and $y_{p.g'}^{(B)}$ (i.e., between-group components in different groups), $y_{pig}^{(W)}$ and $y_{pi'g'}^{(W)}$ (i.e., within-

group components in different groups), and $y_{p.g}^{(B)}$ and $y_{pig}^{(W)}$ (i.e., any cross-level correlation) are

set to be uncorrelated, respectively (yet, $y_{pig}^{(W)}$ and $y_{pi'g}^{(W)}$ are allowed to be correlated). Then, the

variance-covariance matrix of y_{pig} can be decomposed into between-group and within-group

components:

$$cov(y_{pig}) = cov\left(y_{p.g}^{(B)}\right) + cov\left(y_{pig}^{(W)}\right) \tag{18}$$

$y_{B,p.g}$ and $y_{W,pig}$ can be further decomposed as follows:

$$y_{p.g}^{(B)} = \mu + \Lambda^{(B)}\eta_{..g}^{(B)} + \epsilon_{p.g}^{(B)}$$
$$y_{pig}^{(W)} = \mu_g + \Lambda^{(W)}\eta_{.ig}^{(W)} + \epsilon_{pig}^{(W)} \tag{19}$$

Where μ is the grand mean over all groups, μ_g is a group-specific intercept which is

typically set to 0 for model identification (they will not be retained in further equations), $\eta_{B,..g}$ is

a vector of between-group level latent variables which follow a multivariate-normal distribution

with a mean of 0 and a variance-covariance matrix $\Psi^{(B)}$ (i.e., $\eta_{B,..g} \sim MVN(0, \Psi^{(B)})$); yet, for

some cases the grand mean is set to 0 and $\boldsymbol{\eta}_{B,.g}$ gets the group-varying mean structure), $\boldsymbol{\eta}_{W,.ig}$ is

a vector of within-group level latent variables which follow a multivariate-normal distribution

with a mean of 0 and a variance-covariance matrix $\boldsymbol{\Psi}^{(W)}$ (i.e., $\boldsymbol{\eta}_{W,.ig} \sim MVN(\mathbf{0}, \boldsymbol{\Psi}^{(W)})$), $\boldsymbol{\Lambda}^{(B)}$ and

$\boldsymbol{\Lambda}^{(W)}$ are corresponding factor loadings matrices, respectively, $\epsilon_{p.g}^{(B)}$ is a between-group level

measurement error that follows a multivariate-normal distribution with a mean of 0 and a

variance-covariance matrix $\boldsymbol{\Theta}^{(B)}$ (i.e., $\epsilon_{p.g}^{(B)} \sim MVN(\mathbf{0}, \boldsymbol{\Theta}^{(B)})$), and $\epsilon_{pig}^{(W)}$ is a within-group level

measurement error that follows a multivariate-normal distribution with a mean of 0 and a

variance-covariance matrix $\boldsymbol{\Theta}^{(W)}$ (i.e., $\epsilon_{pig}^{(W)} \sim MVN(\mathbf{0}, \boldsymbol{\Theta}^{(W)})$). Here $\boldsymbol{\eta}_{..g}^{(B)}, \boldsymbol{\eta}_{.ig}^{(W)}, \epsilon_{p.g}^{(B)}, \epsilon_{pig}^{(W)}$ are set

to be uncorrelated with each other (i.e., $\boldsymbol{\eta}_{..g}^{(B)} \perp \boldsymbol{\eta}_{.ig}^{(W)} \perp \epsilon_{p.g}^{(B)} \perp \epsilon_{pig}^{(W)}$).

The variance-covariance matrix of y_{pig} also can be further decomposed as follows:

$$cov(y_{pig}) = cov\left(y_{p.g}^{(B)}\right) + cov\left(y_{pig}^{(W)}\right)$$

$$= cov\left(\mu + \boldsymbol{\Lambda}^{(B)}\boldsymbol{\eta}_{..g}^{(B)} + \epsilon_{p.g}^{(B)}\right) + cov\left(\mu_g + \boldsymbol{\Lambda}^{(W)}\boldsymbol{\eta}_{.ig}^{(W)} + \epsilon_{pig}^{(W)}\right) \qquad (20)$$

$$= \boldsymbol{\Lambda}^{(B)}\boldsymbol{\Psi}^{(B)}\boldsymbol{\Lambda}^{(B)\prime} + \boldsymbol{\Theta}^{(B)} + \boldsymbol{\Lambda}^{(W)}\boldsymbol{\Psi}^{(W)}\boldsymbol{\Lambda}^{(W)\prime} + \boldsymbol{\Theta}^{(W)}$$

Traditionally, the degree of similarity within the same cluster is indexed by the intraclass

correlation (ICC) for each observed variable. ICC is defined as the ratio between-group level

variance and the total variance of a variable (Muthén & Satorra, 1995; Mehta & Neale, 2005).

ICC represents the expected correlation between two randomly chosen subjects within the same

group. The larger the intraclass correlation, the larger the deviation from the assumption of

independence between observations and the larger the distortion of conventional statistical

methods that assume independent observations. In the MSEM context, Muthén (1991, 1994)

provided an error-free version of ICC which is defined as the ratio between the between-level

latent factor variance and total latent factor variance. For example, for a single-factor two-level

CFA with latent variable η_1, ICC is given as:

$$ICC = \frac{\Psi^{(B)}}{\Psi^{(B)} + \Psi^{(W)}} = \frac{var\left(\eta_1^{(B)}\right)}{var\left(\eta_1^{(B)}\right) + var\left(\eta_1^{(W)}\right)} \tag{21}$$

However, when more than one latent variable is involved in the model and latent

variables are allowed to covary, the aforementioned ICC cannot be used since total variance is no

longer the sum of the variances of the different components (Anumendem, 2011). As an

alternative approach, Raudenbush and Bryk (2002), in the context of studying school effect on

academic achievement using the two-level LCM, suggested a way to look at the variance ratio

for each latent variable separately. For example, in a 2-factor two-level CFA with latent variables

η_1 and η_2, the percentage of the variance between clusters on latent variables can be given as:

$$\% \ variance \ between \ clusters \ on \ \eta_1 = \frac{var\left(\eta_1^{(B)}\right)}{var\left(\eta_1^{(B)}\right) + var\left(\eta_1^{(W)}\right)}$$

$$\% \ variance \ between \ clusters \ on \ \eta_2 = \frac{var\left(\eta_2^{(B)}\right)}{var\left(\eta_2^{(B)}\right) + var\left(\eta_2^{(W)}\right)} \tag{22}$$

In the context of the two-level LCM with academic achievement data for children nested

in schools, the percentage of the variance between clusters represent the school effects or "the

percentage of variation that lies between schools for both the initial status and learning rate"

regarding children's academic achievement (Raudenbush & Bryk, 2002). In this study, we will use the index to measure individual dependency within the same cluster in clustered longitudinal data.

Multilevel LCM-SR

Now I will introduce multilevel LCM-SR (MLCM-SR), a parameterization of LCM-SR within the MSEM framework to address individual dependency in clustered longitudinal data. Consider a simple example of a two-level LCM-SR with clustered longitudinal dataset where T waves of repeated measures x and y are observed for each of ng participants nested within G groups. Then we have a total of N (i.e., $\sum_{g=1}^{G} n_g = N$) participants. This model is depicted in Figure 8 (e.g., T=4).

Figure 8. *A two-level latent curve model with structured residuals (T=4).*
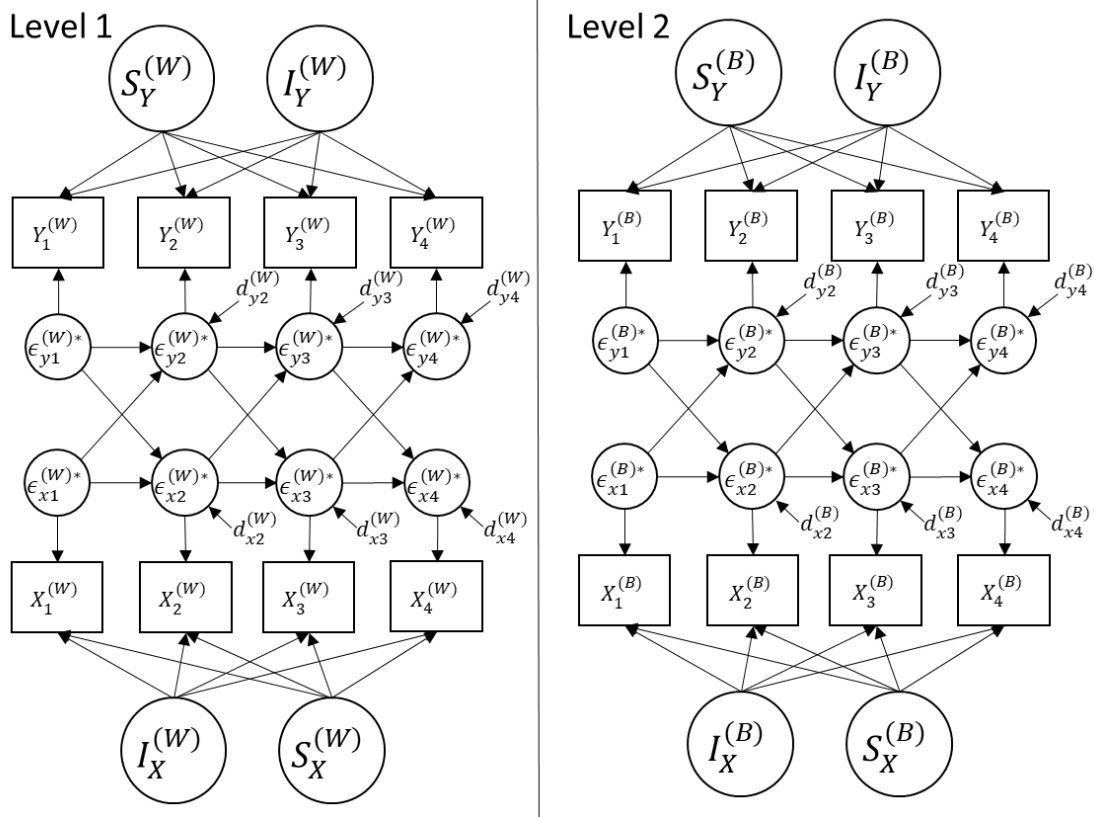
Note. Growth factor covarinace structure and covariances between $d_{xti}^{(B)}$ and $d_{yti}^{(B)}$ and covariances between between $d_{xti}^{(W)}$ and $d_{yti}^{(W)}$ are omitted.

Let y_{tig} and x_{tig} be the t-th element of T repeated measures (t = 1, 2, 3, ..., T) for the i-th participant (i.e., a within-group level unit; i= 1, 2, 3, ..., n_g) within the g-th group (i.e., a between-group level unit; g= 1,2,3,...G). Here, y_{tig} and x_{tig} can be decomposed into its between-group component and within-group component as follows:

$$y_{tig} = y_{t.g}^{(B)} + y_{tig}^{(W)}$$

$$x_{tig} = x_{t.g}^{(B)} + x_{tig}^{(W)}$$

(23)

$$t = 1, 2, 3, \ldots, T; i = 1, 2, 3, \ldots, n_g; g = 1, 2, 3, \ldots, G$$

The between-group component can be further decomposed as follows:

$$y_{t.g}^{(B)} = I_{y..g}^{(B)} + (t-1)S_{y..g}^{(B)} + \epsilon_{yt.g}^{(B)*}$$

$$x_{t.g}^{(B)} = I_{x..g}^{(B)} + (t-1)S_{x..g}^{(B)} + \epsilon_{xt.g}^{(B)*}$$

$$I_{y..g}^{(B)} = \mu_{Iy...}^{(B)} + \zeta_{Iy..g}^{(B)}$$

$$S_{y..g}^{(B)} = \mu_{Sy...}^{(B)} + \zeta_{Sy..g}^{(B)}$$

$$I_{x..g}^{(B)} = \mu_{Ix...}^{(B)} + \zeta_{Ix..g}^{(B)} \tag{24}$$

$$S_{x..g}^{(B)} = \mu_{Sx...}^{(B)} + \zeta_{Sx..g}^{(B)}$$

$$\epsilon_{yt.g}^{(B)*} = \rho_{yyt}\epsilon_{y(t-1).g}^{(B)*} + \rho_{yxt}\epsilon_{x(t-1).g}^{(B)*} + d_{xt.g}^{(B)}$$

$$\epsilon_{xt.g}^{(B)*} = \rho_{xxt}\epsilon_{x(t-1).g}^{(B)*} + \rho_{xyt}\epsilon_{y(t-1).g}^{(B)*} + d_{yt.g}^{(B)}$$

Where the notation with subscript y indicates the elements for the repeated measures y, and the one with subscript x indicates the elements for the repeated measures x. Here, the between-group component $y_{t.g}^{(B)}$ is modeled as the LCM-SR where the unit of analysis is G groups – it is expressed as a combination of the random intercept $I_{y..g}^{(B)}$, the random slope $S_{y..g}^{(B)}$, and a time-specific residual unique to group g, $\epsilon_{yt.g}^{(B)*}$. The grand mean μ is fixed to zero so that the growth factors can have the mean structure. Then, the random intercept $I_{y..g}^{(B)}$ is composed of the overall mean of the random intercept across groups $\mu_{Iy...}^{(B)}$ and a group-specific deviation from the mean $\zeta_{Sy..g}^{(B)}$. Then the random slope $S_{y..g}^{(B)}$ is composed of the overall mean of the random

slope across groups $\mu_{Sy...}^{(B)}$ and a group-specific deviation from the mean $\zeta_{Sy..g}^{(B)}$. The between-group level growth factors represent cluster-level growth – aggregated growth across groups and between-group variability in the growth. Finally, a group- and time-specific residual $\epsilon_{yt.g}^{(B)*}$ is used to model cross-lagged relations in the CLPM part. Note that the autoregressive parameter (ρ_{yyt}) and the cross-lagged parameter (ρ_{yxt}) do not have superscripts on them. This indicates that the relationship between components is the same across between-group and within-group levels. The parametrization is done this way based on the practical purpose of making two-level models that are directly comparable to single-level models when collapsing the two-level model structure. This is done so that representation of cross-lagged parameters as the causal effect would still hold (Usami, Murayama, Hamaker, 2019). The actual expression of a two-model as a single-level model is shown below. Imposing equal constraints on these parameters can be empirically tested to evaluate whether cross-level invariance exists (Muthén & Muthén, 1998-2017). The between-group component $x_{t.g}^{(B)}$ is also defined in the same way.

The within-group component can also be further decomposed as follows:

$$y_{tig}^{(W)} = I_{y.ig}^{(W)} + (t-1)S_{y.ig}^{(W)} + \epsilon_{ytig}^{(W)*}$$

$$x_{tig}^{(W)} = I_{x.ig}^{(W)} + (t-1)S_{x.ig}^{(W)} + \epsilon_{xtig}^{(W)*}$$

$$I_{y.ig}^{(W)} = \zeta_{Iy.ig}^{(W)}$$

$$S_{y.ig}^{(W)} = \zeta_{Sy.ig}^{(W)} \tag{25}$$

$$I_{x.ig}^{(W)} = \zeta_{Ix.ig}^{(W)}$$

$$S_{x.ig}^{(W)} = \zeta_{Sx.ig}^{(W)}$$

$$\epsilon_{yt.g}^{(W)*} = \rho_{yyt}\epsilon_{y(t-1)ig}^{(W)*} + \rho_{yxt}\epsilon_{x(t-1)ig}^{(W)*} + d_{ytig}^{(W)}$$

$$\epsilon_{xt.g}^{(W)*} = \rho_{xxt}\epsilon_{x(t-1)ig}^{(W)*} + \rho_{xyt}\epsilon_{y(t-1)ig}^{(W)*} + d_{xtig}^{(W)}$$

Where the within-group component $y_{tig}^{(W)}$ is modeled as the LCM-SR where the unit of analysis is N participants. The only difference in modeling between the within-group level and between-group level is that the mean of within-group level grow factors is fixed to zero for model identification reasons. The within-group level growth factors represent individual level growth – intra-individual growth and inter-individual variability in the growth above and beyond the influence of cluster-level growth. Again, the autoregressive parameter (ρ_{yyt}) and the cross-lagged parameter (ρ_{yxt}) do not have superscripts on them. The within-group component $x_{tig}^{(W)}$ is also defined in the same way.

The model can be rearranged as follows:

$$y_{tig} = \left(I_{y.ig}^{(B)} + I_{y.ig}^{(W)}\right) + (t-1)\left(S_{y..g}^{(B)} + S_{y.ig}^{(W)}\right) + \left(\epsilon_{yt.g}^{(B)*} + \epsilon_{ytig}^{(W)*}\right)$$

$$\epsilon_{yt.g}^{(B)*} + \epsilon_{ytig}^{(W)*} = \rho_{yyt}\left(\epsilon_{y(t-1).g}^{(B)*} + \epsilon_{y(t-1)ig}^{(W)*}\right) + \rho_{yxt}\left(\epsilon_{x(t-1).g}^{(B)*} + \epsilon_{x(t-1)ig}^{(W)*}\right) + \left(d_{xt.g}^{(B)} + d_{xtig}^{(W)}\right)$$

(26)

The rearrangement shows that the same model structure is specified at both the within-group and between-group levels. Other parameterizations of the model are possible, but, the difference in model structures at each level can affect estimation results (Wu, Kwok, 2012), which is not the focus of this study. Thus, in this study, we will use the same model structure across levels since the focus is on comparing the aggregated and disaggregated approaches to controlling the dependency between individuals within the same cluster. The univariate MLCM-SR requires three waves of data to identify the model when stationarity of parameters can be

assumed, four waves if not. The bivariate MLCM-SR also requires three waves of data when stationarity of parameters can be assumed, four waves if not. The model identification is examined by looking at whether the SEM program can compute a proper solution or not (Kenny & Milan, 2012).

The Current Study

The purpose of this study is to explore the impact of dependency among observations on the results when using the LCM-SR, and how to appropriately analyze the clustered longitudinal data for more accurate inference. To do this, the MLCM-SR (disaggregated approach) was introduced and compared with the single level LCM-SR considering nesting effects (aggregated approach), and the single level LCM-SR ignoring nesting effects (conventional approach). This study serves an exploratory purpose in that a new modeling approach was introduced and tested.

In this study, two different mini studies were conducted: one using the simulated data example and the other using the actual data example. In the simulated data example (Study 1), the model performance was evaluated in terms of differences in convergence rate, likelihood ratio test statistic values, practical model fit indexes, bias in parameter estimates as well as the 95% coverage rate. Finally, empirical power or Type I error rate associated with the cross-lagged parameter estimates between residuals was assessed. In the actual data example (Study 2), models were compared in terms of model fits, the statistical significance of individual parameter estimates under given alpha=0.05, and interpretation of the results.

The primary questions to be addressed in this study are as follows:

1.   Study 1: How does the performance of the MLCM-SR and the alternative modeling strategies differ across varying conditions of…

    a.   Number of clusters?

    b.   Percentage of variance between clusters?

    c.   R-square of T1 repeated measure at between-group level?

    d.   Magnitude of the one-way CL parameter?

2.   Study 2: Do different modeling strategies lead to different conclusions when dependency among observations is present in the dataset?

CHAPTER 3. STUDY 1

In Study 1, a Monte Carlo simulation was conducted to evaluate the relative performance of the selected models under various conditions pertaining to individual dependency in the multilevel longitudinal data and the nature of the reciprocated relationship. The method used in Study 1 is introduced first, followed by the results.

Method

Data generation

For the simulation study, a random sample of T= 4 equally spaced repeated measures on the observed variables x and y with balanced cluster design (i.e., Cluster size is the same for all clusters) and no missing data were generated. All data were generated based on a two-level LCM-SR with autoregressive (AR) (1) & cross-lagged (CL) (1) process. In this study, both within- and between-models were specified as having the same model structure since the difference in the model structure itself can affect estimation results (Wu & Kwok, 2012), which is not the focus of this study. The LCM-SR requires the stationarity assumption and control for any confounder for the CL effects to represent causal effect (Usami, Murayama, & Hamaker, 2019). Thus, in this model, the AR and CL effects are constrained to be equal over time, as are the residual variances and the residual covariance to achieve stationarity assumption (e.g., $\rho_{xx2} = \rho_{xx3} = \rho_{xx4} = \rho_{xx}$). Also, in this model, no additional confounders were assumed. Finally, all covariances between innovations within the same time point are constrained to be 0 for convenience of model estimation (i.e., $cov\left(d_{xtig}^{(W)}, d_{ytig}^{(W)}\right) = cov\left(d_{xtig}^{(B)}, d_{ytig}^{(B)}\right) = 0$).

The central interests of this study pertain to the impact of ignoring/considering variability in the between-group level on the model fits, parameter estimates, and corresponding

conclusions we make about the relationship between growth processes over time. Consequently, four relevant design factors were considered in this study: a number of clusters, percentage of the variance between clusters on latent variables, R-square of repeated measures at between-group level, and the magnitude of a one-way CL effect.

Number of clusters (NC)

The number of clusters or between-level sample size (or the highest-level sample size when there are more than two levels) is an important factor that determines the accuracy of the parameter estimates and their standard errors for the between model (Hox & Maas, 2001; Maas & Hox, 2005; Hox, Maas, & Brinkhuis, 2010; Hox, 2013). Maas and Hox (2005) found that, if the model is simple and the interest of the study is primarily focusing on fixed effects (i.e., regression coefficients) and their standard errors, the highest-level sample size of 20 may be sufficient for accurate estimation. However, if the interest is on random effects (i.e., variance estimates), the sample size must be much larger. Maas and Hox (2005) recommended the sample size to be of at least 50 groups even for the simplest random-effects models. At the same time, for large-scale surveys with nationally representative samples, the highest-level sample size of more than 1000 is not uncommon (Crosby & Mendez, 2016). Thus, in this study, we will use three different numbers of clusters (i.e., 50, 200, and 500) to evaluate whether different sample sizes will affect the parameter estimation of the models. On the other hand, cluster size did not affect the estimation results very much (Hox, Maas, & Brinkhuis, 2010). Thus, in this study, the cluster size is fixed to 5 for all conditions.

Percentage of variance between clusters on each latent variable (PVB)

In multilevel modeling, the ICC for each observed variable is used as one of the design factors in simulation studies because it is known to affect the accuracy of parameter estimates and their standard errors in the between model (Hox & Maas, 2001; Lai & Kwok, 2015; Lüdtke et al., 2008; Preacher, Zhang, & Zyphur, 2011) and convergence rates (yet there is controversy over the effect of ICC on convergence rate; see Maas & Hox, 2005). However, since there is more than one latent variable involved in the two-level LCM-SR, using the ICC for testing biases in this model would not be appropriate. Thus, the percentage of the variance between clusters on each latent variable (PVB) was used instead (Raudenbush & Bryk, 2002; Anumendem, 2011). Typically, in educational studies, ICC values from 10% to 30% were observed using cross-sectional multilevel models (see Raudenbush & Bryk, 2002). When fitting the growth curve model to the multilevel longitudinal data, a relatively lower percentage of the variance between clusters on the intercept or initial status and a relatively higher percentage of the variance between clusters on the slope or learning rates per academic year were observed (8% and 58.3%, respectively; see Raudenbush & Bryk, 2002). In this study, the percentage of the variance between clusters on each latent variable was set to .2 (low) and .5 (high).

R-square of T1 repeated measure at between-group level (R2B)

R-square of repeated measures in the LCM represents proportions of variance explained by the growth factors at the given time point. It is also referred to as growth curve reliability or reliability of the observed variable given the growth curve (Meredith & Tisak, 1990; Grimm & Widaman, 2010; Diallo, Morin, & Parker; 2014; Diallo & Morin, 2015; Diallo & Lu, 2017). Growth curve reliability is usually based on the first measurement occasion since this is where the baseline or zero-point is often located. R-square then represents the ratio of variance attributable to the intercept (i.e., true variance) to the total variance (i.e., variance attributable to

the intercept plus residual variance). R-square values in the LCM are a function of the time

score, the variances and covariance of the growth factors, and the variances of the time-specific

residuals. For example, in the linear LCM, R-square can be computed as follows:

$$R^2(y_{ti}) = \frac{var(I_{yi}) + (t-1)^2 var(S_{yi})}{var(I_{yi}) + (t-1)^2 var(S_{yi}) + var(\epsilon_{yti})} \qquad (27)$$

R-square of repeated measures at between-group level (R2B) is known to affect the

accuracy of parameter estimates, their standard errors, and convergence rate (Diallo, Morin, &

Parker; 2014; Diallo & Lu, 2017). In this study, two different values of R2B of T1 repeated

measures were used in order to reflect medium and large proportions of variance explained by

the growth factors: .5 (medium) and .75 (high). On the other hand, in this study, the R-square of

T1 repeated measure at the within-group level (R2W) is set to .75 for all conditions.

The magnitude of one-way CL or dominance condition (CL2)

In the CLPM and its extensions, cross-lagged (CL) parameters or CL effects determine

the relationship between two or more variables – firstly, they determine whether the variables

influence each other (i.e., significant CL effects on each other are estimated). Secondly, they also

determine which variable is causally dominant (i.e., CL effects from one variable are greater than

CL effects from the other). Third, they determine whether a variable has a positive or negative

influence on the other variable (Hamaker, Kuiper, & Grasman, 2015). Capturing the actual cross-

lagged relationship underlying the variables is a major objective of the CLPM and its extensions.

Failing to do so and thus making incorrect inferences about the causal relationship is problematic

– especially when researchers wish to use the results from the models as a basis for future

interventions. Therefore, in this study, the magnitude of the x to y cross-lagged parameter was

manipulated while constraining the y to x cross-lagged parameter to reflect dominance and non-dominance conditions. Specifically, two levels of x-to-y cross-lagged parameter ρ_{yxt} was considered along the null condition: .1 (y dominance), .3 (non-dominance), and 0 (the null condition). The x-to-y cross-lagged parameter ρ_{xyt} was constrained to .3, following Cohen's criteria on correlation effect sizes (Cohen, 1988).

All other parameters in the model are fixed for all conditions. Correlations among latent variables at the within-group and between-group levels are fixed as follows:

$$corr\begin{bmatrix} I_{y..g}^{(B)} \\ S_{y..g}^{(B)} \\ I_{x..g}^{(B)} \\ S_{x..g}^{(B)} \end{bmatrix} = corr\begin{bmatrix} I_{y..g}^{(W)} \\ S_{y..g}^{(W)} \\ I_{x..g}^{(W)} \\ S_{x..g}^{(W)} \end{bmatrix} = \begin{bmatrix} 1 & & & \\ .3 & 1 & & \\ .15 & .15 & 1 & \\ .15 & .15 & .3 & 1 \end{bmatrix}$$

Following Cohen's criteria on correlation effect sizes (Cohen, 1988), a correlation value of 0.3 was set to reflect a moderate relationship between growth factors within each construct, whereas a correlation value of 0.15 was set to reflect a weaker relationship between growth factors across constructs. In addition, following an example given in Bovaird (2007), a variance of the random intercept and variance of random slope for both constructs at the within-group level are fixed to 10 and 4, respectively, in all conditions. Finally, the variance of time-specific innovations for x and y at within-group and between-group levels are set to ½ times the variance of within-group level T1 residuals for x and y, respectively, in all conditions.

Given the above information, the population covariance structure of growth factors and residuals at within-group and between-group levels (which serves as a basis for data generation) can now be computed based on these values and the values from the design factors. For example,

when a PVB value of .5 (high) is given, the covariance structure of growth factors at within-group and between-group levels can be computed using the following equations:

$$\boldsymbol{D}^{(W)} = \begin{bmatrix} var^{\frac{1}{2}}\left(I_{y..g}^{(W)}\right) & & & \\ & var^{\frac{1}{2}}\left(S_{y..g}^{(W)}\right) & & \\ & & var^{\frac{1}{2}}\left(I_{x..g}^{(W)}\right) & \\ & & & var^{\frac{1}{2}}\left(S_{x..g}^{(W)}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{10} & & & \\ & \sqrt{4} & & \\ & & \sqrt{10} & \\ & & & \sqrt{4} \end{bmatrix} = \begin{bmatrix} 3.16 & & & \\ & 2 & & \\ & & 3.16 & \\ & & & 2 \end{bmatrix}$$

$$\boldsymbol{\Psi}^{(W)} = \boldsymbol{D}^{(W)} * corr \begin{bmatrix} I_{y..g}^{(W)} \\ S_{y..g}^{(W)} \\ I_{x..g}^{(W)} \\ S_{x..g}^{(W)} \end{bmatrix} * \boldsymbol{D}^{(W)}$$

$$= \begin{bmatrix} 3.16 & & & \\ & 2 & & \\ & & 3.16 & \\ & & & 2 \end{bmatrix} * \begin{bmatrix} 1 & & & \\ .3 & 1 & & \\ .15 & .15 & 1 & \\ .15 & .15 & .3 & 1 \end{bmatrix} * \begin{bmatrix} 3.16 & & & \\ & 2 & & \\ & & 3.16 & \\ & & & 2 \end{bmatrix}$$

$$\cong \begin{bmatrix} 10 & 1.9 & 1.5 & .95 \\ 1.9 & 4 & .95 & .6 \\ 1.5 & .95 & 10 & 1.9 \\ .95 & .6 & 1.9 & 4 \end{bmatrix}$$

$$\boldsymbol{D}^{(B)} = \boldsymbol{D}^{(W)} * \sqrt{PVB} = \begin{bmatrix} 3.16 & & & \\ & 2 & & \\ & & 3.16 & \\ & & & 2 \end{bmatrix} * 1$$

$$= \begin{bmatrix} 3.16 & & & \\ & 2 & & \\ & & 3.16 & \\ & & & 2 \end{bmatrix}$$

$$\boldsymbol{\Psi}^{(B)} = \boldsymbol{D}^{(B)} * corr \begin{bmatrix} I_{y..g}^{(B)} \\ S_{y..g}^{(B)} \\ I_{x..g}^{(B)} \\ S_{x..g}^{(B)} \end{bmatrix} * \boldsymbol{D}^{(B)}$$

$$= \begin{bmatrix} 3.16 & & & \\ & 2 & & \\ & & 3.16 & \\ & & & 2 \end{bmatrix} * \begin{bmatrix} 1 & & & \\ .3 & 1 & & \\ .15 & .15 & 1 & \\ .15 & .15 & .3 & 1 \end{bmatrix} * \begin{bmatrix} 3.16 & & & \\ & 2 & & \\ & & 3.16 & \\ & & & 2 \end{bmatrix}$$

$$\cong \begin{bmatrix} 10 & 1.9 & 1.5 & .95 \\ 1.9 & 4 & .95 & .6 \\ 1.5 & .95 & 10 & 1.9 \\ .95 & .6 & 1.9 & 4 \end{bmatrix}$$

Where $\boldsymbol{D}^{(W)}$ and $\boldsymbol{D}^{(B)}$ are 4*4 diagonal matrices with standard deviations or the square root of latent variable variance at within-group and between-group levels, respectively, in the diagonal and zeros in the other cells, and $\boldsymbol{\Psi}^{(W)}$ and $\boldsymbol{\Psi}^{(B)}$ are population variance-covariance matrices of growth factors at within-group and between-group levels, respectively.

Also, when ρ_{yxt} value of .3 (non-dominance), R2B value of .75 (high) are given, the variance of T1 residual for y at within-group and between-group levels can be computed using the following equations:

$$R^2\left(y_{1ig}^{(W)}\right) = \frac{var\left(I_{y.ig}^{(W)}\right)}{var\left(I_{y.ig}^{(W)}\right) + var\left(\epsilon_{y1ig}^{(W)*}\right)} = .75$$

$$var\left(\epsilon_{y1ig}^{(W)*}\right) = \frac{1 - .75}{.75} * var\left(I_{y.ig}^{(W)}\right) = \frac{1}{3} * 10 = \frac{10}{3}$$

Similarly,

$$R^2\left(y_{1.g}^{(B)}\right) = \frac{var\left(I_{y..g}^{(B)}\right)}{var\left(I_{y..g}^{(B)}\right) + var\left(\epsilon_{y1.g}^{(B)*}\right)} = .75$$

$$var\left(\epsilon_{y1.g}^{(B)*}\right) = \frac{1 - .75}{.75} * var\left(I_{y..g}^{(B)}\right) = \frac{1}{3} * 10 = \frac{10}{3}$$

We can also compute the variance of T1 residuals for x at within- and between-group levels in the same way:

$$R^2\left(x_{1ig}^{(W)}\right) = \frac{var\left(I_{x.ig}^{(W)}\right)}{var\left(I_{x.ig}^{(W)}\right) + var\left(\epsilon_{x1ig}^{(W)*}\right)} = .75$$

$$var\left(\epsilon_{x1ig}^{(W)*}\right) = \frac{1 - .75}{.75} * var\left(I_{x.ig}^{(W)}\right) = \frac{1}{3} * 10 = \frac{10}{3}$$

$$R^2\left(x_{1.g}^{(B)}\right) = \frac{var\left(I_{x..g}^{(B)}\right)}{var\left(I_{x..g}^{(B)}\right) + var\left(\epsilon_{x1.g}^{(B)*}\right)} = .75$$

$$var\left(\epsilon_{x1.g}^{(B)*}\right) = \frac{1 - .75}{.75} * var\left(I_{x..g}^{(B)}\right) = \frac{1}{3} * 10 = \frac{10}{3}$$

We can now compute the variance of time-specific innovations for x and y at within--and between-group levels using the information above:

$$var\left(d_{xtig}^{(W)*}\right)_{t\geq 2} = \frac{10}{3} * \frac{1}{2} \cong 1.67$$

$$var\left(d_{ytig}^{(W)*}\right)_{t\geq 2} = \frac{10}{3} * \frac{1}{2} \cong 1.67$$

$$var\left(d_{xtig}^{(B)*}\right)_{t\geq 2} = \frac{10}{3} * \frac{1}{2} \cong 1.67$$

$$var\left(d_{ytig}^{(B)*}\right)_{t\geq 2} = \frac{10}{3} * \frac{1}{2} \cong 1.67$$

Here, innovation variance at time 2 (T2) is set equal to innovation variance at each time point since innovation variances are set equal across time by design.

Consequently, the variance-covariance structure for residual at the first timepoint and innovation at later time points in the current condition ($\rho_{yxt} = .3$, PVB = .5, & R2B = .75) is expressed as follows:

$$var\begin{bmatrix} \epsilon_{y1ig}^{(B)*} \\ d_{y2i}^{(B)} \\ d_{y3i}^{(B)} \\ d_{y4i}^{(B)} \\ \epsilon_{x1ig}^{(B)*} \\ d_{x2i}^{(B)} \\ d_{x3i}^{(B)} \\ d_{x4i}^{(B)} \end{bmatrix} = var\begin{bmatrix} \epsilon_{y1ig}^{(W)*} \\ d_{y2i}^{(W)} \\ d_{y3i}^{(W)} \\ d_{y4i}^{(W)} \\ \epsilon_{x1ig}^{(W)*} \\ d_{x2i}^{(W)} \\ d_{x3i}^{(W)} \\ d_{x4i}^{(W)} \end{bmatrix} = \begin{bmatrix} 3.33 & & & & 0.01 & & & \\ & 1.67 & & & & 0.01 & & \\ & & 1.67 & & & & 0.01 & \\ & & & 1.67 & & & & 0.01 \\ 0.01 & & & & 3.33 & & & \\ & 0.01 & & & & 1.67 & & \\ & & 0.01 & & & & 1.67 & \\ & & & 0.01 & & & & 1.67 \end{bmatrix}$$

In the very same condition, the mean structure of growth factors at the between-group level is given as follows:

$$E \begin{bmatrix} I_{y..g}^{(B)} \\ S_{y..g}^{(B)} \\ I_{x..g}^{(B)} \\ S_{x..g}^{(B)} \end{bmatrix} = \begin{bmatrix} \mu_{Iy...}^{(B)} \\ \mu_{Sy...}^{(B)} \\ \mu_{Ix...}^{(B)} \\ \mu_{Sx...}^{(B)} \end{bmatrix} = \begin{bmatrix} 25 \\ 5 \\ 25 \\ 5 \end{bmatrix}$$

Where the overall mean of the random intercept across groups $\mu_{Iy...}^{(B)}$ and $\mu_{Ix...}^{(B)}$ are fixed at

25 and 4, respectively, for all conditions, following the example from Bovaird (2007). Using the

given mean structure and covariance structures, we can generate the multivariate normal data and

then assemble them into multilevel longitudinal data using equations (25) and (26).

In this study, a 2 (50 or 500 number of clusters) * 2 (.1 or .3 percentage of variance

between clusters on latent variables) * 2 (.5 or .75 R-square of repeated measures at between-

group level) * 2 (dominance or non-dominance CL) factorial design was employed to generate

data. For each cell condition, 10000 replications were generated using R based on the data-

generating model (i.e., two-level LCM-SR) with varying conditions.

For data analysis, Mplus 8.5 (Muthén & Muthén, 1998-2017) was used to investigate the

adequacy and robustness of three modeling approaches to address individual dependency in

multilevel longitudinal data under varying conditions. More specifically, Mplus has two built-in

routines for analyzing multilevel data (i.e., TYPE=TWO-LEVEL and TYPE=COMPLEX). First,

the TYPE=TWO-LEVEL routine was used for the disaggregated approach where the full two-

level LCM-SR was analyzed. Then, the TYPE=COMPLEX routine was used for the aggregated

approach in which the single-level LCM-SR was analyzed with consideration for

nonindependence of observations. The maximum likelihood estimation method with robustness

to non-normality and non-independence of observations (MLR estimation method in the Mplus

framework) was employed for both models. As for the third approach where the single-level

LCM-SR ignoring the nesting effects was analyzed, a regular ML estimation method was used. In the overall data analysis process, the R package MplusAutomation (Hallquist & Wiley, 2018) was used to run batches of simulations, extract results from the output files, and summarize the results.

The results were discussed as follows: First, models are evaluated in terms of differences in convergence rate and model fit indices. Then, the models are compared based on relative parameter bias, relative standard error (SE) bias, and 95% coverage rate for each parameter. In addition, empirical power or Type 1 error rate associated with the CL2 parameter was compared between the models.

Regarding model fit indices, the Chi-square test of exact fit and alternative model fit indices were used to compare the models. These alternative model fit indices include information criteria, absolute fit criteria, and incremental fit criteria. Information criteria are only interpretable when comparing two different models and are often used for comparing non-nested models. In this study, Akaike Information Criteria (AIC; Akaike, 1973) and Bayesian Information Criteria (BIC; Schwarz, 1978) were used. Note that in this study, AIC and BIC are not defined as a function of the Chi-square but as a function of the log maximum likelihood under the null hypothesis only (Hoyle., 2012; Muthén, 1998-2004). Incremental fit criteria (also called relative fit criteria) are analogous to R-square and so a value of zero indicates the worst possible model, while a value of one indicates the best possible model. They are based on the fit of a model relative to the worst possible model or the null model (e.g., constrain all the variables in the model to have no correlation and to have free means and variances). In addition, in this study, Comparative Fit Index (CFI; Bentler, 1990) and Tucker Lewis Index (TLI; Tucker & Lewis, 1973) were used. Finally, the absolute model fit criteria make a comparison directly to a

saturated or just-identified model with a value of zero indicating a perfectly fitting model. In this study, Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980) and Standardized Root Mean Square Residual (SRMR; Hu & Bentler, 1999) were used for model comparison. The latter is especially useful in multilevel modeling context in that it is computed separately for each level so that it can be used to locate the sources of misfit when the model is not fitting well (Asparouhov & Muthén, 2018). When considering model fit indices, it was assumed that the models favored by fit indices are deemed more valid, and if a model index doesn't indicate a preference, a more parsimonious model would be preferred.

In this study, relative point estimate bias, relative standard error (SE) bias, empirical power or type 1 error rate, and coverage were measured by using the following equations:

- Relative point estimate bias $= \dfrac{\text{Average parameter estimate} - \text{Population parameter value}}{\text{Population parameter value}}$

- Relative SE bias $= \dfrac{\text{Average SE of parameter estimate} - \text{empirical SD of parameter estimates}}{\text{Empirical sd of parameter estimates}}$

- Empirical power or Type 1 error rate $= \dfrac{\text{Number of cases with 95\% CI not covering zero}}{\text{Number of replications}}$

- 95\% coverage rate $= \dfrac{\text{Number of cases with 95\% CI covering population parameter value}}{\text{Number of replications}}$

Where average parameter estimate refers to the mean of parameter estimates over all replications within each cell condition. Average SE of a parameter estimate is the average of estimated SEs across replications within each cell condition. Empirical standard deviation (SD) of parameter estimates refers to the standard deviation of parameter estimates across replications within each cell condition, parameter estimate$_i$ refers to individual parameter estimate for one replication, and #rep refers to number of replications within each cell condition.

In this study, the relative bias gives two important pieces of information: the first is the magnitude of estimation bias, and the second is the direction of this bias (i.e., whether the certain parameter was overestimated or underestimated). The point estimate bias indicates how well the estimate represents the true nature of the relationship among variables. The SE bias indicates the adequacy with which our estimates are reliable.

When computing the SE bias, the empirical standard deviation of parameter estimates across replications was taken as the true sampling variance for the estimate. Indeed, with a sufficiently large number of replications, the empirical standard deviation can be seen as the true value for the variability across replications of the parameter estimates (Muthén, 2002). As for interpreting bias values, Hoogland and Boomsma (1998) suggested a guideline for where an absolute value of the bias less than .05 could be considered to represent a lack of bias. On the other hand, Muthén, Kaplan, and Hollis (1987) suggested the more lenient criterion that absolute values of the bias less than .10 to .15 might be considered negligible. In this study, Hoogland and Boomsma's criterion was used.

Finally, in this study, the empirical Type 1 error rate was defined as the proportion of replications for which 95% confidence interval (CI) that did not cover zero when the population value is indeed equal to zero, power defined as the proportion of cases 95% CI not covering zero when the population value is indeed different from zero. 95% coverage rate is defined as the proportion of replications for which the 95% CI covered the population value.

In Monte Carlo experiments, when the null hypothesis is true, that is, a population parameter is set equal to zero, each test should reject the null at about the nominal rate of 5%. If the type 1 error rates do not hold at $\alpha = .05$, the following empirical power for alternative

hypothesis tests would be biased. If type 1 error rates are not known, then discussing power would be meaningless since bias in power cannot be addressed. In this study, the null hypothesis is only set for the CL2 parameter (CL2 =0). Therefore, the empirical type 1 error and the empirical power were only discussed regarding the CL2 parameter.

Results

Model convergence

First, the frequency of cases where the models showed convergence issues were examined. SEM estimation methods are based on iterative techniques in which parameter estimates successively change until the model reaches a pre-specified, minimum convergence criterion (Bandalos & Gagné, 2012). However, there are cases where the model does not converge, converges but gives parameter estimates that are unobtainable (improper solutions; Chen, Bollen, Paxton, Curran, & Kirby, 2001), or they occur together. In this study, for a given replication the model was considered to have successfully converged if it did not produce any warnings or error messages.

Table 1 shows the frequency of non-convergence, improper solutions, and proper solutions across all the replications for each model and each condition, and then the overall number of proper solutions across the models and for each condition. Results for complex and default models are put together since they showed the same number of non-convergence and improper solutions. In fact, the complex model is the extension of the default model with corrections to the standard errors and Chi-square test of model fit that take into account stratification, non-independence of observations, and unequal probability of selection (Muthén & Muthén, 1998-2017).

The results showed that in general, all the models showed high rates of non-convergent solutions or improper solutions for conditions with low sample size (nc=50; N=250). Note that these results are from the models with zero constraints on innovation covariances in addition to the equality constraints imposed on the residual structure. As the sample size increases the number of non-convergent solutions or improper solutions decreased (around 1%).

There was a difference between the two-level model and the complex/default model in terms of non-convergence and improper solution rates. In general, the complex and default models showed more non-convergent solutions than the 2level model. Yet in most cases, the differences were small since only a small fraction of replications showed non-convergent solutions (up to 5% of all replications in each condition). On the other hand, when comparing the numbers of improper solutions, the 2level model showed more improper solutions than the complex/default model, and the difference was huge in most cases. In small sample size conditions, the two-level model showed improper solutions for more than half of the total replications, whereas complex and default models showed a lot less, even though the numbers were still substantial (more than 10%). In some of the middle sample size conditions, the two-level model still showed a substantial number of improper solutions. In the large sample size conditions, all the models showed low numbers of non-convergent or improper solutions.

Convergence issues can be caused by sampling variability, a poorly specified model, poor starting values, a lack of identification, or many other factors. In this example, we can see that the number of non-convergent or improper solutions is greatly impacted by sample size (the number was the highest in conditions with nc=50 and then significantly dropped as NC increased). In addition to sample size, other design factors had their impact as well. For example, between-level variance had varying effects depending on the utilized model. For the two-level

model, the number of improper solutions was larger on average in conditions with PVB=0.2 than PVB=0.5. For the other models, it was the opposite - the number of improper solutions was smaller on average in conditions with PVB=0.2 than PVB=0.5. This is understandable in that the two-level models are meant to have the two-level model structure. If the model is fit to the data which are likely to have a single-level model structure, the model would not perform well. In the follow-up analysis to find the source of improper solutions, it was found that most improper solutions are caused by between-group level components (e.g., the non-positive definite covariance matrix for between-group level latent variables). On the contrary, other models assume a single-level structure and if there is much variance going on in the between-level, the models would not be a good fit for the data. R2B and CL also had their impact on the number of non-convergent or improper solutions in that the numbers were bigger on average in conditions with R2B=0.5 than R2B=0.75, and that the numbers were bigger on average in conditions with cl=0.1 or 0 than cl=0.3. Yet, R2B made a substantial difference while CL conditions did not. This is also understandable in that low R2B itself means that the growth portion of the model did not explain much variance in the data, which implies that the results would be more unstable.

Researchers must decide if they will generate other samples to replace those that did not converge or perhaps base results only on the samples that do converge. Replacing the non-converged samples with new ones has the advantage of maintaining a balanced design. However, in studies of extreme conditions, non-convergence may be so pervasive that replacement of nonconvergent samples would be unrealistic. Most researchers agree, however, that non-convergent solutions should be screened out before analyzing data from simulation studies (Bandalos & Gagné, 2012).

In this study, the first 5000 successfully converged solutions across the models in each condition were included for further analysis. However, in most of the small sample size conditions, the two-level model did not converge in more than half of all replications, hence lowering the overall number of convergent solutions in each condition. Therefore, in small sample size conditions, only results for the complex model and the default model were included for further comparison. Assessment of individual outcomes was performed using only these successfully converged replications.

Table 1. *A frequency of non-convergence, improper solutions, and proper solutions*

		PVB=0.2						PVB=0.5					
Non-convergence		R2B=0.5			R2B=0.75			R2B=0.5			R2B=0.75		
NC	Model	cl=0	cl=0.1	cl=0.3	cl=0	cl=0.1	cl=0.3	cl=0	cl=0.1	cl=0.3	cl=0	cl=0.1	cl=0.3
	2level	211	196	2	194	187	5	188	160	3	188	198	8
50	complex/default	284	288	28	180	187	3	552	525	147	460	425	71
	2level	1	1	1	1	2	1	1	1	1	1	1	1
200	complex/default	5	1	1	1	3	1	43	30	1	15	19	1
	2level	1	1	1	1	1	1	1	1	1	1	1	1
500	complex/default	1	1	1	1	1	1	1	1	1	1	1	1
Improper solutions		PVB=0.2						PVB=0.5					
		R2B=0.5			R2B=0.75			R2B=0.5			R2B=0.75		
NC	Model	cl=0	cl=0.1	cl=0.3	cl=0	cl=0.1	cl=0.3	cl=0	cl=0.1	cl=0.3	cl=0	cl=0.1	cl=0.3
	2level	7064	7076	6736	6582	6664	6198	5083	5079	4599	2909	3029	2118
50	complex/default	1921	2022	1073	1080	1078	277	4092	4337	3783	2314	2446	1263
	2level	1608	1540	1176	384	390	252	838	771	477	27	44	4
200	complex/default	124	103	5	16	22	1	893	836	293	160	164	6
	2level	136	143	57	9	4	6	50	51	7	1	1	1
500	complex/default	1	2	1	1	1	1	114	83	5	8	3	1
Proper solutions		PVB=0.2						PVB=0.5					
		R2B=0.5			R2B=0.75			R2B=0.5			R2B=0.75		
NC	cl2	cl=0	cl=0.1	cl=0.3	cl=0	cl=0.1	cl=0.3	cl=0	cl=0.1	cl=0.3	cl=0	cl=0.1	cl=0.3
	2level	3147	3120	3266	3612	3523	3807	5105	5081	5404	7279	7169	7890
	complex/default	8363	8266	8955	9100	9109	9726	6460	6188	6364	8146	7979	8808
50	total	2981	2930	3160	3539	3459	3780	4038	3890	4247	6462	6219	7244
	2level	8393	8461	8825	9617	9612	9749	9163	9230	9524	9974	9957	9997
	complex/default	9881	9898	9996	9985	9981	10000	9150	9194	9708	9855	9855	9995
200	total	8352	8431	8825	9612	9605	9749	8579	8680	9335	9842	9832	9993
	2level	9865	9858	9944	9992	9997	9995	9951	9950	9994	10000	10000	10000
	complex/default	10000	9999	10000	10000	10000	10000	9887	9918	9996	9993	9998	10000
500	total	9865	9858	9944	9992	9997	9995	9850	9878	9990	9993	9998	10000

Model fit

Next, model fit values are compared between the models. When comparing models in terms of the fit we looked at the average value and the standard deviation of fit values. All the results are presented in figures with lines where lines represent the model result with a ± 1 standard deviation bar around them. Note that lines for the two-level are broken in some figures, which indicates the results for the two-level model are not given in those specific conditions. This is because the results for the two-level model in small sample size conditions are excluded from the current analysis.

Figure 9 and Figure 10 show the AIC and BIC values for each of the models separately across different conditions. When looking at AIC mean values and standard deviations, it is seen that AIC values are largely influenced by sample size. The results in the small sample size conditions showed lower values on average than other results, and the results in the middle sample size conditions showed lower values than the rest of the results. This is because, by definition, AIC values are proportional to sample size (Muthén, 1998-2004). Among conditions with the same sample size, AIC mean values were higher when PVB=0.5 or R2B =0.5 than when PVB=0.2 or R2B =0.75, but the magnitude of change was minimal. As for CL2 values, there was no noticeable difference between conditions with different CL2 values. Finally, AIC values showed that two-level models are slightly better fitting than other models in general, while the complex and default model showed the same results since AIC values for the complex model are not adjusted from those for the default model. There was no systematic difference found in AIC's standard deviations. A similar pattern of results was found when interpreting BIC values. This is because they are both direct functions of -2 times loglikelihood (Akaike, 1987; Schwartz, 1978).

Figure 9. *AIC values for each of the models across study conditions*
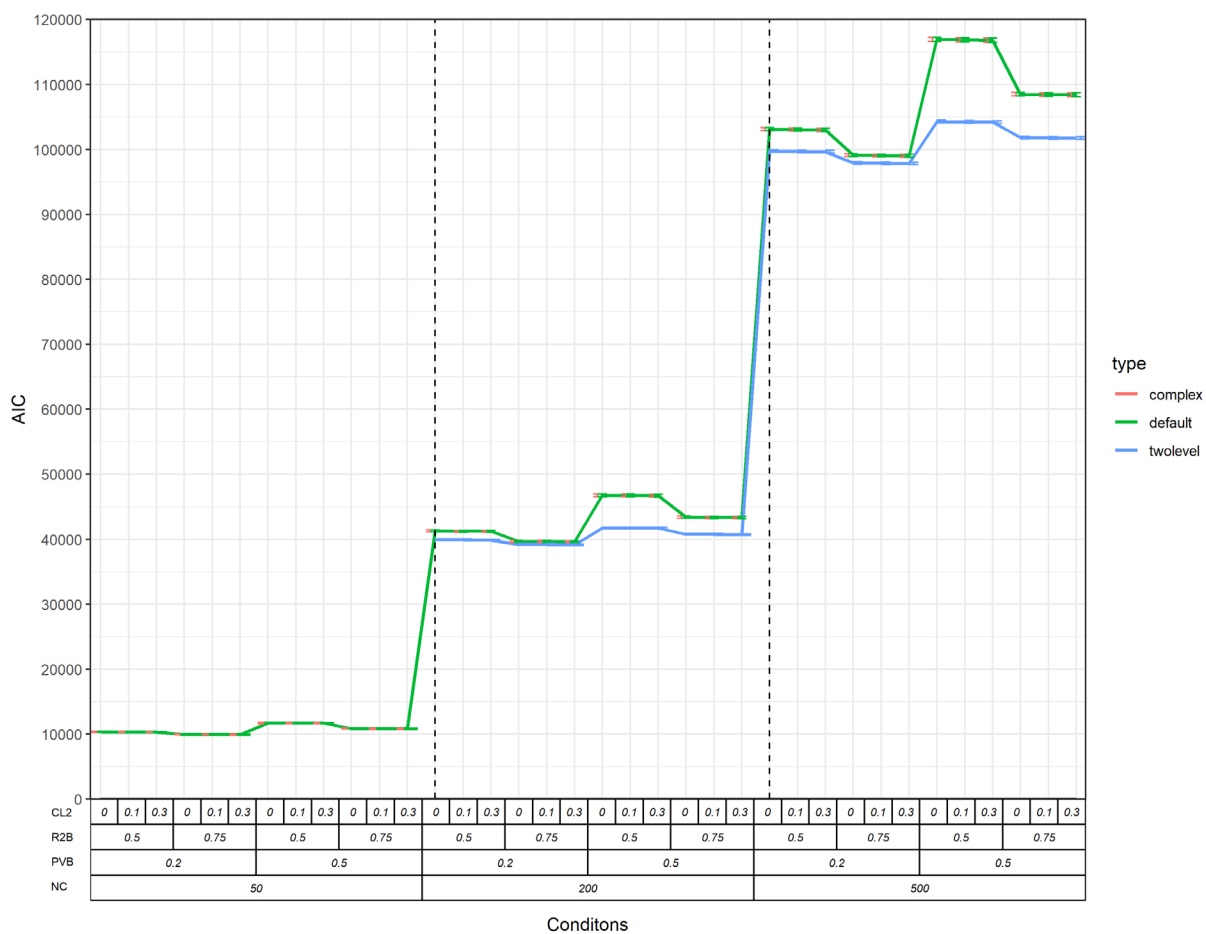
Figure 10. *BIC values for each of the models across study conditions*

However, when looking at Chi-square values, which are given in Figure 11, the results showed an interesting pattern of difference. When the hypothesized model is correctly specified to be the true model, the likelihood ratio test statistic approaches a central Chi-square distribution where the mean values are the degrees of freedom of the model (Alavi et al., 2020). Here, Chi-square mean values of the two-level model are near the degrees of freedom (i.e., 44) across all conditions since they are all deemed to be correctly specifying the true model. Note the line is cut off on the left side of the plot because the results are not given for small sample conditions due to high rate of improper solutions. In addition, those for the complex model are also close to their degrees of freedom because they are adjusted for non-normality and non-independence of

observations. As for the default model, however, the influence of non-normality (due to sampling variability) and non-independence (due to design factors) is reflected in unadjusted Chi-square mean values and standard deviations in that they were almost twice as much on average ~~as~~ to those in the complex model. On average, they were bigger when PVB=0.5 or R2B=0.5 than when PVB=0.2 or R2B=0.75 (there was no noticeable difference found between conditions with different CL values). As for Chi-square standard deviation values, those for the default model were the biggest, followed by the ones for the two-level models and then the ones from the complex model. As a result, there was a clear tendency that Chi-square values favored the complex model. For the complex model, there was no discernible difference found in Chi-square

---

Bentler, 1999), which indicated a good fit of the model to the sample along with standard

deviations that are close to 0. However, in small sample size conditions, there was a discernible

difference found between the complex and default models. Also, the RMSEA mean values for

the default model were also above the conventional cutoff scores. This implies that the

magnitude of the difference is inversely proportional to sample size, which is different from Chi-

square results. In addition, other results were similar to those of Chi-square in that RMSEA were

also larger when PVB=0.5 or R2B=0.5 than when PVB=0.2 or R2B=0.75, on average.

Figure 12. *RMSEA values for each of the models across all other study conditions*
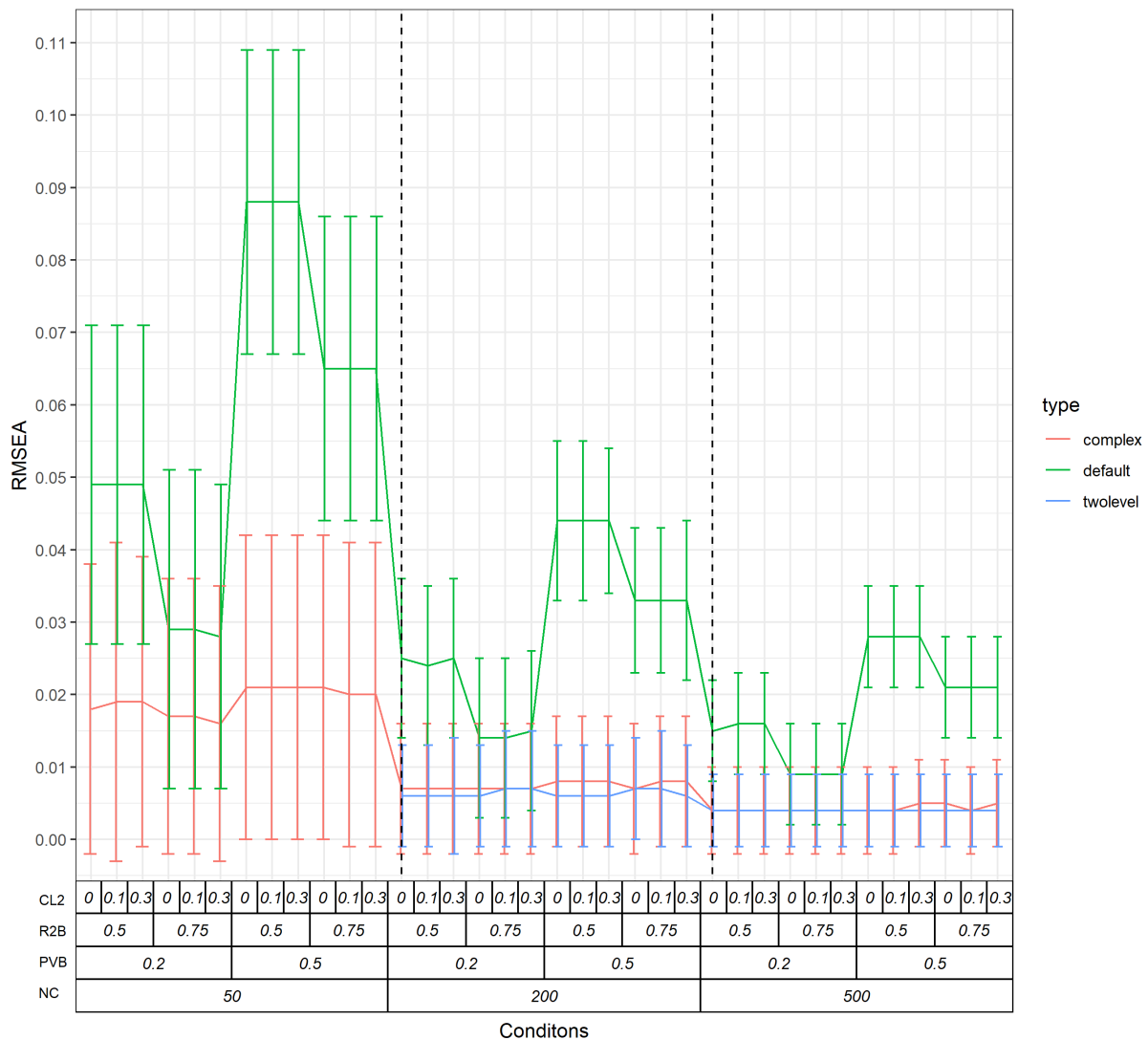
Figure 13 and Figure 14 depicts the CFI and TLI values for each of the models separately across different conditions. CFI and TLI favored the two-level and complex models over the default model and there was also a difference in mean values and standard deviations between the two-level and complex models. However, the differences were just minor fluctuations, and all average values and standard deviations are equal to or close to 1 and 0, respectively.

Figure 13. *CFI values for each of the models across all other study conditions*
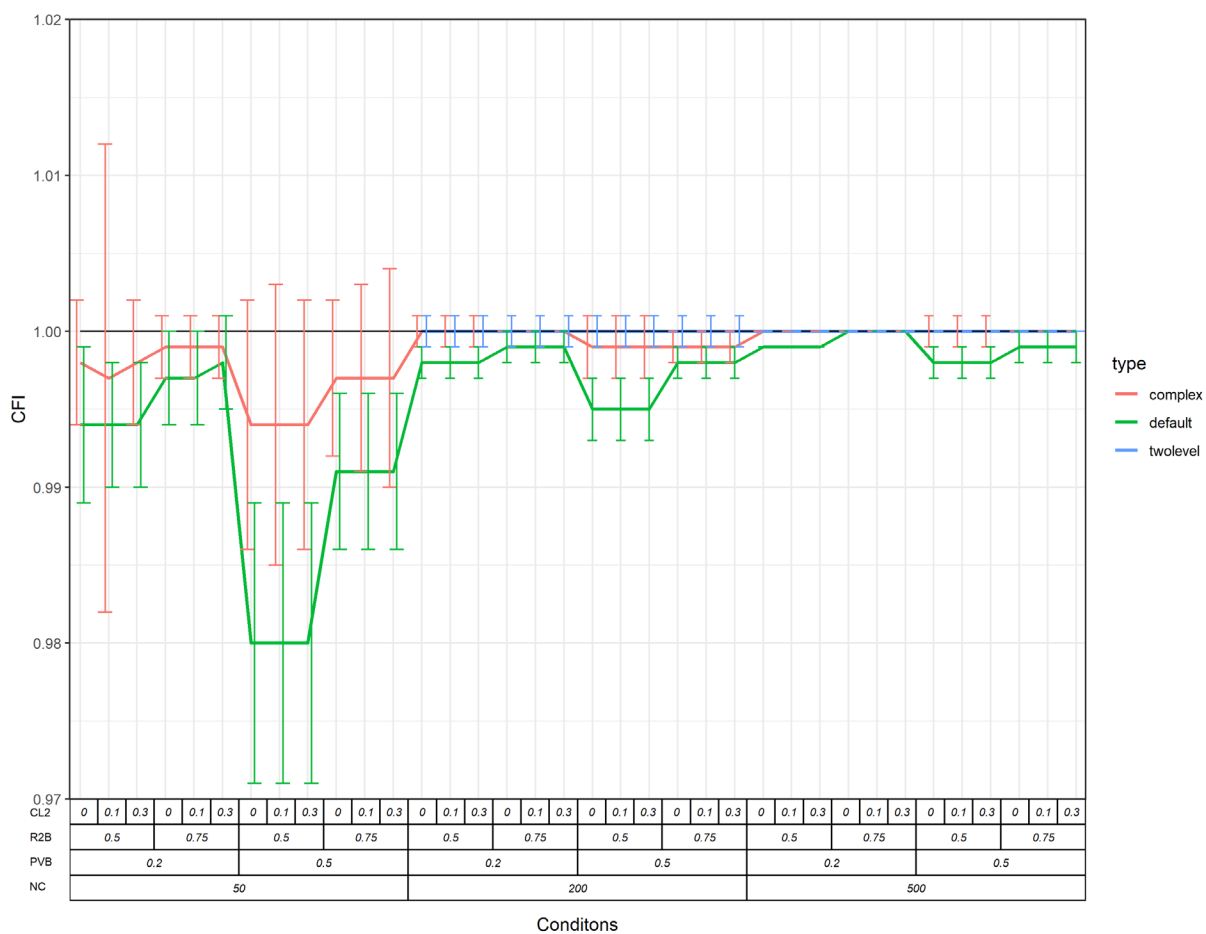
Figure 14. *TLI values for each of the models across all other study conditions*

The chart shows TLI (y-axis ranging from 0.97 to 1.02) plotted against Conditions (x-axis). The bottom table shows the following condition rows:

CL2	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3	0	0.1	0.3
R2B	0.5			0.75			0.5			0.75			0.5			0.75			0.5			0.75			0.5			0.75			0.5			0.75		
PVB	0.2						0.5						0.2						0.5						0.2						0.5					
NC	50												200												500											

Legend (type): complex (red), default (green), twolevel (blue)
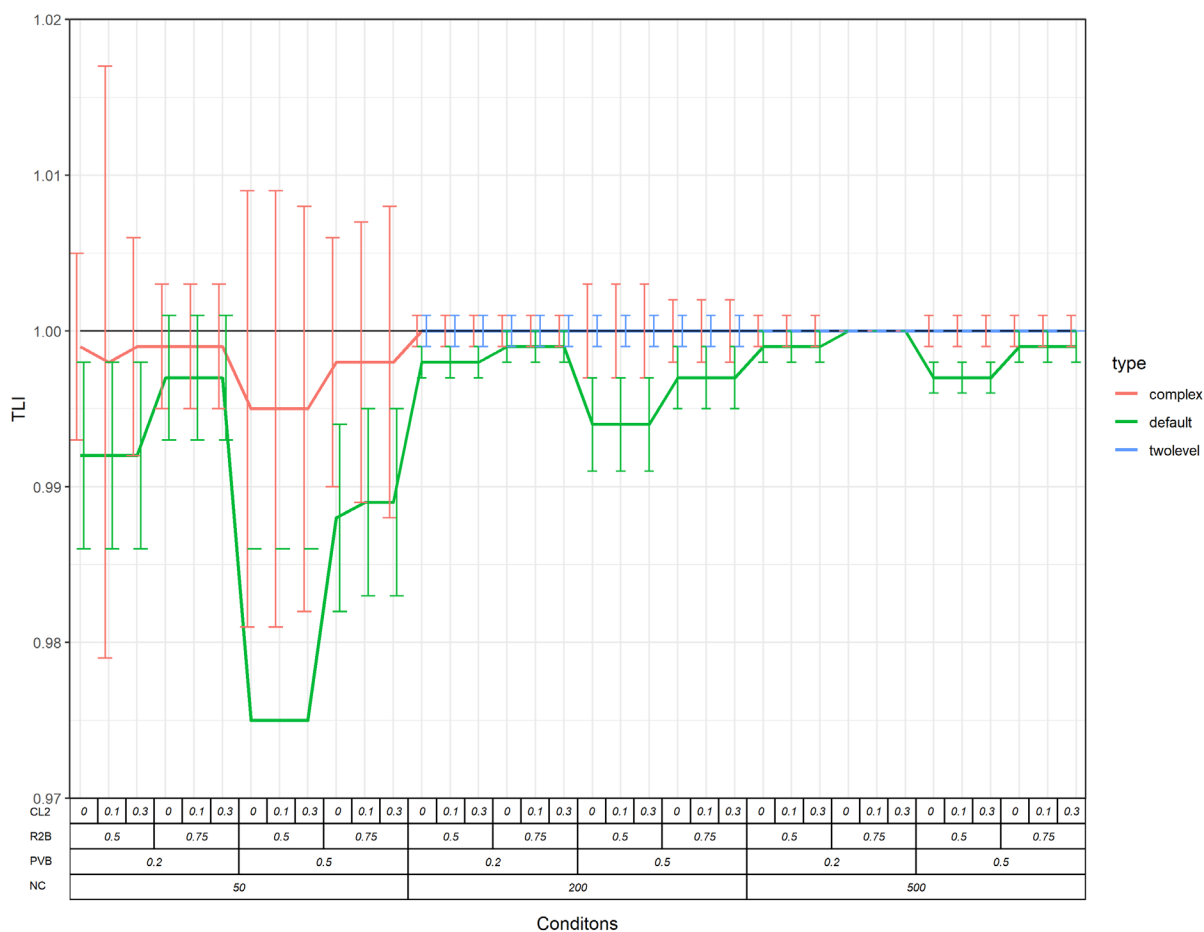
Finally, SRMR values for the complex model and default models were the same since SRMR is not a function of chi-squares. The results are depicted in Figure 15. The results showed that SRMR values increase as the sample size gets smaller, and SRMR values were also larger when PVB=0.5 and R2B=0.5 than when PVB=0.2 or R2B=0.75, on average. Yet, they were all below the conventional cutoff score (SRMR <.05; Hu & Bentler, 1999). When looking at SRMR values for the two-level model, there were two lines plotted. Here, the line on the top is the SRMR-Between values and the one on the bottom is the SRMR-Within values, which indicates that between-level components showed more degree of misfit than within-level components. Nevertheless, all the mean values were again below the conventional cutoff values.

Figure 15. *SRMR values for each of the models across all other study conditions*
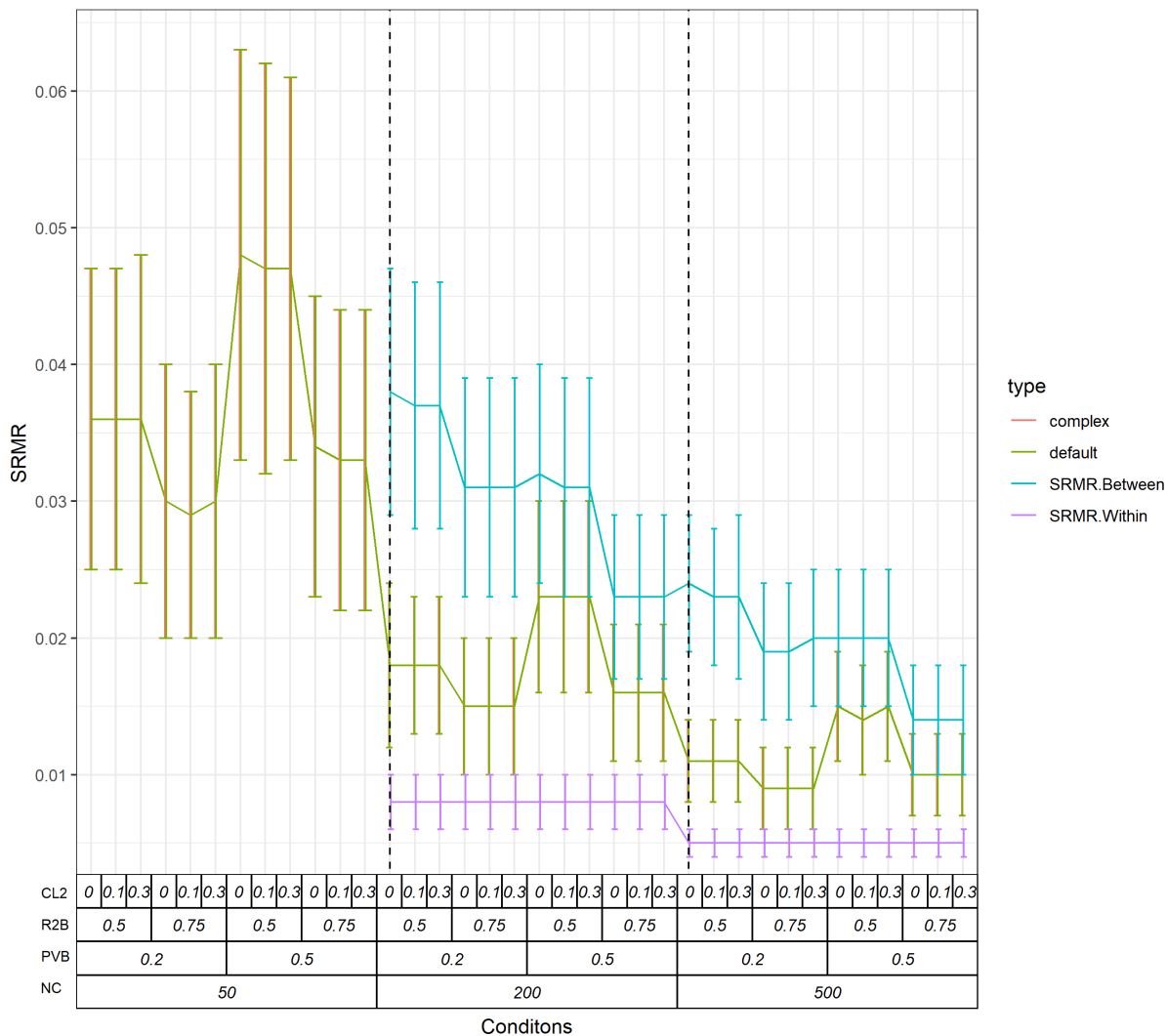
Overall, the results showed that information criteria, in general, are largely influenced by sample size and gave little information about that two-level model is better fitting than others. Absolute fit criteria showed that the two-level model and the complex model are well-controlling for non-normality and non-independence. Finally, relative fit criteria give no useful information at all. Note that in the study, all the models fitted to the data have no systematic model misspecification.

Bias, Coverage, Power, and Type 1 error

Bias, 95% coverage, empirical power, and empirical Type 1 error rates for the default &

complex model were checked and compared to the two-level model. As for the two-level model

results, fixed effect parameters from the two-level model were directly comparable to those from

the default and complex model, while random effect parameters were not since they were

estimated separately for each level. Therefore, the Mplus model constraints command was used

to combine separate parameter estimates at within- and between-level and compute total

expected variance estimates across levels. Covariance parameters were excluded from

comparison since they cannot be combined.

Results for parameter bias are presented in Figure 16 through Figure 19. The arrangement

of design factors for each graph is the same as the ones from model fit results. Note that the

results for the complex and default models are the same since the point estimates are not related

to Chi-square correction. Also note that lines in the CL2 parameter graph are broken because,

when computing the relative point estimate bias, the population parameter is used as a

denominator. In some conditions, the population parameter for CL2 is 0, which results in

positive or negative infinities. Yet, the remaining graph showed that there is not much bias

except in one condition where bias was over 0.1. Other results showed that in general, there was

no or little bias in point estimates for every parameter in middle and large sample size

conditions. However, there was a noticeable fluctuation among parameters for the complex and

default models in small sample size conditions. Specifically, there was a larger negative bias

when PVB=0.5 and R2B=0.5 than when PVB=0.2 or R2B=0.75, on average. For the latent

variable variance structure, the results showed upward bias in the same conditions. Only the

latent variable mean structure showed no bias at all and across all conditions. There was no discernible difference found between the models.

Figure 16. *Parameter bias on autoregressive (AR) and cross-lagged (CL) processes of the model*

Figure 17. *Parameter bias on the mean structure of the model*
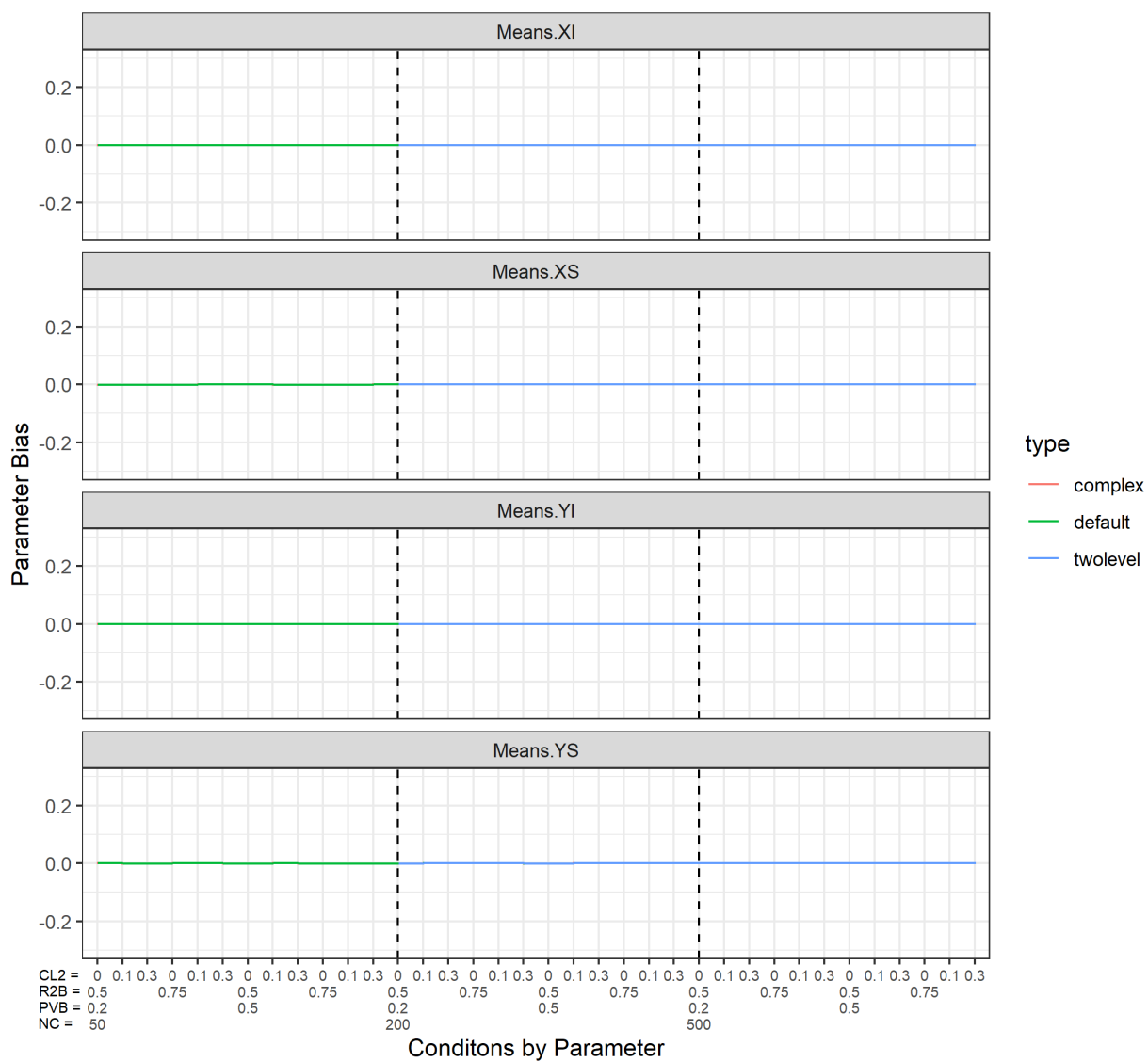
Figure 18. *Parameter bias on variance structure of the model*

Figure 19. *Parameter bias on residual variances of the model*

Results for SE bias are presented in Figure 20 through Figure 23. The results showed that there were discernible differences found for every parameter between the models. For the AR/CL parameters, there was a large negative bias found on the result for the default model, whereas the results for the complex and two-level models did not show much bias except for AR parameters in small sample size conditions. The complex model showed an upward bias in the conditions with PVB=0.5, R2B=0.5, and CL2=0.1 or 0.3 than other conditions. The negative bias for the

default model was larger when PVB=0.5 and R2B=0.5 and smaller when PVB=0.2 and

R2B=0.75.

Figure 20. *SE bias on autoregressive (AR) and cross-lagged (CL) processes of the model*
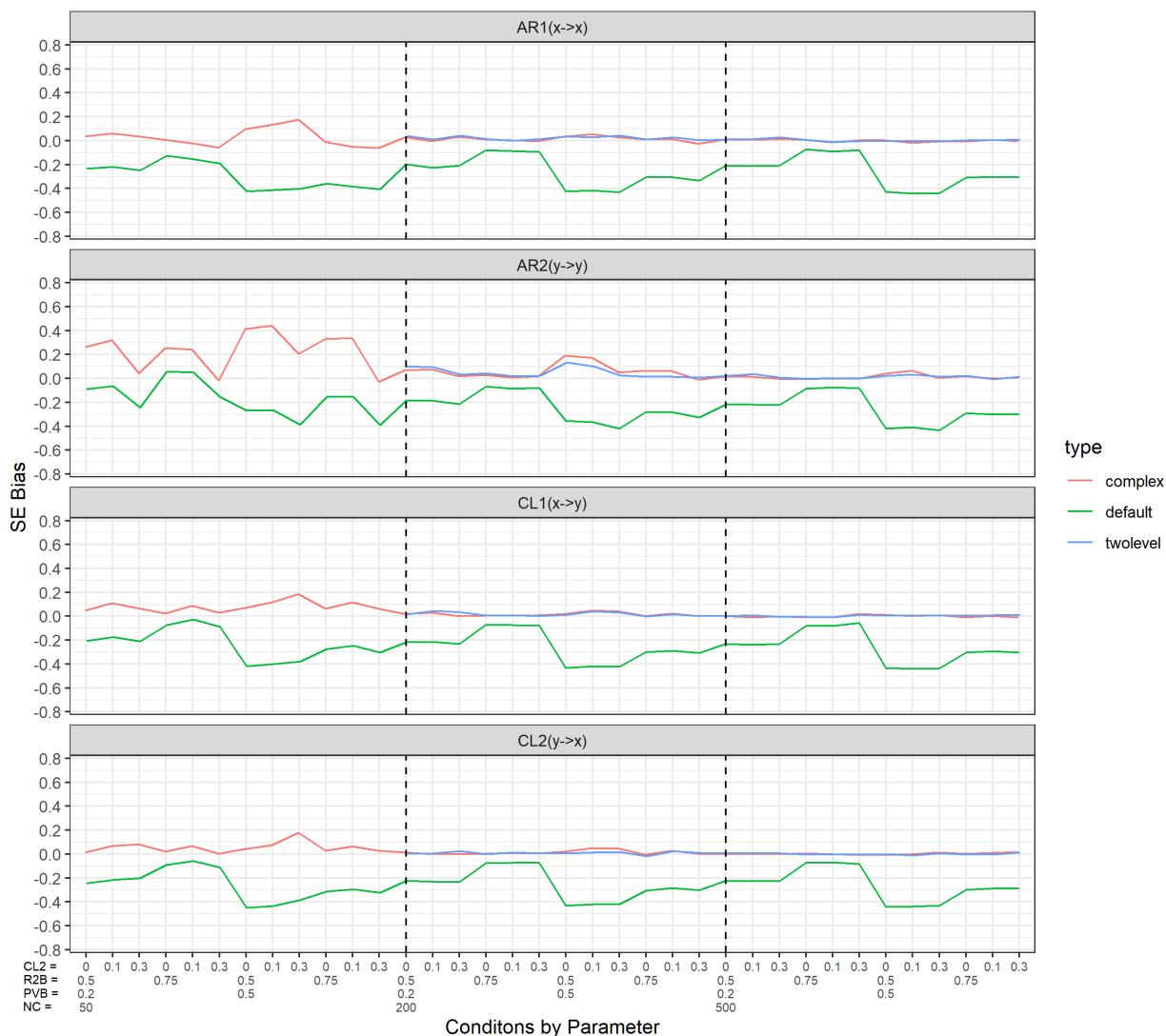
As for the mean structure of growth factors, results from the two-level and complex

models were almost identical and showed no bias. On the other hand, the default model showed a

large negative SE bias and the downward bias become larger in the conditions with R2B=0.5

than in other conditions.

Figure 21. *SE bias on the mean structure of the model*
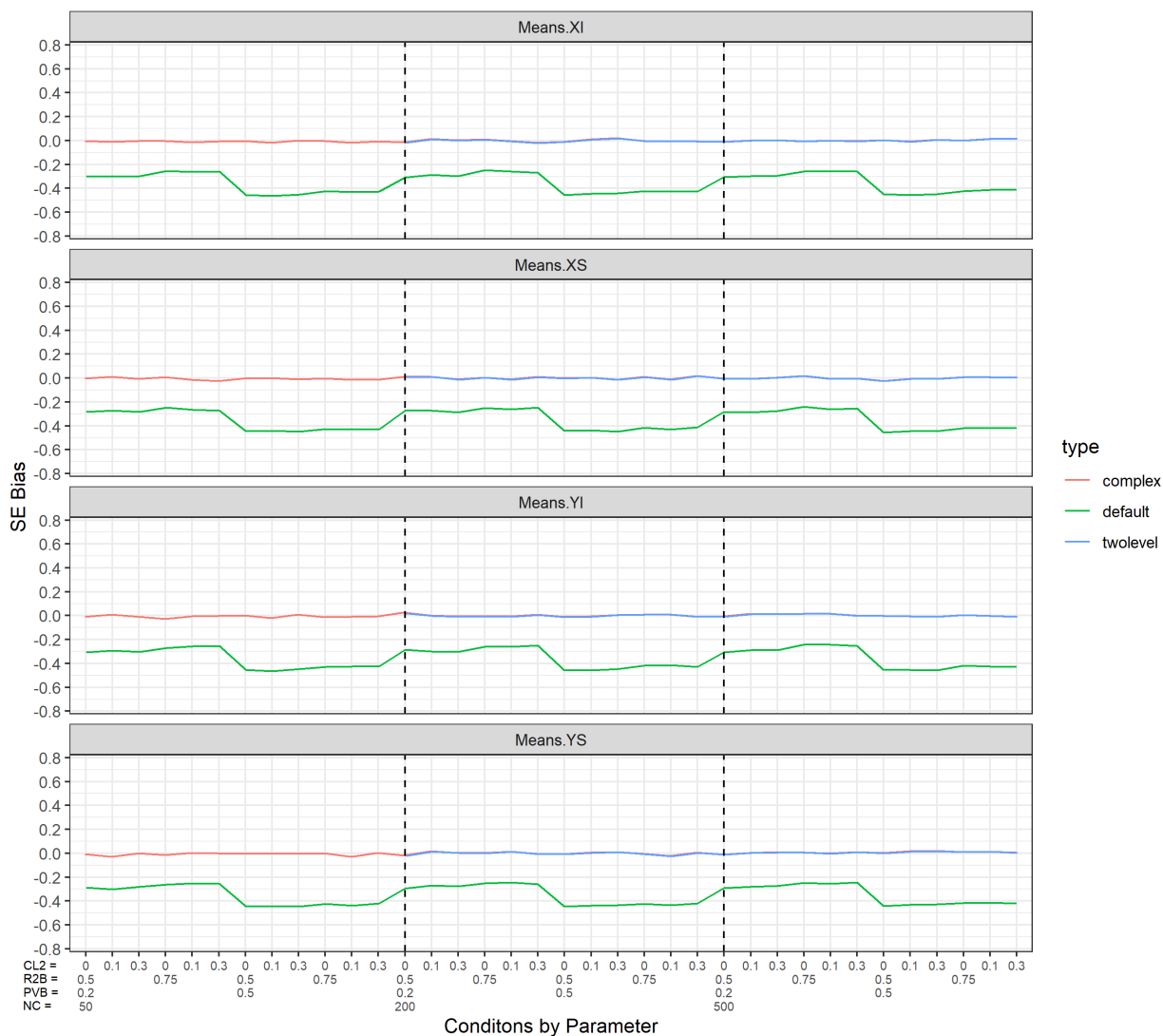
As for the latent variable variance structure, the results from the two-level model showed

no bias. The complex model also showed almost no SE bias except that a positive SE bias was

observed for the x-intercept and y-intercept variances in small sample size conditions with

PVB=0.5 and R2B=0.5. The results from the default model also showed a negative SE bias in

general, especially in conditions with PVB=0.5 and R2B=0.5. The default model also showed

large negative SE biases for the latent variable covariance components, which is especially larger in conditions with PVB=0.5 and R2B=0.5.

Figure 22. *SE bias on variance structure of the model*
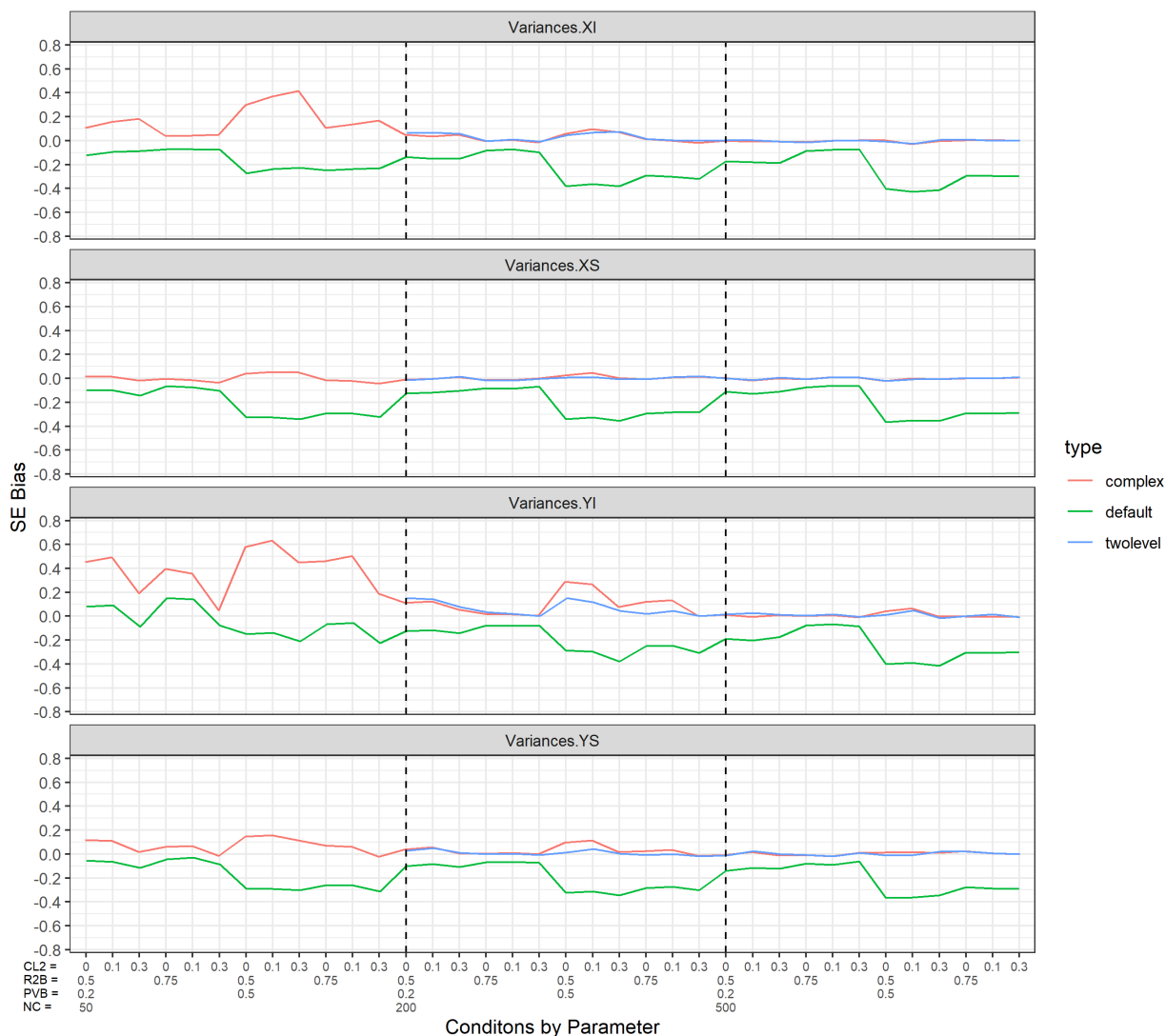
Regarding the residual variance structure, results showed a similar pattern as in the latent variable variance structure. The complex model results showed no SE bias except for small sample size conditions with PVB=0.5, where upward bias was observed in general. The default

model results showed large negative SE biases for all variance components, especially in

conditions with PVB=0.5 and R2B=0.5.

Figure 23. *SE bias on residual variances of the model*
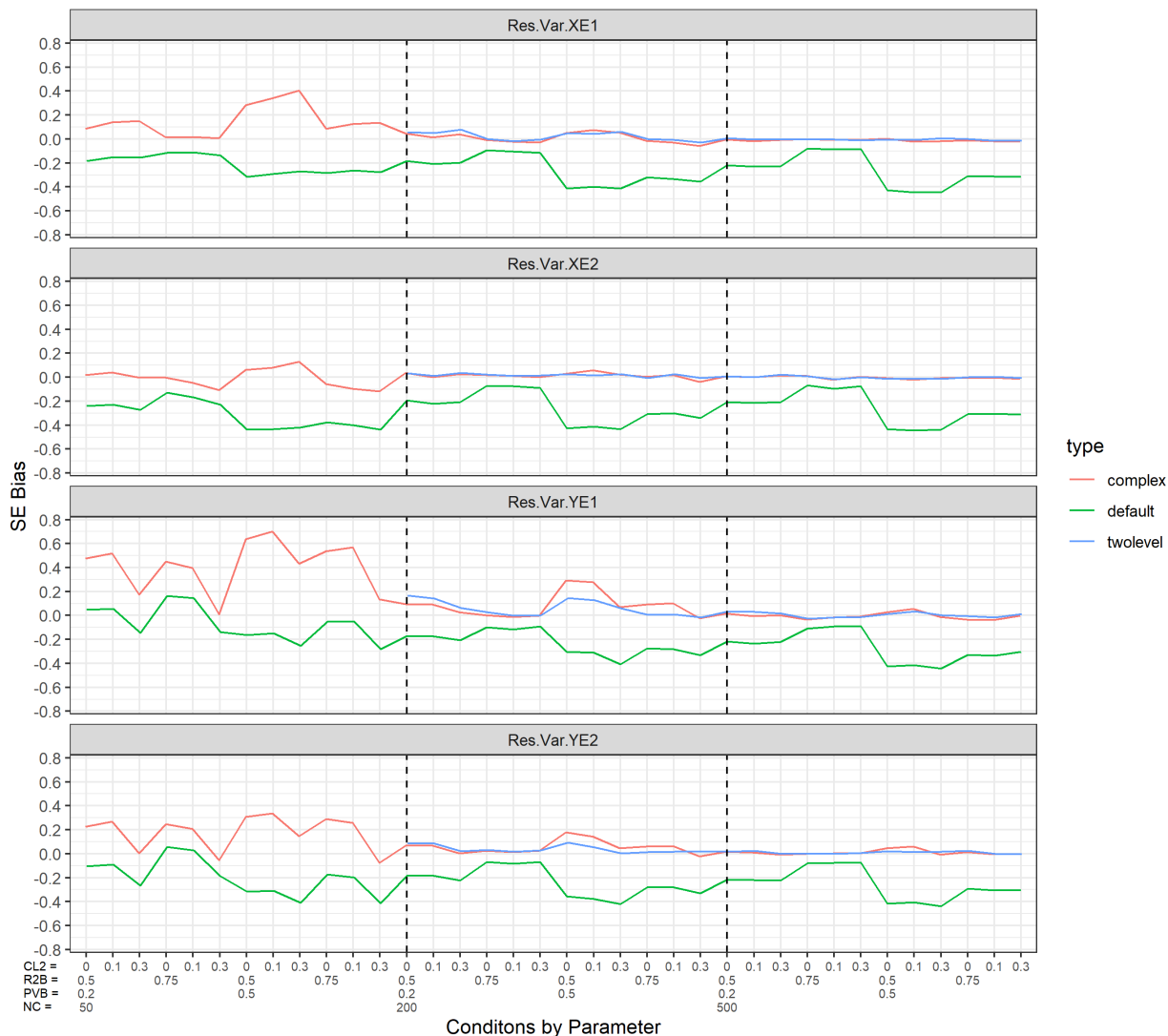
Results for the 95% coverage are presented in Figure 24 through Figure 27. The results

showed that for the AR/CL2 parameters, the two-level and complex models showed a decent

coverage rate of over 90% except for the small sample size condition with PVB=0.5, R2B=0.5,

and L2=0.3 where the complex mode showed the coverage rate below 90%. On the other hand,

the default model showed a low coverage rate in conditions with PVB=0.5 and R2B=0.5, in

general.

Figure 24. *95% coverage on autoregressive (AR) and cross-lagged (CL) processes of the model*
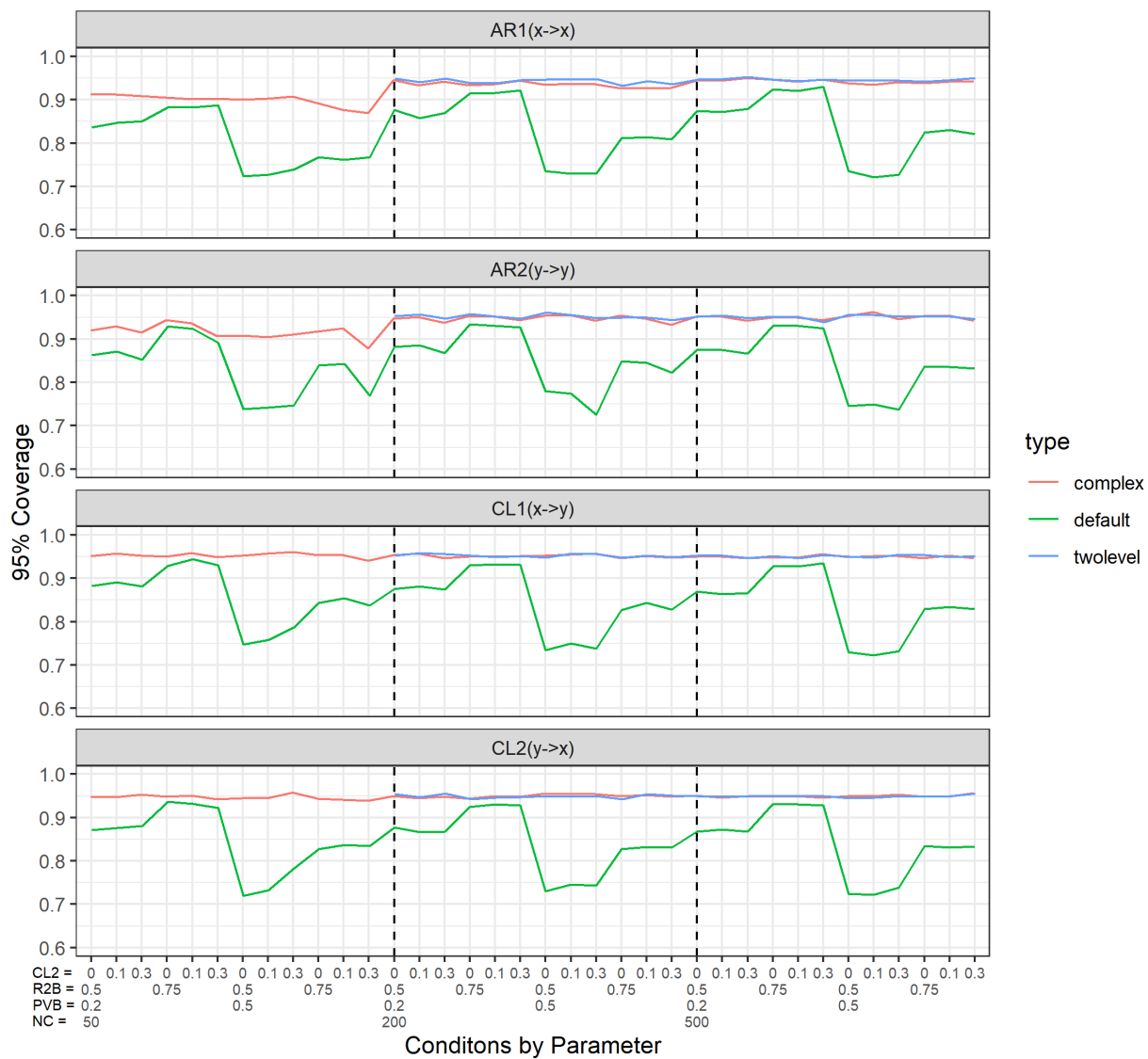
As for the latent variable mean structure, the pattern of difference was more solid than the AR/CL parameters. The two-level and complex models showed a decent coverage rate across all conditions. On the other hand, the default model showed a low coverage rate in conditions with PVB=0.5, in general.

Figure 25. *95% coverage on the mean structure of the model*
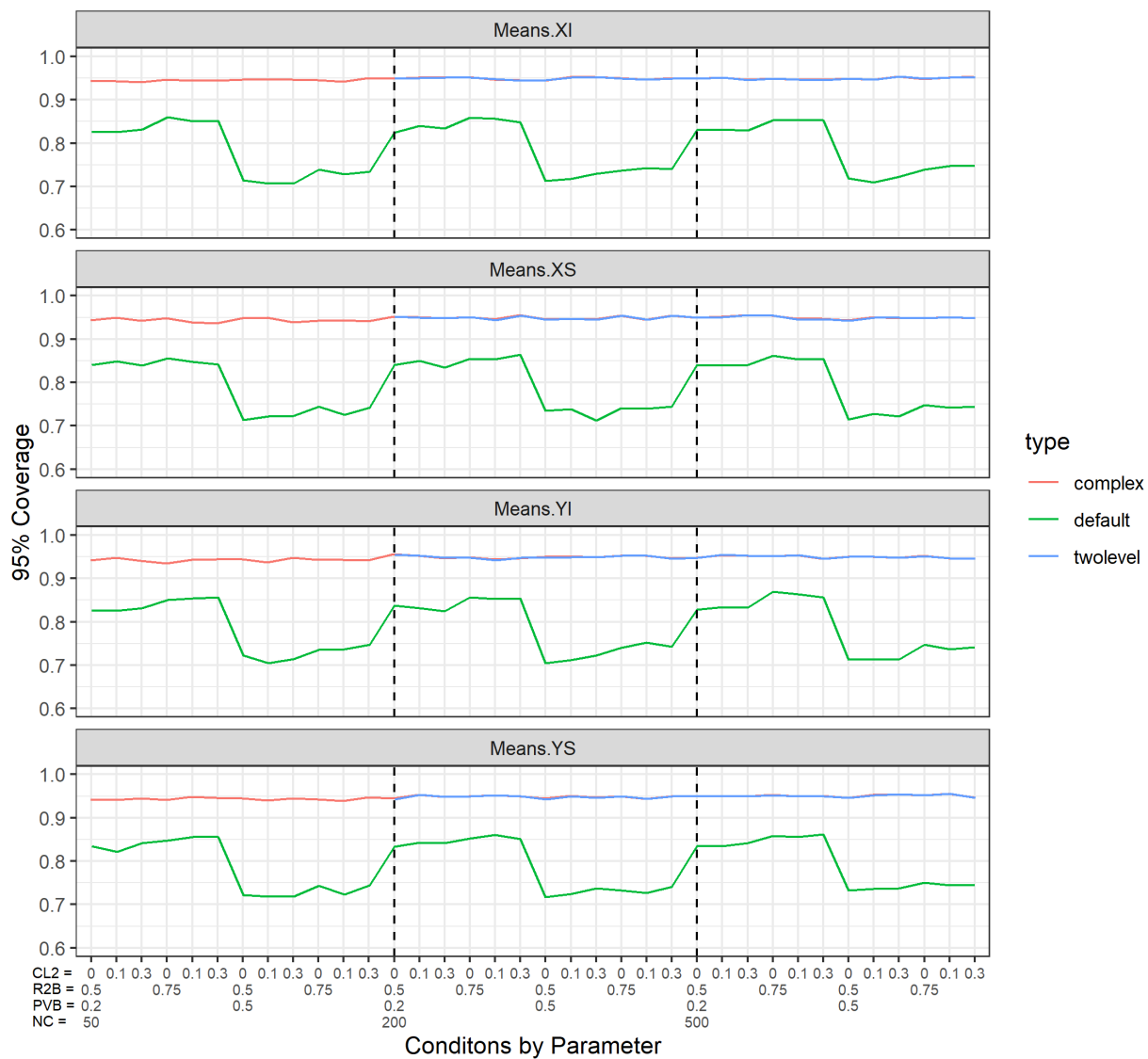
As for the latent variable variance structure, a similar pattern of difference was observed as for the mean structure. The two-level and complex models showed a decent coverage rate across all conditions. On the other hand, the default model showed a low coverage rate in conditions with PVB=0.5 yet the coverage rate was higher than that of the mean structure.

Figure 26

95% coverage on variance structure of the model
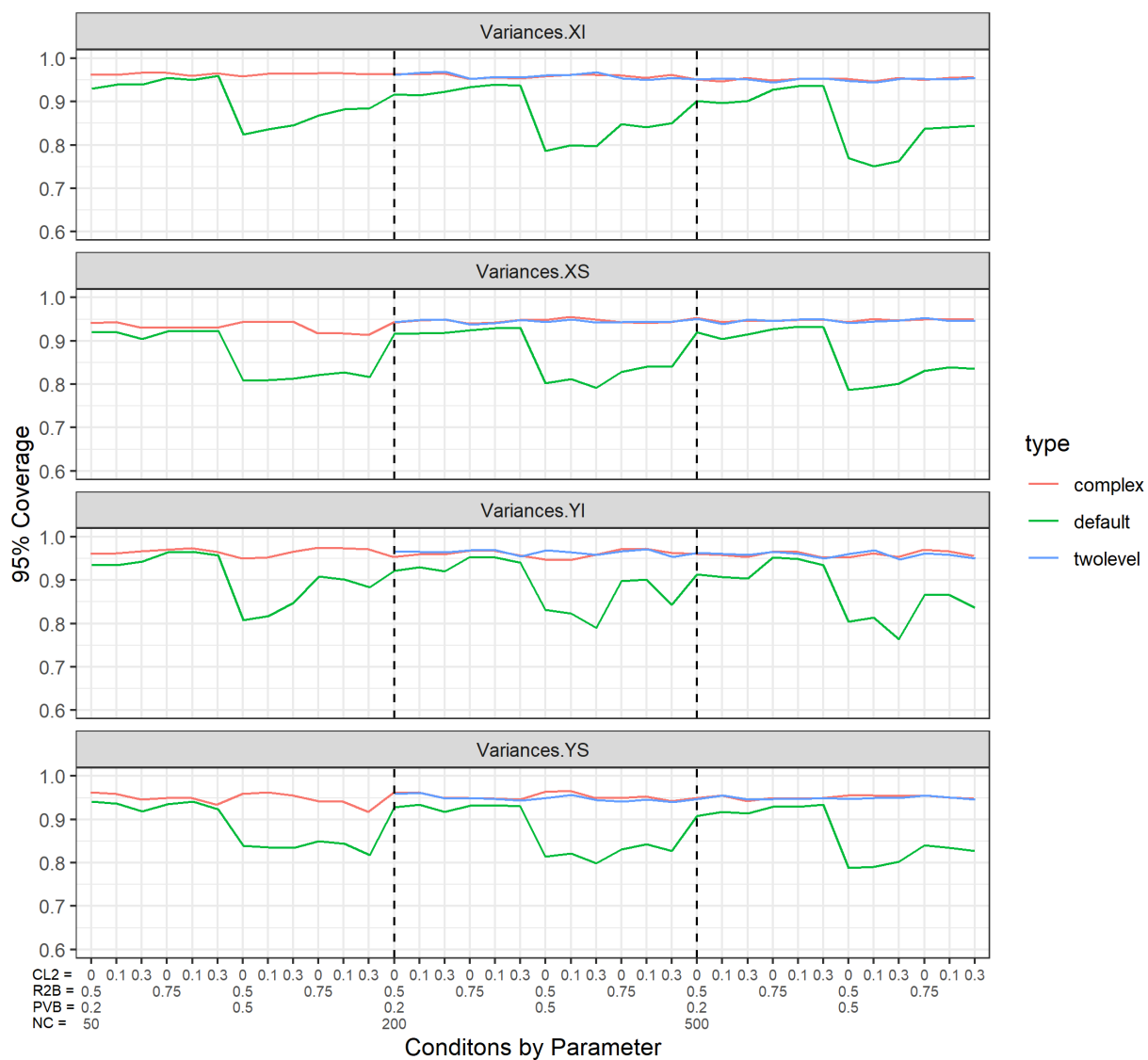
For the residual variance structure, the results showed that the two-level and complex models showed a decent coverage rate of over 90% except for the small sample size condition with PVB=0 and R2B=0.5, where the complex mode showed the coverage rate to be below 90%. The default model showed a low coverage rate and especially lower in conditions with PVB=0 and R2B=0.5.

Figure 27. *95% coverage on residual variances of the model*
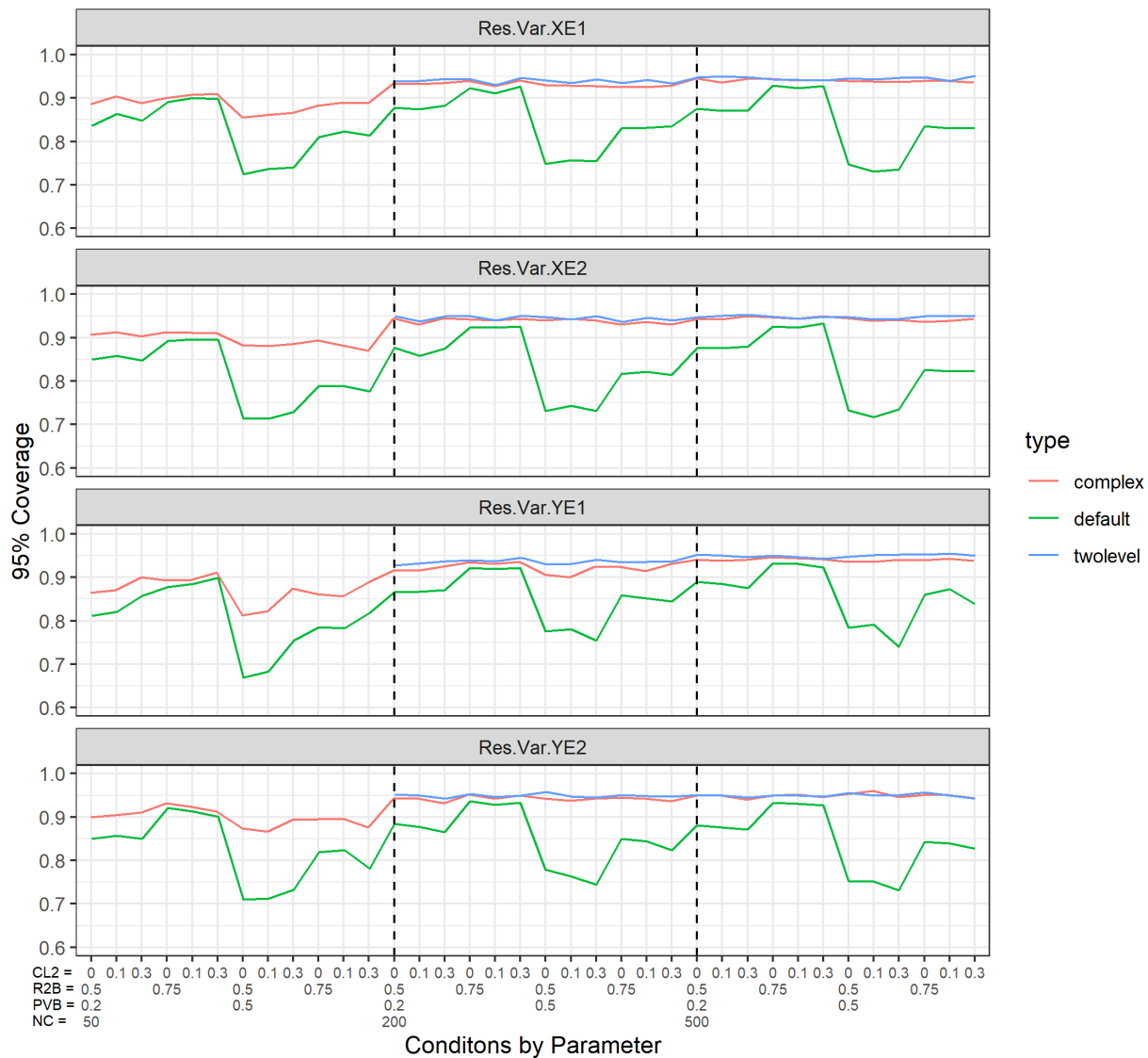
Finally, results for empirical power and type 1 error for the CL2 parameter are presented in Figure 28. Note that the line is broken again for the two-level model but a different reason. Here, the arrangement of design factors for each graph is different than before. This is done to make the results easier to understand. Recall that when the parameter is set at 0 (the null condition), the model gives a type 1 error rate since there is no effect to detect. In other conditions, the results show power to detect the effects. The results showed that in the conditions

with CL2=0, the two-level model and the complex model gave a nominal type 1 error rate

(around 0.05), whereas the default model gave an upwardly biased type 1 error rate with

noticeable spikes in small sample size conditions with PVB=0 and R2B=0.5. This indicates that

power is biased for the default model since it is based on an inflated type 1 error rate. The results

also showed that the sample size did not affect variability in type 1 error rates. Then, in the

conditions with CL2=0.1, the sample size had its influence on the results in that as the sample

size increases the power also showed increasing patterns with varying degrees depending on the

other conditions. The default model was overpowered in detecting the effects when compared to

the complex model. The two-level model showed a decent level of power for detecting the

effects in general although the results were not given in small sample size conditions. In the

conditions with CL2=0.3, the default model was overpowered in detecting the effects when

compared to the complex model in small sample size conditions. In middle and large sample size

conditions, all the models showed as high power as 1 across all other conditions.

Figure 28. *Empirical power/ type 1 error rate of CL2 parameter*
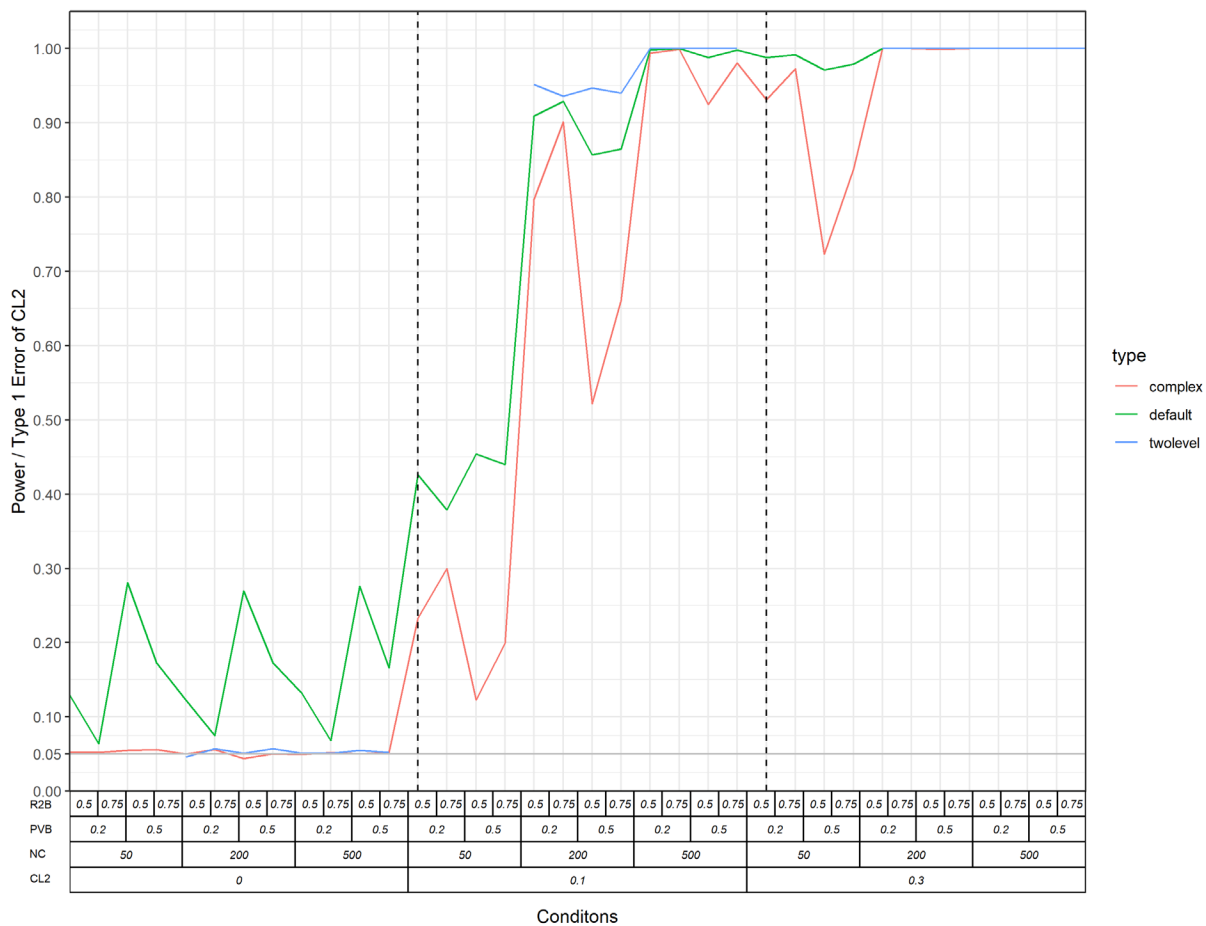
Overall, there was a difference found in results between the models in certain conditions. Specifically, standard errors were overestimated for the complex model and underestimated for the default model when compared to the two-level model, which might lead to different conclusions in statistical inference between the models. The size of SE bias for the complex model was small whereas that for the default model was large. SE bias for the complex model was less severe whereas SE bias for the default model was more severe when there is a larger percentage of variance at the between-group level. Finally, in terms of power and type 1 error rates for the cross-lagged parameter, it was found that if the effect size is zero or close to zero the sample size did not affect the results and only variance conditions affect the results. It was found to be the conditions with the small effect size when the sample size matters. If the effect size is large enough, then the sample size again had no impact on the results and all the models showed a high level of power whether they were overpowered or underpowered.

CHAPTER 4. STUDY 2

In Study 2, an actual data example was analyzed to demonstrate and compare the performance of selected models. The models were then compared in terms of model fits, the statistical significance of individual parameter estimates under given alpha=0.05, and consequent interpretation of the results. A description of the data and analysis method is introduced first, followed by the results.

The focus of Study 2 is on the bivariate processes of executive function and elementary school children's progress in academic achievement. The relationship between executive function and academic achievement has been widely studied across various populations. Several meta-analyses synthesizing individual studies indicated that although in varying degrees depending on gender, race, ethnicity, and age groups, there is a positive relationship between children's executive functioning and their academic achievement in terms of overall association (Allan, Hume, Allan, Farrington, & Lonigan, 2014; Jacob & Parkinson, 2015). In addition, when looking at the within-person bidirectional associations between children's executive functioning and learning outcomes directly, there was consistent evidence for bidirectional associations between executive functioning and learning outcomes (Bohlmann, Maier, & Palacios, 2015; Connor et al., 2016; Daneri & Blair, 2017; Fuhs, Nesbitt, Farran, & Dong, 2014; Weiland, Barata, & Yoshikawa, 2014; Welsh, Nix, Blair, Bierman, & Nel- son, 2010). In 2019, Willoughby, Wylie, and Little looked at the between and within-person associations between 2 types of executive functions (working memory, cognitive flexibility) and 2 types of academic achievement (math, reading) by applying the LCM-SR on ECLS-K:2011 data (Willoughby, Wylie, and Little, 2019). In their research, they have found a strong between-person association between executive functioning and achievement and a weak to a nil within-person, time-varying

association. In their study, the multilevel aspect of the data has been addressed using the single-level LCM-SR with the robust maximum likelihood estimation method. In Study 2, the data were re-analyzed using the MLCM-SR, along with the alternative modeling strategies for comparison. According to the results of Study 1, when the size of cross-lagged effects is small (less than 0.1), the default model would show overpower to detect the effects under certain conditions. Since in the original study the default model could not detect the effect most of the time, there are chances that the two-level and complex models would not show the difference as well. However, even though a sample was drawn from the same dataset and the same variables were analyzed using the same model, Willoughby, Wylie, and Little (2019) and Study 2 will differ in terms of the target population, sample sizes, years of assessment, and analytic focus and procedure. Therefore, in Study 2, it is hypothesized that depending on the level of design factors in Study 1, the selected models will show a difference in terms of model fits, the statistical significance of individual parameter estimates under given alpha=0.05, and consequent interpretation of the results.

Method

Illustrative Data

Data for Study 2 were derived from the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K:2011; Tourangeau et al., 2019). The ECLS-K:2011 is the third and latest study in the Early Childhood Longitudinal Study (ECLS) program, which provides descriptive information on children's development and early learning experience, and their school progress. The longitudinal design of the study, multiple sources of information, and the comprehensive set of data collection instruments enable researchers to study a wide variety

of research questions about how child, home, school, and neighborhood factors relate to children's cognitive, social, emotional, and physical development.

In the ECLS-K:2011, a nationally representative sample of 18,170 children was selected from about 1310 (public and private) schools attending both full-day and part-day kindergarten in 2010-11 and then followed through the 2015–16 school year, where most of the children are expected to be in fifth grade. Currently, the data file from the beginning of kindergarten through the end of fifth grade are publicly available online (https://nces.ed.gov/ecls/kindergarten2011.asp).

In the ECLS-K:2011, a clustered, multi-stage sampling design and procedure, which involves sampling primary sampling units (PSUs; which is counties) and schools with probabilities proportional to the number of children, and selecting a fixed number of children per school, was employed (For a detailed description of the sampling design, see Tourangeau et al. 2015). The sample design is highly related to the dispersion of the children in the sample, and the user's manual directly warns that if statistical analyses are conducted with the assumption of simple random sampling for collected data, then the calculated standard errors will be incorrect.

Study 2 evaluated children's executive functioning and their reading achievement each year from 1st through 5th grade. A measure of reading achievement was collected for each grade through a direct cognitive assessment of children's reading skills. The reading test scores were then calculated using item response theory (IRT) scoring procedures, where scores at different assessments (different in time and assessment tools) are computed on the same scale ("vertically scaled" or "linked") so they can be used in longitudinal analysis. Interested readers in IRT and equating methods of assessment scales are referred to De Ayala (2009) and Kolen and Brennan

(2014), respectively. The reliability of reading achievement IRT scores was 0.86 – 0.95 across the data collection period (Spring 2013 – Spring 2016). In addition, A measure of executive function was also collected for each grade through a direct child assessment battery, in which three subtests measured cognitive flexibility, working memory, and inhibitory control, respectively. In this study, the Numbers Reversed subtest of the Woodcock-Johnson III (WJ III) Tests of Cognitive Abilities, which measures children's working memory, was used (Woodcock, McGrew, and Mather 2001). This is because it is the only subtest that was administered uniformly across all rounds of collection, kindergarten through fifth grade. In addition, for the working memory scores, W-ability score, a special transformation of the Rasch ability scale which provides a common scale of equal intervals and represents both a child's ability and the task difficulty, is available. W score is particularly useful for the measurement of growth and can be considered as a growth scale. Its mean is reflective of the average performance for 10-year-old children (Tourangeau et al., 2015). Higher scores for both variables reflect higher ability in reading skills or working memory, respectively.

The bivariate correlations of repeated measures for both variables are given in Table 2. It is shown that there was a stronger relationship found among repeated measures within each variable rather than across variables. In addition, longitudinal trends are plotted for a random sample of n=200 participants in Figure 29 and Figure 30. A large variance was observed for both scores across measurement occasions. There was a smooth increasing trend on average across occasions for reading achievement and working memory scores.

In this study, schools were chosen as a cluster variable. In educational research, schools are usually considered as the third level in the hierarchical data structure where children are nested within teachers, and teachers are then nested within schools (Raudenbush and Bryk 2002).

However, since only two levels are considered in this study and teachers usually change every year while children tend to stay in the same school across years, schools were included for analysis. All other sampling units and sampling weights which make the sample nationally representative were ignored. This is because analyses done in this study were for illustrative purposes and not for actual dissemination of generalizable findings of the population.

In this study, since data missingness is not of interest, only complete cases were included for analysis. For the same reason, cases with cross-classified cluster structures (e.g., children who attended more than one school during the given period due to transfer) were also excluded. Finally, only cases with a cluster size greater than 5 for estimation accuracy were selected (Maas & Hox, 2005; McNeish & Stapleton, 2016). The resulting sample size is 6614 clustered within 652 schools.

Table 2. *Correlation table for total eligible children (N=6414) on reading achievement and working memory scores*

	1	2	3	4	5	6	7	8	9	10
T1 WM	1									
T2 WM	.497**	1								
T3 WM	.456**	.554**	1							
T4 WM	.427**	.517**	.599**	1						
T5 WM	.425**	.501**	.582**	.642**	1					
T1 reading	.501**	.431**	.446**	.441**	.465**	1				
T2 reading	.483**	.449**	.448**	.439**	.463**	.848**	1			
T3 reading	.481**	.434**	.442**	.428**	.433**	.764**	.834**	1		
T4 reading	.471**	.445**	.446**	.455**	.473**	.773**	.833**	.834**	1	
T5 reading	.464**	.424**	.427**	.432**	.462**	.727**	.792**	.812**	.848**	1

*** Correlation is significant at the 0.01 level (2-tailed).

Figure 29. *Longitudinal plots for a random sample of n=200 children on working memory scores*
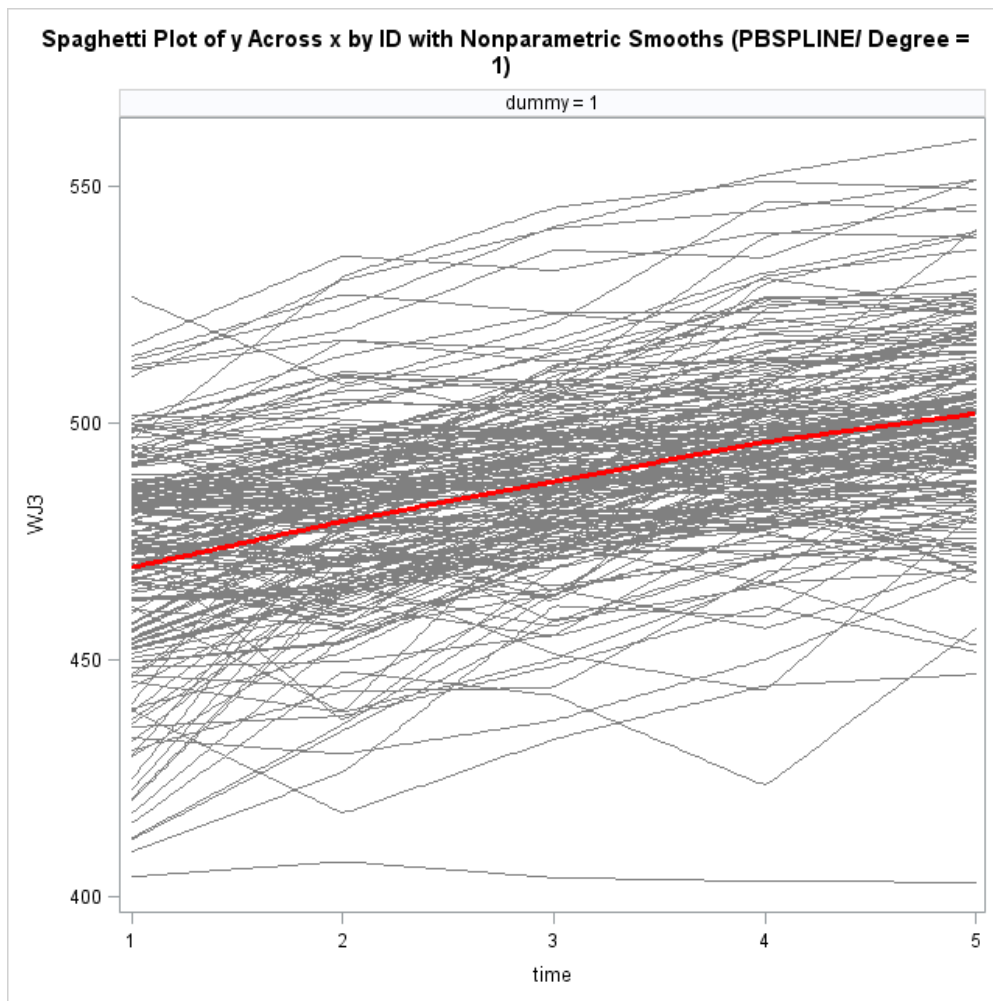
Figure 30. *Longitudinal plots for a random sample of n=200 children on reading achievement scores*

Spaghetti Plot of y Across x by ID with Nonparametric Smooths (PBSPLINE/ Degree = 1)
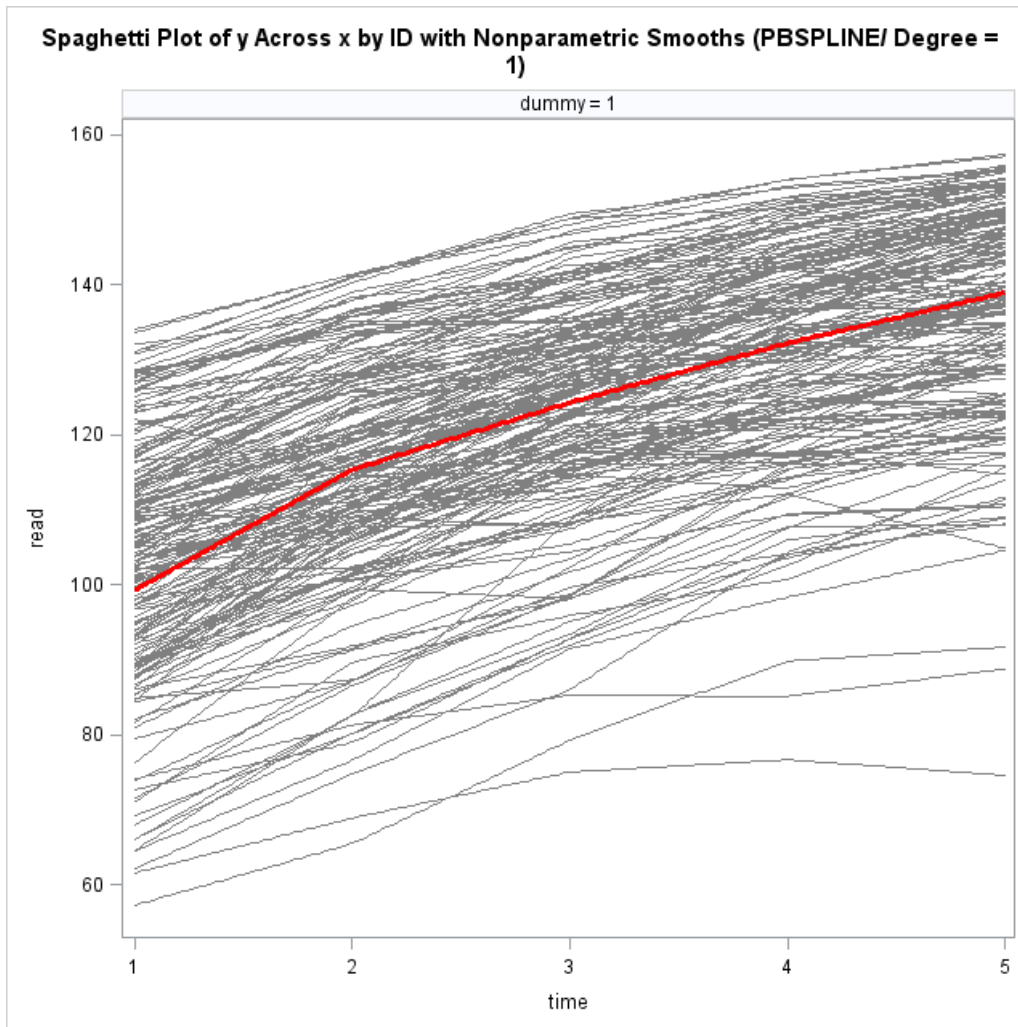
Analytic procedure

Curran et al (2014) gave a general model-building strategy for LCM-SR, which consists of two big steps and several smaller steps. First, the optimally fitting models for each construct were established separately. More specifically, two univariate latent curve models were estimated for each outcome to find the optimal function of time for each construct. Next, autoregressive effects among the structured residuals are introduced to the models without constraints. Then, the equality constraints on the autoregressive effects are imposed. Second, bivariate latent curve models were estimated for both constructs simultaneously. Next, latent

curve factors and time-specific residuals at the same time points are allowed to covary with one another, and those covariances are set to be equal across time except for the first time points. Finally, autoregressive components among the time-residuals are introduced followed by cross-lagged components. At each step, adding components and imposing equality constraints on each component are tested via LRT to decide whether or not to proceed further with those added components or constraints.

Yet, in this study, the model-building strategy differed from the suggested one for the following reasons. First, this study was done for illustrative purposes. Since the focus of the study was in comparing models in terms of SE bias resulting from ignoring the dependency among observations, retaining the same growth factor structure and residual structure across the models was desired. Therefore, all the parameters in the LCM-SR were introduced and retained, whether the model fit significantly improved or not, as long as the result gave a proper solution. In fact, Curran et al. (2014) also made similar decisions when building a model for illustration in their study.

Second, in line with Study 1, one of the main features of the LCM-SR is its ability in making causal claims, and the LCM-SR demands of stationarity assumption are met for all the components. The focus of this study is comparing a difference among the models in making causal inference. Therefore, equality constraints on the parameter were imposed whether the model fit gets significantly degraded or not, as long as the result gave a proper solution. In addition, the given strategy itself did not consider the multilevel context. Thus, the percentage of variance was checked first to make sure that applying a two-level model to the data is appropriate.

After the model building process, estimation results of final models are then compared between models in terms of fit indices, parameter estimates, SE, and p-values. More specifically, model fit indices were evaluated to gauge which model appears to be doing the best at fitting the data. In addition, the cross-lag parameters were also examined to see (1) which process is indicated as being causally dominant and (2) what is the size of these effects concerning one another. The between-person growth factor components (LCM part), especially the size of the variance which implies the extent to which individuals are different from one another in trait-like stability or inter-individual differences in change, were also examined. Finally, the covariance in the between-person growth factor components was also examined. All models were fit using Mplus 8.5 (Muthén & Muthén, 1998-2017).

Results

First, the percentage of variance in initial status (intercept) and rate of change (slope) that lies between schools and R-square of repeated measures were checked for both variables by fitting univariate two-level LCM to the data. This was done to measure data dependency within the same cluster in the data and determine whether using a multilevel model is appropriate. For the reading achievement variable, a linear slope model was found to be the optimally fitting model. For this variable, since the result gave an improper solution and the slope factor variance was not significant at the between-group level, it was fixed to 0 to proceed with a proper solution. Therefore, the proportion of variance in the intercept that lies between schools was $25.94/242.08 \cong 11\%$ whereas the slope's corresponding proportion of variance was $0/9.26 = 0$. For the working memory variable, a linear slope model was found to be the optimally fitting model. For this variable, the result gave an improper solution, and the slope factor variance was not significant at the between-group level. Thus, the slope variance was fixed to 0. The

proportion of variance in the intercept that lies between schools was 44.17/194.42 ≅ 19% whereas the slope's corresponding proportion of variance was 0/3.37 = 0. R-square for each of the repeated measures at the between-group level was 0.54-0.93 for T1-T5 repeated measures of working memory and 0.25-0.98 for T1-T5 repeated measures of reading achievement. Note that R-square was low for T1 repeated measures only and high as 0.8 for all other time points. R-square for each of the repeated measures at the within-group level were 0.45-0.67 for working memory and 0.81-0.89 for reading achievement. Overall, there is some level of variability in the initial status that lies between schools, while there was no slope variance nested for both variables. Therefore, the two-level model structure was retained for further analysis. Note that the level of variability is lower than Study 1 conditions.

For the next step, two univariate LCMs were combined into a two-level bivariate LCM. Then, residual structure components were added to the model to form the full two-level LCM-SR. The model fit information at each step is given below in Table 3.

Table 3. *Model fit information of the two-level model at each step of the model building process*

	Bivariate LCM	T1-T5 residual covariance added	AR/CL added	Residual/ innovation covariance equality	Residual/ innovation variance equality
# Parameters	37	47	51	45	33
H0 LL	-260494	-260353	-260286	-260290	-260503
H1 LL	-259323	-259323	-259323	-259323	-259323
AIC	521062.5	520800.7	520673.2	520670.5	521072.9
Δ AIC		-261.825	-127.459	-2.676	402.326
BIC	521314	521120.1	521019.9	520976.3	521297.2
Δ BIC		-193.856	-100.271	-43.51	320.815
ChiSqM_DF	83	73	69	75	80
Δ ChiSqM_DF		-10	-4	6	5
ChiSqM	2322.464	2054.529	1936.484	1932.456	2275.954
Δ ChiSqM		-267.935	-118.045	-4.028	343.498
RMSEA	0.064	0.064	0.064	0.061	0.062

Δ RMSEA			0	0	-0.003	0.001
CFI	0.954	0.959	0.961	0.962	0.955
Δ CFI		0.005	0.002	0.001	-0.007
TLI	0.95	0.95	0.95	0.954	0.953
Δ TLI		0	0	0.004	-0.001
SRMRW	0.027	0.025	0.024	0.025	0.035
SRMRB	0.26	0.255	0.285	0.287	0.337

Note that the LCM in the first column refers to the LCM whose residuals are modeled as single-indicator latent variables when using the ML estimation method in the Mplus framework (Asparouhov & Muthén, 2021). Also, note that T2- T5 residual covariances turned into innovation covariances when the AR/CL components are added at the third step. The results showed that model fits improved slightly as residual structure components are added and even constrained to be equal. However, at the fifth step where equality constraints are imposed on the innovation variances, the model fits got degraded, which is a sign of misfit in the model. The level-specific SRMR indicated that the source of misfit came from the between-group level. As discussed previously, the residual equality constraints were to be retained. Even though the final model was not the best fitting model, it was retained as-is for model comparison. The Chi-square goodness of fit tests was significant across all steps since they are a direct function of sample size, in which even a small misfit can be detected (Jöreskog, 1969). Other model fit indices indicated that the final model is a good fit for the data.

The parameter estimates for the final model are also given below in Table 4. The results showed significant autoregressive processes within each variable (AR1 estimate = 0.087, $p < .05$; AR2 estimate = 0.086, $p < .05$) but no significant cross-lagged processes across the variables (CL1 estimate = 0.028, $p > .05$; CL2 estimate = 0.006, $p > .05$). As for the rest of the parameters, most variance components were significant, which leaves room for exploration. For example, there was

a large amount of variance found in the intercept at both the within and between-group levels for both variables. This could be explained by time-invariant, child-level covariates, and time-invariant, school-level covariates. At the same time, there was a large amount of variance found in the residuals and innovations. which could be explained by time-variant, child-level covariates and time-variant, school-level covariates.

Table 4. *Model estimation results for the two-level model*

	Estimate	S.E.	Est./S.E.	P-Value	Estimate	S.E.	Est./S.E.	P-Value
AR/CL								
AR1 (XE_t <- XE_t-1)	0.087	0.015	5.908	0				
CL1 (XE_t <- YE_t-1)	0.028	0.019	1.46	0.144				
AR2 (YE_t <- YE_t-1)	0.086	0.01	8.751	0				
CL2 (YE_t <- XE_t-1)	0.006	0.004	1.446	0.148				
	WITHIN				**BETWEEN**			
Growth Factor Structure	Estimate	S.E.	Est./S.E.	P-Value	Estimate	S.E.	Est./S.E.	P-Value
Mean								
X Intercept					476.279	0.434	1096.794	0
X Slope					7.204	0.111	64.753	0
Y Intercept					106.001	0.336	315.225	0
Y Slope					7.92	0.059	134.661	0
Variance								
X Intercept	220.807	10.796	20.452	0	24.797	2.963	8.37	0
X Slope	4.51	1.016	4.439	0	0	0	999	999
Y Intercept	189.019	4.271	44.253	0	33.053	2.789	11.849	0
Y Slope	2.624	0.175	14.956	0	0	0	999	999
Covariance								
X Intercept with X Slope	-6.7	2.634	-2.543	0.011	0	0	999	999
X Intercept with Y Intercept	136.964	5.153	26.578	0	25.74	2.526	10.191	0
X Intercept with Y Slope	-5.279	0.869	-6.078	0	0	0	999	999
X Slope with Y Intercept	-2.741	1.226	-2.237	0.025	0	0	999	999
X Slope with Y Slope	-0.236	0.273	-0.865	0.387	0	0	999	999
Y Intercept with Y Slope	-11.485	0.718	-15.998	0	0	0	999	999
Residual structure								
Variance								
T1 Residual for X	335.573	11.065	30.327	0	39.662	5.536	7.164	0
T1 Residual for Y	53.294	1.801	29.589	0	111.52	4.828	23.1	0
T2-T5 Innovation for X	193.366	4.611	41.934	0	3.448	0.798	4.322	0
T2-T5 Innovation for Y	35.295	0.583	60.542	0	1.404	0.184	7.649	0

Covariance

T1 Residual Covariance	15.951	2.541	6.276	0	58.736	4.982	11.79	0
T2-T5 Innovation Cov.	3.514	0.886	3.968	0	0.335	0.281	1.193	0.233

The model fit information for the complex and default models is given in Table 5. Since some of the model fits which are not a function of Chi-square are the same for both models, they are shown only once at the top of the table.

Table 5. *Model fit information of the complex and default models at each step of the model building process*

	Bivariate LCM	T1-T5 residual covariance added	AR/CL added	T2-T5 innovation covariance equality	T2-T5 innovation variance equality
# Parameters	24	29	32*	29	23
H0 LL	-262592	-262444	-262377	-262388	-262640
H1 LL	-259828	-259828	-259828	-259828	-259828
AIC	525232.8	524945.4	524817.3	524833.9	525326.9
Δ AIC		-287.35	-128.116	16.557	493.015
BIC	525395.9	525142.5	525034.8	525031	525483.2
Δ BIC		-253.366	-107.724	-3.835	452.234
SRMR	0.084	0.077	0.092	0.097	0.181
COMPLEX					
ChiSqM_DF	41	36	33	36	42
ChiSqM	5011.897	4724.956	4626.742	4646.011	4901.241
Δ ChiSqM		-286.941	-	19.269	255.23
RMSEA	0.135	0.14	0.145	0.139	0.132
Δ RMSEA		0.005	-	-0.006	-0.007
CFI	0.909	0.914	0.916	0.916	0.911
Δ CFI		0.005	-	0	-0.005
TLI	0.9	0.893	0.885	0.895	0.905
Δ TLI		-0.007	-	0.01	0.01
DEFAULT					
ChiSqM_DF	41	36	33	36	42
ChiSqM	5528.176	5230.826	5096.711	5119.267	5624.282
Δ ChiSqM		-297.35	-	22.556	505.015
RMSEA	0.142	0.148	0.152	0.146	0.142
Δ RMSEA		0.006	-	-0.006	-0.004
CFI	0.894	0.9	0.902	0.902	0.892

Δ CFI		0.006	-	0	-0.01
TLI	0.884	0.875	0.867	0.878	0.885
Δ TLI		-0.009	-	0.011	0.007

Note. When only equality constraints are imposed the number of free parameters should be 29+4=33. Yet since the constraint was imposed on the slope variance of Y, the number of free parameters is now 33-1 = 32.

The results showed that model fits improved slightly as the residual structure components are added at the 2nd step. Yet at the third step where the AR/CL components are added to the model, the model gave an improper solution. As discussed previously, the residual structure components and the constraints were to be retained. Since information about the slope variance is given in the prior steps, it was fixed to the given value, which is 2.1. Then the model gave a proper solution. Since additional constraints are added on the third step, the models at the second and third steps are not nested. Even further, the two-level model as well as the other models are not nested either. Yet, the difference in AIC and BIC showed that there was an improvement in the model results. At the fourth and the fifth step, AIC and Chi-square values were degraded.

However, at the fifth step where equality constraints are imposed on the innovation variances, the model fits got degraded, which is a sign of misfit in the model. Even though the final models were not the best fitting models, they were retained as-is for model comparisons. In terms of model fits, when compared between the complex and default models, the complex model is better fitting than its default counterpart. Yet in absolute terms, both models did not fit the data well (for the complex model, $\chi2(42) = 4901.24$, $p<.05$; RMSEA = .13, CFI=.91, TLI=.91; for the default model, $\chi2(42) = 5624.28$, $p<.05$; RMSEA = .14, CFI=.89, TLI=.89). When both models are compared to the two-level model, the two-level model was found to be better fitting in terms of AIC and BIC (for the two-level model, AIC= 521072.9, BIC= 521297.2; for the complex and default model, AIC= 525326.9, BIC= 525483.2; for AIC and

BIC, lower values indicate better fitting). However, considering the difference in the number of free parameters, the difference in the fit values observed here are not so meaningful.

The parameter estimates for the complex and default models are given below in Table 6. For comparison purposes, the aggregated parameter estimates for the two-level model are provided as well. Covariance parameters were not included in the table because they are not comparable to each other. The results showed that in terms of the AR/CL parameter estimates, the three models were not very different. The implicit assumption of the model parameterization was that the nature of the AR/CL process is the same across the levels of nesting. However, as shown in the result, the amount of R-square for repeated measures at the within-group and between-group levels is not necessarily the same, which could give different AR/CL parameter estimates at both levels. This could have introduced a difference in the AR/CL parameter estimates between the models but there was only a small difference in this example. The estimates for the CL terms were small across the models and they were all not significant (i.e., the data did not show evidence for any causal effects). There was a difference in other parameter estimates among the models, which may be due to the constraint on the slope variance for the complex and default models. There were differences in the estimated standard errors between the models in that the standard errors were overestimated for the complex model and underestimated for the default model when compared to the two-level model. However, the size of the differences was small. Therefore, it is deemed that the conclusions drawn from the models would not differ between the models.

Table 6. *Model estimation results for the two-level, complex, and default models*

	Two-level				Complex				Default			
	Estimate	S.E.	Est./S.E.	P-Value	Estimate	S.E.	Est./S.E.	P-Value	Estimate	S.E.	Est./S.E.	P-Value
AR/CL												
AR1 (XE_t <- XE_t-1)	0.087	0.015	5.908	0	0.082	0.015	5.556	0	0.082	0.013	6.401	0
CL1 (XE_t <- YE_t-1)	0.028	0.019	1.46	0.144	0.015	0.019	0.77	0.442	0.015	0.019	0.8	0.424
AR2 (YE_t <- YE_t-1)	0.086	0.01	8.751	0	0.063	0.008	7.466	0	0.063	0.008	8.34	0
CL2 (YE_t <- XE_t-1)	0.006	0.004	1.446	0.148	-0.001	0.005	-0.106	0.915	-0.001	0.005	-0.117	0.907
Growth Factor Structure												
Mean												
X Intercept	476.279	0.434	1096.794	0	475.048	0.38	1251.721	0	475.048	0.286	1659.309	0
X Slope	7.204	0.111	64.753	0	7.664	0.092	82.862	0	7.664	0.081	94.685	0
Y Intercept	106.001	0.336	315.225	0	103.042	0.355	290.345	0	103.042	0.213	484.665	0
Y Slope	7.92	0.059	134.661	0	8.968	0.054	167.307	0	8.968	0.042	215.379	0
Variance												
X Intercept	245.604	11.28	21.773	0	253.741	12.007	21.132	0	253.741	9.76	25.999	0
X Slope	4.51	1.016	4.439	0	4.965	1.036	4.793	0	4.965	0.925	5.365	0
Y Intercept	222.072	4.915	45.181	0	224.769	5.44	41.314	0	224.769	4.505	49.893	0
Y Slope	2.624	0.175	14.956	0	2.1	0	999	999	2.1	0	999	999
Residual structure												
Variance												
T1 Residual for X	375.235	13.83	27.132	0	360.014	12.342	29.169	0	360.014	9.296	38.728	0
T1 Residual for Y	164.814	5.558	29.654	0	123.349	3.824	32.257	0	123.349	2.946	41.87	0
T2-T5 Innovation for X	196.814	4.663	42.209	0	196.276	4.704	41.723	0	196.276	3.219	60.976	0
T2-T5 Innovation for Y	36.699	0.58	63.302	0	38.76	0.601	64.543	0	38.76	0.488	79.429	0

CHAPTER 5. DISCUSSION

The purpose of this study is to explore the impact of dependency among observations on the results when using the LCM-SR, and how to appropriately analyze the clustered longitudinal data for more accurate inference. To do this, the MLCM-SR (disaggregated approach) was introduced and compared with the single level LCM-SR considering nesting effects (aggregated approach), and the single level LCM-SR ignoring nesting effects (conventional approach). The discussion follows the original research questions. Then, limitations, future directions, and implications are presented.

RQ1: How would the MLCM-SR model performance fare when compared to alternative modeling strategies?

All the models showed high rates of non-convergence or improper solutions in certain conditions, especially in low sample size conditions, even though covariance constraints are added after the equality constraints on residual components across time. The low number of proper solutions from the models is somewhat expected as reported by Orth et al. (2021) who analyzed 10 samples of actual data (sample size varied from 404 to 8259) from longitudinal studies and found that in most datasets, even the single level LCM-SR did not converge at all or did not converge properly.

There was a difference between the two-level and complex/default models in terms of non-convergence rates and improper solution rates. Specifically, a higher rate of non-convergence was observed from the complex/default model results whereas a higher rate of improper solutions was observed from the two-level model. Yet, the total number of proper solutions was higher for the complex/default model than for the two-level model in general. This

is also expected in that the two-level model has a more complex model structure than the complex/default model (Bates et al., 2015).

Asparouhov and Muthén (2021) pointed out that the high rate of non-convergence/improper solutions of the model is due to empirical under-identification (Newsom, 2012; Kenny & Milan, 2012). Therefore, a series of additional analyses using a large scale (nc=1000; N=5000) simulated data were run to check if there was any specific issue with pre-specified population parameters in data generation models with varying conditions (e.g, size of CL parameters and variance parameters). The results showed that, when all the constraints were removed, a high rate of improper solutions were observed even with large datasets when CL parameters were high and innovation variance parameters were below 1. However, when the equality constraints were imposed, all the solutions converged properly. Therefore, a high rate of improper solutions was more likely due to sampling variability (i.e., the data do not support the complex model structure).

In another experiment, additional waves of data were added to the model following Asparouhov's and Muthén's suggestion (for example, T=8). The results using the large-scale datasets showed improvement in that most replications converged properly. However, in small sample size conditions, the problem got worse in that the model structure got more complex with additional latent variables to address (recall that residual components were modeled as latent variables when using ML estimation method in the Mplus framework; Asparouhov and Muthén, 2021). In the following experiment, the models were also run without residual structure at the between-group level, following the multilevel modeling tradition. The results showed similar improvement in the large sample conditions (N=5000) as well as in the small sample condition (N=250). The difference in the total number of proper solutions between the models was not held

constant but varied across conditions. A low number of proper solutions in the small sample condition (nc=50) for the two-level model resulted in its exclusion from further analysis.

There was also a difference between the two-level and complex/default models in terms of model fit, bias, coverage, power, and type 1 error rates. In general, bad model fit, severe bias, low coverage rate, and low power were found in conditions with a large percentage of variance as well as a large residual variance at the between-group level (Diallo & Lu, 2017). The severity of bias increased as the sample size decreased. The default model showed the highest level of bias followed by the complex model. The direction of bias was also different between the two models in that the complex model showed a small upward bias whereas the default model showed a large downward bias. This difference resulted in the inflated power and type 1 error rate for the default model and the deflated power and nominal level type 1 error rate for the complex model. The two-level model showed little or no bias in general, thus showing a decent level of power and a nominal level of type 1 error rate. The results imply that the two-level model should be used in presence of dependency between observations in the clustered data. This suggestion should be considered with caveats regarding the sample size since the results for the two-level model were not discussed in small sample size conditions. In any case, the complex model should be considered over the default model for its more accurate results when compared to the latter.

RQ2: Do different modeling strategies lead to different conclusions when dependency among observations is present in the dataset?

In Study 1, it was shown that there would be a difference in the results between the models in that standard errors were overestimated for the complex model and underestimated for the default model when compared to the two-level model, which leads to different conclusions

between the models. Yet in Study 2, even though there was a difference in the standard errors found between the models, using different modeling strategies did not lead to different conclusions. This is somewhat similar to the results from Bailey et al. (2020) where the researchers tested for the nesting effects at the school level in the reciprocal relationship between reading and math achievement in the ECLS:K 2011 dataset and found no effect. The results could be explained as follows under the conjecture that the design factors and the outcomes in Study 1 have a linear relationship.

First, the sample size was large (N= 6614) to offer enough power for all the models to detect any effects, as long as they exist. Study 1 only discussed that relative bias is maintained across different sample size conditions and overlooked that the actual magnitude of bias decreases as the sample size increases. Therefore, even though there were differences in the standard errors between the models, since the size of the differences was small, it did not lead to different conclusions. Second, there was a small percentage of variance at the between-level that lies between schools. Study 1 tested extreme conditions where the percentage of variance at the between-level is high as 0.5 and most variability in the results came from the conditions with the largest percentage of variance at between-group level conditions. The sample size was again large enough to ignore the difference in the standard errors. Third, with specific regards to the AR/CL parameters, the effect size was small to show a difference between the models in standard errors as well as point estimates.

Limitation and future directions

The present study has several limitations and opportunities for future research. First, the dataset in Study 2 was chosen because the purpose of the study was in part to revisit the former study results with the proposed modeling approach (Willoughby, Wylie, and Little, 2019) and

because the data meet the implicit inclusion criteria (e.g., the dataset should provide information about early child development with more than 4 waves of data with an equal time interval between repeated measures. It should be appropriately scaled for longitudinal modeling). Yet, the results showed that the data characteristics, including the sample size, were not within the range of values that were tested in Study 1. This in turn made the results from Study 1 less informative in which the interpretation was largely based on conjectures. In Study 1, the relatively high values in design factor conditions and relatively small sample sizes were chosen to obtain a clear insight into the effect of variation in design factors. Therefore, further exploration with adjusted conditions in a simulation study is needed.

Second, in this study, only the simplest form of the model was considered, which is the linear growth model with AR(1) & CL(1) processes. One can extend the model to allow for different growth forms (e.g., quadratic growth) or higher-order AR/CL processes to best fit the data in future research. Third, in this study, no time-varying or time-invariant predictors were considered at both levels of the model. According to the Rubin Causal Model, to draw accurate estimates of causal effects, all the potential confounders must be accounted for (Usami, Murayama, Hamaker, 2019; Rubin, 1974). Yet in this study, it was assumed that there are neither observed nor unobserved confounders in the process. Therefore, in future studies, researchers might want to consider the impact of omitting confounders at the between-group level on the AR/CL parameter estimates. Fourth, in this study, only one clustering variable (i.e., school) was considered in the model. However, in reality, there are multiple sources of clustering (e.g., family, teacher, or community). Therefore, future studies might consider integrating higher-level or cross-classification model structures into the model. Fifth, all the variables analyzed in this study were considered to be continuous variables with multivariate normal distributions. In

future research, the behavior of the model under the non-optimal conditions for the data (e.g., non-normality, data missingness) should also be considered. Sixth, in this study, only 4 waves were considered for data simulation based on the fact that many panel datasets have at most 4 waves of data (Orth et al., 2021). Yet, this posits a large limitation in the generalizability of the results. In future research, incorporating more than one length of waves and more than 4 repeated measures in the simulation design would be desired. Finally, discrete-time variable and equal spacing of observed data were assumed based on the fact that the effect size of the AR/CL parameters are influenced by the interval between the observations ("the lag-problem"; Usami, Murayama, Hamaker, 2019; Gollob and Reichardt, 1987). Also, equality constraints are imposed on all residual structure components for causal inference. However, at cost of claiming causal inference, researchers can relax these assumptions and turn to a continuous time modeling approach and use unequally spaced repeated measures data (Deboeck & Preacher, 2016; Voelkle; Oud, Davidov, & Schmidt, 2012).

Implications for practitioners

If a researcher has large-scale longitudinal data with more than 4 waves, then the any model would work fine depending on the study purposes. For individual researchers, however, it is hard to collect a large-scale longitudinal data like the one used in this study. Based on the study results and discussion, the following implications on modeling approaches to addressing observation dependency when using the LCM-SR are outlined. When the clustered data have a small number of clusters such as 50 or when clustering effects are only considered as a nuisance, the complex model should be used to account for data dependency. Yet, if one is interested in between-group level time-invariant predictors (e.g., school district), then one can opt for the two-level LCM-SR without residual structure at the between-group level, following the multilevel

modeling convention. One can simply fix the residual variance and covariance at zero, which should allow the model estimation results to be more reliable and accurate unless there is a large between-level residual variance (Diallo & Lu, 2017). If one is interested in between-group level time-varying predictors (e.g., school-level policy change), then one might want to fix the latent variable variance structure to the numbers from the LCM, as shown in Study 2.

CHAPTER 6. CONCLUSION

The present study explored the impact of dependency among observations on the results when using the LCM-SR and options to appropriately analyze the clustered longitudinal data for more accurate inference. The LCM-SR and its multilevel extensions have the potential for discovering valuable information in developmental and educational research. One example is the relationship between children's psychological development and their interaction with the environment.

Children's psychological development has been understood based on contextual/ ecological system theories, which posit that development is shaped by environmental factors within multiple layers of ecological systems (Bronfenbrenner, 1977; Pianta & Walsh, 1996; Rimm-Kaufman & Pianta, 2000; Bronfenbrenner & Morris, 2006). For example, during early childhood, parents help children build and refine their knowledge and skills, serving as a foundation for children's well-being and healthy development. After entering school, teachers and peers play a critical role as learning environments in promoting children's cognitive, language, social, and emotional development.

Yet, the relationship between children and their environment is not just a unidirectional flow from an environment to a child – it is more like a reciprocal or bidirectional interaction. In their bioecological theory, Bronfenbrenner & Morris said, "(e)specially in its early phases, … human development takes place through processes of progressively more complex reciprocal interaction between an active, evolving biopsychological human organism and the persons, objects, and symbols in its immediate external environment" (Bronfenbrenner & Morris, 2006, p.797). As for interpersonal interaction, this means that initiatives do not come from one side only but have some degree of reciprocity in the exchange. For instance, in parent-child

interaction, children can make parents feel happy, sad, fulfilled, or angry, and then parents can give adequate feedback to their children as well.

In sum, children's psychological development occurs through a variety of interaction processes in their immediate environment such as interaction with their parents, caregivers, relatives, siblings, peers, and so on. Their influence in the development, which yet might change in nature or strength, is not limited to the formative years but continue to be effective as children grow older (Bronfenbrenner & Morris, 2006). As a result, both characteristics of children themselves and characteristics of their environment are of substantive interest by researchers. To study and capture such interaction dynamics over time that are consistent with developmental theories, and to study the predictive or causal relationship between them, the LCM-SR can be implemented as one of the options. And whenever clustering structure is involved in child data, the use of the modeling approaches discussed here is recommended.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52(3),* 317–332.

    https://doi.org/10.1007/BF02294359

Alavi, M., Visentin, D. C., Thapa, D. K., Hunt, G. E., Watson, R., & Cleary, M. (2020). Chi-

    square for model fit in confirmatory factor analysis. *Journal of Advanced Nursing, 76(9),*

    2209–2211. https://doi.org/10.1111/jan.14399

Anumendem, D. N. (2011). Advanced techniques in multilevel growth curve modeling.

    Application to educational effectiveness research. Doctoral Dissertation, Katholieke

    Universiteit Leuven. (Unpublished).

Asparouhov, T. & Muthén, B. (2005). Multivariate statistical modeling with survey data.

Asparouhov, T. & Muthén, B. (2010). Resampling methods in Mplus for complex survey data.

Asparouhov, T. & Muthén, B. (2018). SRMR in Mplus.

Asparouhov, T. & Muthén, B. (2021). Residual Structural Equation Models.

Bailey, D. H., Oh, Y., Farkas, G., Morgan, P., & Hillemeier, M. (2020). Reciprocal effects of

    reading and mathematics? Beyond the cross-lagged panel model. *Developmental*

    Psychology, 56(5), 912–921. https://doi.org/10.1037/dev0000902

Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In R.

    H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 92–108). The Guilford

    Press.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal*

    of Educational and Behavioral Statistics, 28(2), 135–167.

    https://doi.org/10.3102/10769986028002135

Berry, D., & Willoughby, M. T. (2017). On the Practical Interpretability of Cross-Lagged Panel

    Models: Rethinking a Developmental Workhorse. *Child Development, 88(4),* 1186–1206.

    https://doi.org/10.1111/cdev.12660

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

    https://doi.org/10.1002/9781118619179

Bornstein, M. H., Putnick, D. L., & Esposito, G. (2017). Continuity and Stability in

    Development. *Child Development Perspectives, 11(2),* 113–119.

    https://doi.org/10.1111/cdep.12221

Bovaird, J. A. (2007). Multilevel structural equation models for contextual factors. In T. D.

    Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling Contextual Effects in Longitudinal*

    Studies (p. 149–182). Lawrence Erlbaum Associates Publishers.

Bureau of Labor Statistics, U.S. Department of Labor. (2019). *National Longitudinal Survey of*

    Youth 1979 Cohort, 1979-2016 (rounds 1-27). Produced and distributed by the Center for

    Human Resource Research (CHRR), The Ohio State University.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American*

    Psychologist, 32(7), 513–531.

Bronfenbrenner, U., & Morris, P. A. (2006). The Bioecological Model of Human Development.

    In *Handbook of Child Psychology: Theoretical Models of Human Development* (p. 793-

    828).

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in

    structural equation models: Causes, consequences, and strategies. *Sociological Methods*

    & Research, 29(4), 468–508. https://doi.org/10.1177/0049124101029004003

Crosby, D. & Mendez, J. (2016). *Using Existing Large-Scale Data to Study Early Care and Education Among Hispanics: Project Overview and Methodology*. Washington, DC: The National Research Center on Hispanic Children and Families.

Curran P. J. (2003). Have Multilevel Models Been Structural Equation Models All Along?. *Multivariate Hehavioral Research, 38(4),* 529–569. https://doi.org/10.1207/s15327906mbr3804_5

Curran, P. J., Howard, A. L., Bainter, S., Lane, S. T., & McGinley, J. S. (2014). The Separation of Between-person and Within-person Components of Individual Change Over Time: A Latent Curve Model with Structured Residuals. *Journal of Consulting and Clinical Psychology, 82(5),* 879–894. https://doi.org/10.1037/a0035297

De Ayala, R. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

Diallo, T. M. O., Morin, A. J. S., & Parker, P. D. (2014). Statistical power of latent curve models to detect quadratic growth. *Behavior Research Methods, 46(2),* 357–371. https://doi.org/10.3758/s13428-013-0395-1

Diallo, T. M. O., & Morin, A. J. S. (2015). Power of latent curve models to detect piecewise linear trajectories. *Structural Equation Modeling, 22(3),* 449–460. https://doi.org/10.1080/10705511.2014.935678

Diallo, T. M. O., & Lu, H. (2017). Consequences of Misspecifying across-cluster time-specific residuals in multilevel latent curve models. *Structural Equation Modeling, 24(3),* 359–382. https://doi.org/10.1080/10705511.2016.1247647

Early Childhood Data Collaborative. (2010). *Building and using coordinated state early care and education data systems: A framework for state policymakers*. Washington, DC: Early Childhood Data Collaborative.

Fan, X., & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. *Educational and Psychological Measurement, 58,* 701-735.

Feldon, D. F., & Litson, K. (2021). Modeling Theories and Theorizing Models: an Attempted Replication of Miller-Cotto & Byrnes' (2019) Comparison of Working Memory Models Using ECLS-K Data. *Educational Psychology Review, 33(4),* 1907–1934. https://doi.org/10.1007/s10648-021-09596-8

Fischer, K. W., & Bullock, D. (1984). Cognitive development in school-age children: Conclusions and new directions. In W. A. Collins (Ed.), *Development during middle childhood: The years from six to twelve*, 70-146. Washington, DC: National Academies Press (US).

Graham, J. M. (2008). The General Linear Model as Structural Equation Modeling. *Journal of Educational and Behavioral Statistics, 33(4),* 485–506. https://doi.org/10.3102/1076998607306151

Grimm, K. J., & Widaman, K. F. (2010). Residual structures in latent curve modeling. *Structural Equation Modeling, 17(3),* 424–442. https://doi.org/10.1080/10705511.2010.489006

Grimm, K. J., Helm, J., Rodgers, D., & O'Rourke, H. (2021). Analyzing cross-lag effects: A comparison of different cross-lag modeling approaches. *New Directions for Child and Adolescent Development, 2021(175),* 11–33. https://doi.org/10.1002/cad.20401

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20(1),* 102–116. https://doi.org/10.1037/a0038889

Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling.* The Guilford Press.

Hox, J. (2002). *Multilevel analysis techniques and applications*. Lawrence Erlbaum Associates Publishers.

Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Statistical Analysis* (pp. 281–294). Oxford University Press.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8(2),* 157–174. https://doi.org/10.1207/S15328007SEM0802_1

Hox, J. J., & Maas, C. J. M. (2005). Multilevel Analysis. *Encyclopedia of Social Measurement*, 785–793. https://doi.org/10.1016/B0-12-369398-5/00560-0

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64(2),* 157–170. https://doi.org/https://doi.org/10.1111/j.1467-9574.2009.00445.x

Hsu, H. Y., Lin, J., & Skidmore, S. T. (2018). Analyzing individual growth with clustered longitudinal data: A comparison between model-based and design-based multilevel approaches. *Behavior research methods, 50(2),* 786–803. https://doi.org/10.3758/s13428-017-0905-7

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6(1),* 1–55. https://doi.org/10.1080/10705519909540118

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International; Lawrence Erlbaum Associates, Inc.

Kenny, D. A., & Milan, S. (2012). Identification: A nontechnical discussion of a technical use. In *Handbook of structural equation modeling* (pp. 145–163). Retrieved from http://uconn.illiad.oclc.org/illiad/illiad.dll?Action=10&Form=75&Value=372980%5Cnpapers2://publication/uuid/68EF2707-040C-4CBC-A185-5D1C2FF33723

Kish, L. (1965). Survey sampling. Wiley.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking* (3rd ed.). New York, NY: Springer.

Lai, M.H., & Kwok, O. (2015). Examining the Rule of Thumb of Not Using Multilevel Modeling: The "Design Effect Smaller Than Two" Rule. *The Journal of Experimental Education, 83,* 423 - 438.

Lai, K. (2020). Confidence Interval for RMSEA or CFI Difference Between Nonnested Models. *Structural Equation Modeling: A Multidisciplinary Journal, 27(1),* 16–32. https://doi.org/10.1080/10705511.2019.1631704

Levy, R., & Hancock, G. R. (2007). A framework of statistical tests for comparing mean and covariance structure models. *Multivariate Behavioral Research, 42(1),* 33–66. http://dx.doi.org/10.1080/00273170701329112

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level

effects in contextual studies. *Psychological methods, 13(3),* 203–229.

https://doi.org/10.1037/a0012869

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1(3),* 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom Climate and Contextual Effects: Conceptual and Methodological Issues in the Evaluation of Group-Level Effects. *Educational Psychologist, 47(2),* 106–124. https://doi.org/10.1080/00461520.2012.670488

McKee, L., Colletti, C., Rakow, A., Jones, D. J., & Forehand, R. (2008). Parenting and child externalizing behaviors: Are the associations specific or diffuse? *Aggression and Violent Behavior, 13(3),* 201–215. https://doi.org/10.1016/j.avb.2008.03.005

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28(2),* 295–314. https://doi.org/10.1007/s10648-014-9287-x

Meredith, W., & Tisak, J. (1990). Latent curve analysis. Psychometrika, 55(1), 107–122. https://doi.org/10.1007/BF02294746

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological Methods, 21(2),* 151–163. https://doi.org/10.1037/met0000038

Mund, M., & Nestler, S. (2019). Beyond the Cross-Lagged Panel Model: Next-generation statistical tools for analyzing interdependencies across the life course. *Advances in Life Course Research, 41*, 100249. https://doi.org/10.1016/J.ALCR.2018.10.002

Muthén, B.O. (1998-2004). *Mplus Technical Appendices.* Los Angeles, CA: Muthén & Muthén.

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA: Muthén & Muthén.

Newsom, J. T. (2012). *Improper solutions in SEM. Unpublished paper.* Available at: http://web.pdx.edu/~newsomj/semclass/ho_improper.pdf.

Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. Journal of *Personality and Social Psychology, 120(4),* 1013-1034. http://dx.doi.org/10.1037/pspp0000358

Patterson, G. R., & Yoerger, K. (2002). A developmental model for early- and late-onset delinquency. In J. B. Reid, G. R. Patterson, & J. Snyder (Eds.), *Antisocial behavior in children and adolescents: A developmental analysis and model for intervention* (p. 147–172). American Psychological Association. https://doi.org/10.1037/10468-007

Pianta, R. C., & Walsh, D. J. (1996). *High-risk children in schools: Constructing sustaining relationships.* New York: Routledge.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling, 18(2),* 161–182. https://doi.org/10.1080/10705511.2011.557329

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent curve modeling*. SAGE Publications, Inc. https://www.doi.org/10.4135/9781412984737

Preacher, K. J. (2010). Latent curve models. In G. R. Hancock & R. O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 185-198). London: Routledge.

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

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling.* Lawrence Erlbaum Associates Publishers.

Rimm-Kaufman, S. E., & Pianta, R. C. (2000). An Ecological Perspective on the Transition to Kindergarten: A Theoretical Framework to Guide Empirical Research. *Journal of Applied Developmental Psychology, 21(5),* 491–511. https://doi.org/10.1016/S0193-3973(00)00051-4

Ross, C. E., & Broh, B. A. (2000). The role of self-esteem and the sense of personal control in the academic achievement process. *Sociology of Education, 73(4),* 270–284. http://dx.doi.org/10.2307/2673234

Roy, M., & Giraldo-García, R. (2018). The Role of Parental Involvement and Social/Emotional Skills in Academic Achievement: Global Perspectives. *School Community Journal, 28(2),* 29–46. Retrieved from https://search.proquest.com/docview/2166969249?accountid=188730

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688-701.

Ryu E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in psychology, 5, 81*, 1-9. https://doi.org/10.3389/fpsyg.2014.00081

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage Publications, Inc.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6(2),* 461–
    464. http://www.jstor.org/stable/2958889

Scott, P. W. (2018) DISAGGREGATING BETWEEN AND WITHIN-PERSON EFFECTS IN
    AUTOREGRESSIVE CROSS-LAG MODELS. Doctoral Dissertation, University of
    Pittsburgh. (Unpublished)

Stapleton, L. M. (2006). Using Multilevel structural equation modeling techniques with complex
    sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A
    Second Course.* Greenwich: Information Age.

Stapleton, L. M., & Kang, Y. (2018). Design Effects of Multilevel Estimates From National
    Probability Samples. *Sociological Methods and Research, 47(3),* 430–457.
    https://doi.org/10.1177/0049124116630563

Tourangeau, K., Nord, C., Lê, T., Sorongon, A.G., Hagedorn, M.C., Daly, P., and Najarian, M.
    (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-
    K:2011), User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic
    Codebook, Public Version (NCES 2015-074).* U.S. Department of Education.
    Washington, DC: National Center for Education Statistics.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L. and
    Najarian, M. (2019). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–
    11 (ECLS-K:2011) User's Manual for the ECLS-K:2011 Kindergarten–Fifth Grade Data
    File and Electronic Codebook, Public Version (NCES 2019-051).* U.S. Department of
    Education. Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2019). *SLDS Issue Brief: Answering Key Questions With an Early Childhood Integrated Data System.* Washington, D.C.: Institute of Education Sciences, National Center for Education Statistics.

U.S. Department of Health and Human Services. (2020). Head Start Family and Child Experiences Survey (FACES), United States, 2014-2017. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR36643.v6

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods, 24(5),* 637-657. http://dx.doi.org/10.1037/met0000210

West, J. (2017). National Longitudinal Studies of Kindergarten Children: Historical Context and Ongoing Contributions. *AERA Open, 3(2),* 233285841770168. https://doi.org/10.1177/2332858417701684

Wilder, S. (2014). Effects of parental involvement on academic achievement: A meta-synthesis. *Educational Review, 66(3),* 377–397. https://doi.org/10.1080/00131911.2013.780009

Wu, J.-Y., & Kwok, O. (2012). Using SEM to Analyze Complex Survey Data: A Comparison between Design-Based Single-Level and Model-Based Multilevel Approaches. *Structural Equation Modeling, 19(1),* 16-35, DOI: 10.1080/10705511.2012.634703.

Yuan K. H. (2005). Fit Indices Versus Test Statistics. *Multivariate behavioral research, 40(1),* 115–148. https://doi.org/10.1207/s15327906mbr4001_5

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi,. A., Pierides, D. C., Koval, P., & Diener, E. (2020). From Data to Causes II: Comparing Approaches to Panel Data Analysis. *Organizational Research Methods*, 23(4), 688–716. https://doi.org/10.1177/1094428119847280