CAN WE DISTINGUISH BETWEEN DIFFERENT LONGITUDINAL MODELS FOR ESTIMATING NONLINEAR TRAJECTORIES?

Ai Ye

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in the Department of Psychology and Neuroscience in the College of Arts and Sciences.

Chapel Hill
2022

Approved by:

Kenneth A. Bollen

David M. Thissen

Patrick J. Curran

ABSTRACT

Ai Ye: Can We Distinguish Between Different Longitudinal Models for Estimating Nonlinear Trajectories?
(Under the direction of Kenneth A. Bollen)

Substantive theory rarely provides specific enough information to guide our selection of the optimal model for longitudinal data. Instead, researchers are more likely to rely on models common to their field, even if it is not appropriate. The purpose of our study is to assess whether researchers can use overall goodness of fit measures from structural equation models to correctly find the data generating model (DGM) from among a broad set of different longitudinal models. I use four different DGM adapted from published empirical studies. We compare goodness-of-fit statistics (e.g., p-value, CFI, RMSEA, etc.) of the DGM with those of six alternative models. Overall, the BIC performed best in selecting the DGM, though no fit statistic was flawless. In the absence of substantive theory, I recommend that researchers begin with the most general longitudinal model and test whether it can be simplified by eliminating parameters.

To the L. L. Thurstone Psychometric Lab

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Longitudinal panel data consists of repeated measures for the same variables on the same cases over time. Such data have been widely used in many research areas across education, psychology, behavioral and social sciences. A primary interest of researchers when analyzing longitudinal data are the characteristics of change in the repeatedly measured variables. The shape, rate, or pattern of such change indicates the nature of the developmental processes in the outcome of interest. In the longitudinal data analysis, many studies have aimed to optimally model the overall shape of the growth trajectories for all subjects based on the hypothesized model (Blozis, 2007; Duncan et al., 1994; Hancock & Lawrence, 2006). Model choice is indeed crucial, because the fitted model with the corresponding set of parameter estimates recovers the growth pattern, which in turn represent the substantive theory that is initially ambiguous. However, the conundrum lies in the fact that there usually is a lack of substantive theory to dictate a particular form of growth model that can guide the selection of the optimal statistical model to fit the data. When this is the case, researchers may use exploratory approach to search for the optimal growth shape informed by their data. However, this data-driven approach poses many challenges to researchers and it often requires both advanced training in the statistical methodology and knowledge in the substantive area of interest. In addition, we do not know whether different longitudinal models can be distinguished from each other. If a wrong model is chosen, the model-implied trajectory could be misleading. Given its importance, the purpose of the present study is to investigate to what extent the optimal longitudinal model can be distinguished from alternative ones in the absent of theoretical hypotheses by using goodness-of-fit statistics and parameter estimates. The present study examines several longitudinal models with different ways of capturing nonlinear trends. Towards this overarching goal, the study is structured as below. This section is followed by a brief review of the longitudinal models under investigation. Following this are sections on current practice and the challenges of selecting optimal longitudinal models, and previous examinations of longitudinal

models. The next sections present the goals of the study, the simulation design, and the analytical procedures. Finally, descriptions of three empirical studies used for data generation are provided.

1.1 Review of Longitudinal Models

In psychology, two common approaches are considered to analyze panel data. The first is the quasi-simplex model or the autoregressive model (AR). AR model has deep roots in both the social science literature and in econometrics research (Anderson, 1960; Duncan, 1969; Humphreys, 1960; Kessler & Greenberg, 1981; Rogosa & Willett, 1985). It can be easily extended to a broader family of time series models called the Autoregressive Integrated Moving Average models (ARIMA; Hamilton, 1994). A second approach is the latent growth curve or latent trajectory model (LGCM). Though these models have a long history, my focus is on LGCM as developed in social and behavioral sciences under the latent variable framework

There are of course other model choices for longitudinal data which overlap with these two basic forms or have some combined features thereof. Just to list a few, latent curve model with structured residuals (Curran et al., 2014), latent dual change score models (Ghisletta & McArdle, 2001; McArdle, 2001; McArdle, 2009), latent state trait models with or without autoregressive components (Cole et al., 2005; Kenny & Zautra, 2001; Steyer et al., 1992; Steyer & Schmitt, 1994), a general panel model that combines fixed effects and random effects models with autoregressive outcome variables (Bollen & Brand, 2010; Wooldridge, 2002), and growth curve models with autoregressive disturbances (Azzalini, 1987; Chi & Reinsel, 1989; Goldstein et al., 1994), etc. Many of these models are proposed under different scenarios, some fully or partially nested within others. This became clear with the development of the Autoregressive Latent Trajectory (ALT) that directly synthesize AR and LGCM, and demonstrated that many of these models are special cases (i.e., restricted forms) of the ALT model (Bollen & Curran, 1999, 2004; Curran & Bollen, 1999, 2001). In what follows, I provide the equations and path diagrams for each of these three longitudinal models, autoregressive, together with its multivariate case autoregressive cross-lagged model, latent growth curve model, and autoregressive latent trajectory model.

1.1.1 Autoregressive Cross-Lagged Model (AR-CL)

A defining feature of the AR model is that current observation y_{it} for person i at time t is determined by the previous values of the same variable $y_{i,t-1}$ also known as the âlaggedâ effect. The

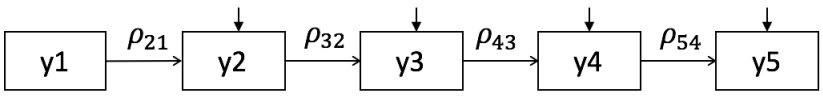

Figure 1.1: Path Diagram of An AR(1) Model over Five Waves

lagged effect is weaker among observations with increasing time lapse because the correlation between two observations typically becomes smaller as they get further apart in time (Guttman, 1954). Usually, a lag-1 assumption is imposed, in which only the first order of lagged effect (i.e., the effect from the adjacent time point) is considered, the score can be partially determined given the previous score. The lagged variable can be easily generalized to latent variables with multiple indicators. Alternatively, a univariate AR model can also be extended to a multivariable case in which one variable might have a lagged effect on other variable(s) net of the AR process within itself, called âcrossed-lagged (CL)â effects. In other words, an AR-CL model allows us to deliberately consider potential bidirectional relationships between two constructs over multiple time points. It does so by a simultaneous examination of longitudinal influences of one variable over another variable and vice versa, while also controlling for concurrent correlations between variables along with the stability of each variable over time (Bollen & Curran, 2006; Selig & Little, 2012).

Figure 1.1 shows an example of AR(1) model of repeated measures over five waves with order-one lagged effect $\rho_{t,t-1}$. Note that this AR(1) lagged coefficient can vary across time (i.e., time-specific) but there is no subscript of person i, indicating homogeneous effect across individuals. The fact that either the AR or CL effects are invariant across individuals in the sample suggests that lagged relations focus exclusively on within-person changes over time. It is the observed score for each individual in conjunction with the AR or CL coefficient that predicts the next value for that individual. Comparisons across individuals can occur by knowing the different AR and CL values of different individuals. In addition, since AR model does not use any time metric to explain the longitudinal dependency, no individual trajectory with an explicit form or shape is modeled. In other words, the patterns of change and shape of trajectory, whether linear or nonlinear, is inherent within the AR process. Trajectories are predicted by knowing the coefficients and the values of the variables for an individual.

Suppose y_{it} is a repeated measure for person i at time j, where $i = 1, ..., N; t = 2, ..T$, the

Equation for the simplex AR(1) model is

$$y_{it} = \alpha_t + \rho_{t,t-1} y_{i,t-1} + \epsilon_{it} \tag{1.1}$$

in which $\rho_{t,t-1}$ is the first-order or lag-1 autoregressive coefficient at time t, α_t is the time-specific intercept term and ϵ_{it} is the residual term for person i at time t. Assumptions in this equation are: for $\forall i$ and $\forall t$, $E(\epsilon_{it}) = 0$, $Cov(\epsilon_{it}, y_{i,t-1}) = 0$; $Cov(\epsilon_{it}, \epsilon_{jt}) = 0 \forall t$ and $i \neq j$, and $Cov(\epsilon_{it}, \epsilon_{jt}) = \sigma^2_{\epsilon_t}$ for $\forall t$ and $i = j$; and nonautocorrelated disturbance $Cov(\epsilon_{it}, \epsilon_{j,t+k}) = 0$ for $\forall k \neq 0$.

1.1.2 Latent Growth Curve Model (LGCM)

In contrast to AR-CL model where the change over time is modeled as each variable depending uniformly on its prior observation rather than having an explicit form of time variable, the LGCM focuses on individual trajectories as a function of time for repeated measures (Bollen & Curran, 2004, 2006; Meredith, 1984; Meredith & Tisak, 1990). Each case in the sample is modeled as an individual trend of change dictated by latent growth factors such as an intercept and a slope in a linear case. The LGCM has been increasingly popular in longitudinal studies because it has attractive property that it shows the inter-individual differences in their intra-individual trajectories over time (Duncan et al., 2013; Meredith & Tisak, 1990; Preacher et al., 2008). Three major types of parameters are estimated to realize that goal: the mean parameters of growth factors dictate the averaged shape of trajectory from the sample, the variance parameters of growth factors represent the variability of individual trajectories around that mean curve, and finally, the covariance parameters of the growth factors reveal the relationship between growth factors. An added advantage of LGCM is the capability to easily incorporate covariates that might explain between-person variabilities in their growth trajectories. Readers interested in fitting an unconditional and conditional LGCM may refer to Bollen and Curran (2006) for more illustrations.

In a univariate unconditional linear LGCM Latent Growth Curve Model

$$y_{it} = \alpha_i + \Lambda_{t2} \beta_i + \epsilon_{it} \tag{1.2}$$

where α_i is the random intercept for person i (different from the intercept term α_t in the previous

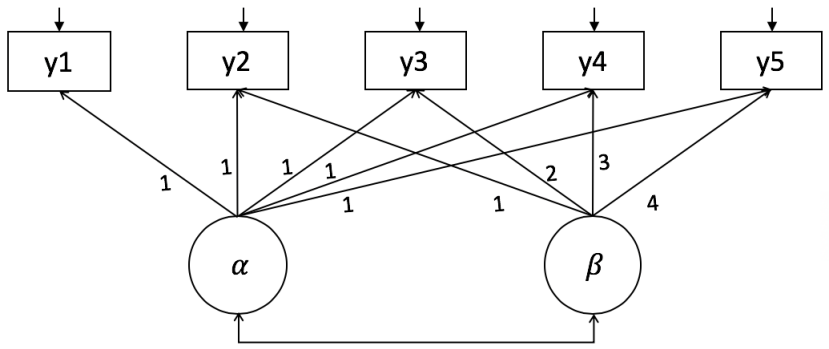

Figure 1.2: Path Diagram of A Linear LGCM over Five Waves

AR model), and β_i is the random slope for person i, so that

$$\alpha_i = \mu_\alpha + \zeta_{\alpha i}$$

$$\beta_i = \mu_\beta + \zeta_{\beta i}$$

where μ_α and μ_β are the mean intercept and slope across all cases. The $\zeta_{\alpha i}$ and $\zeta_{\beta i}$ are residuals with means of zero and uncorrelated with ϵ_{it}.

We assume that $E(\epsilon_{it}) = 0$ for $\forall i$ and $\forall t$, $Cov(\epsilon_{it}, \beta_i) = 0$ and $Cov(\epsilon_{it}, \alpha_i) = 0$ for $\forall i$ and $t = 2, 3, ..., T$; $Cov(\epsilon_{it}, \epsilon_{jt}) = 0$ for $\forall t$ and $i \neq j$, and $Cov(\epsilon_{it}, \epsilon_{jt}) = \sigma^2_{\epsilon_t}$ for $\forall t$ and $i = j$; and nonautocorrelated disturbance $Cov(\epsilon_{it}, \epsilon_{i,t+k}) = 0$ for $\forall k \neq 0$.

Under the framework of structural equation models (SEM), a LGCM can be specified as a polynomial function of repeatedly measured outcome regressed on a time variable. With the initial level and the rate of change (i.e., the intercept and slope of the growth line, respectively) being the latent factor of the SEM, the factor loadings with fixed values represent the value of time (see Figure 1.2). Such linear LGCM can be easily extended to a nonlinear LGCM by incorporating higher-order power functions (e.g., cubic) to represent the nonlinear characteristic of the trajectory over time. For example, in a quadratic term, a curvilinear trend is modeled as a product of a quadratic latent factor and a second-order time metric as its factor loading (see Figure 1.3).

In a univariate unconditional LGCM with a Quadratic term

$$y_{it} = \alpha_i + \Lambda_{t2}\beta_i + \Lambda_{t3}\beta_i^2 + \epsilon_{it} \tag{1.3}$$

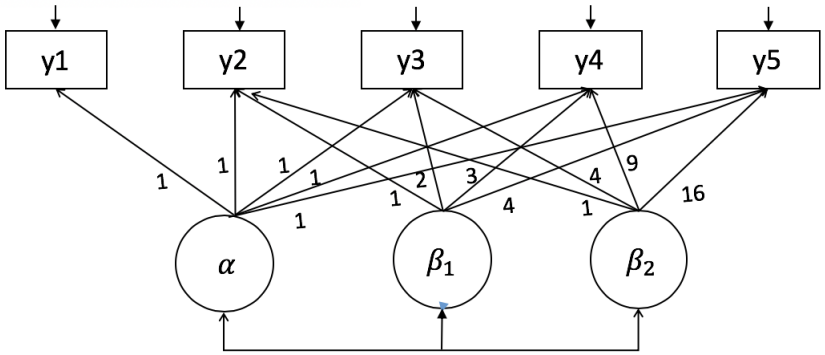

Figure 1.3: Path Diagram of An LGCM with A Quadratic Term over Five Waves

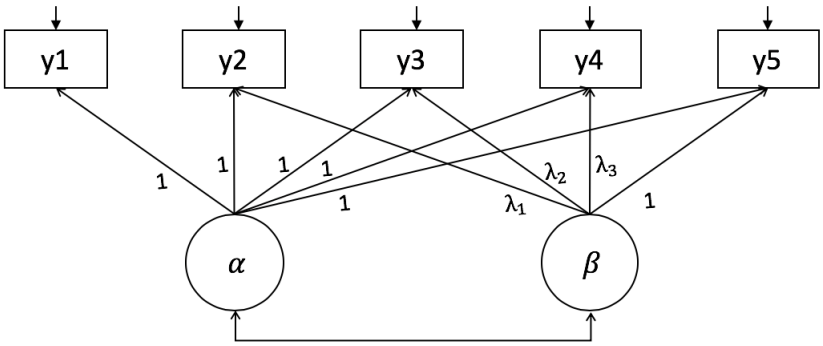

Figure 1.4: Path Diagram of An LGCM with Freed Loading over Five Waves

where Λ_{tp} is the factor loading for each moment of latent variable, the values in Λ_{tp} allow the incorporation of the shape (linear or nonlinear) of trajectories, α_i is the random intercept for person i, β_i is the random slope for person i, and β_i^2 is the random quadratic term for person i, and that

$$\alpha_i = \mu_\alpha + \zeta_{\alpha i}$$

$$\beta_i = \mu_\beta + \zeta_{\beta i}$$

$$\beta_i^2 = \mu_{\beta^2} + \zeta_{\beta^2 i}$$

Note that the trajectory model can be extended in several ways to allow for alternative form of nonlinear patterns of change. For example, Meredith and Tisak (1984, 1990) proposed the latent-basis (“freed loading") LGC model where the curvilinear trajectories modeled by freeing one or more factor loadings. This way, nonlinear patterns are accommodated by retaining the basic form of the model above but estimating all but two of the factor loadings for the slope factor (typically

the first two or the first and last), rather than fixing these to linearly or nonlinearly increasing values (see Figure 1.4). The freed loadings allows for flexible forms as a type of nonlinear âsplineâ that best fits the data between any two time points. A more general form of using the concept of âsplineâ is what is known as piecewise linear growth models using two or more linear piecewise splines, developed under the mixed modeling framework (Raudenbush & Bryk, 2002). The general idea is to break the full range of the time period into a number of pieces using deterministic knots, and fit a regular linear LGCM to each piece of trajectory. The selection of knots often carry subjective, transitional meaning and can sometimes vary randomly across individuals. Yet there exists other functional forms of nonlinear curves, such as random curve models with complex error structures (e.g., Cudeck & Harring, 2007; Diggle et al., 1994). But these are less commonly used due to the difficulty to incorporate in SEM (McArdle, 2001), and hence these are not included in the current investigation.

1.1.3 Autoregressive Latent Trajectory Model (ALT)

In earlier literature, AR-CL and the LGCM were considered competing techniques for the analysis of panel data (Mandys et al., 1994; Rogosa & Willett, 1985), but an increasing number of studies have suggested that this is not necessarily the case (Bollen & Curran, 2004; Curran & Bollen, 1999; Ou et al., 2017). Instead of being mutually exclusive, these two models simply uses different functional forms relying on their corresponding assumptions to capture developmental processes of distinct yet perhaps overlapping nature. A hybrid model that synthesizes these two models is first proposed by Bollen and Curran (Bollen & Curran, 1999, 2004; Curran & Bollen, 1999, 2001), and the authors later formally introduced it in published journal (Bollen & Curran, 2004) and book (Bollen & Curran, 2006), calling it the Autoregressive Latent Trajectory (ALT) model. Bollen and Curran (2004) demonstrated that the AR-CL model and the LGCM are special cases of the ALT model, but the latter inherits the appealing characteristics of both more restrictive models. Specifically, the ALT model processes the AR modelâs ability to include the prediction from the prior knowledge of a variable to its current value (and previous knowledge of another measure in a multivariate case, i.e., AR-CL model), it also shares the property with the LGCM that individual variability is allowed via the random coefficients governing the trajectory process (rather than assuming that all are governed by the same process in the same manner). By incorporating AR process and LGCM, the ALT model leads to a flexible, hybrid model. As such, researchers can

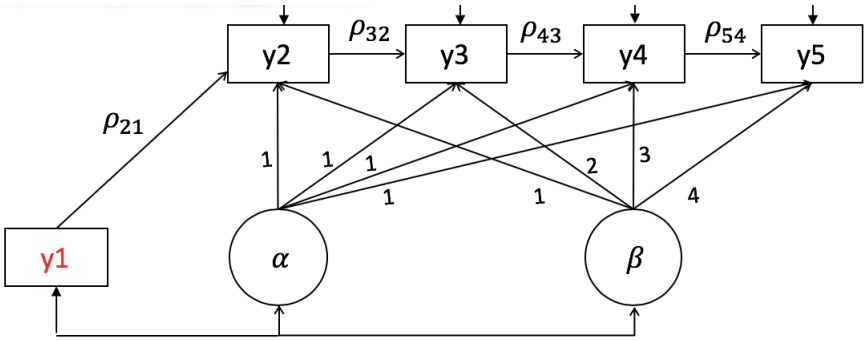

Figure 1.5: Path Diagram of A Linear ALT over Five Waves

compare the ALT model to alternative models or reduced it to more restricted forms in the evaluation of the optimal model that describes a data set (Bollen & Curran, 2004). To overcome complications or avoid nonlinear constraints to the model to estimate for the first wave of data, the simplest is to consider the first repeated measure as predetermined (or "exogenous"). Because ALT is a synthesized model of AR and LGCM, nonlinearity trend could be modeled in the same way as it is in LGCM, that is, either as a quadratic term of a polynomial function, or in a freed loading form (Bauldry & Bollen, 2018).

In a Linear ALT

$$y_{it} = \alpha_i + \Lambda_{t2}\beta_i + \rho_{t,t-1}y_{i,t-1} + \epsilon_{it} \tag{1.4}$$

All assumptions made for the AR(1) and LGCM model apply to the ALT model. The random intercept and slope factors can be expressed the same, but they are *net* the lagged time-specific effects in an ALT model.

We usually treat y_{i1} as predetermined, and thus:

$$y_{i1} = \nu_1 + \epsilon_{i1}$$

where the predetermined y_{i1} correlates with α_i and β_i.

The ALT model can not be identified if the first measure is treated the same way as it is in a LGCM, after the AR is incorporated in the model. The variable measured at the first time point is predetermined, the model needs to be âstarted upâ. There are different ways to start up the model, one way illustrated by Bollen and Curran (2004) is the predetermined ALT in which the first

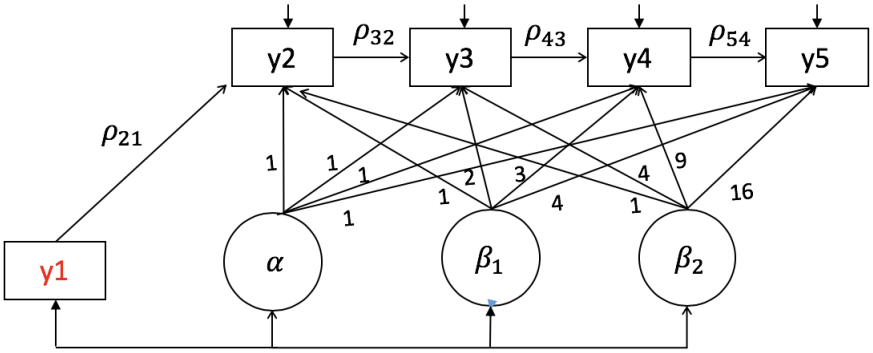

Figure 1.6: Path Diagram of An ALT with Quadratic Term over Five Waves

measure is treated as predetermined ("exogenous") and correlated with the LGC growth factors. Other specifications such as adding nonlinear constraints on the first measure are available (Bollen & Curran, 2004). Studies have explicitly examined what they call the âinitial condition specification (ICS)â under a variety of contexts, such as the ALT (Ou et al., 2017), space-state (Harvey, 1991; Jong, 1991), as well as time series models in SEM (du Toit & Browne, 2007). To keep the scope simple and be consistent with the majority of the literature, the current investigation uses the predetermined ALT.

In a Quadratic ALT

$$y_{it} = \alpha_i + \Lambda_{t2}\beta_i + \Lambda_{t3}\beta_i^2 + \rho_{t,t-1}y_{i,t-1} + \epsilon_{it} \tag{1.5}$$

Recent work has also extended ALT to nonlinear ALT (NL-ALT; Bauldry & Bollen, 2018) and to a yet more general framework of what is called the latent variable ALT (LV-ALT; Bianconcini, 2012). The NL-ALT incorporates nonlinear functions within the LCM component, i.e., as polynomial term or latent basis (freed loadings for slope factor). The LV-ALT is a generalization of the traditional ALT that models the two components of change on repeated latent factors instead of observed variables directly, and the inclusion of a measurement structure to allow for multiple indicators of the latent factors. This permits researchers to evaluate a wide variety of longitudinal models from statistical and psychometric literature, as well as their restrictive forms, as special cases of LV-ALT. It is thus recommended by the authors as a starting model to fit general longitudinal data, especially when there is little theoretical guidance. The current investigation focuses on ALT with observed variables, but our simulation study will involve a restrictive form of LV-ALT to

demonstrate a nesting relation of two models (see the next section below). Future study shall extend the current study to a more in-depth investigation of longitudinal models with latent variables.

1.2 Nesting Relations between the Longitudinal Models

Previous studies have demonstrated the relations between these longitudinal models, and showed model specification to achieve one from the other when they have nesting relations. We say that models are nested when the parameters for one model are a subset or a restricted version of the parameters of another (Bollen & Curran, 2004). The most common form of a nested, restricted model is when a subset of the parameters of the more general model are set to zero. For instance, the linear LGCM is equivalent to the quadratic LGCM when parameters related to the quadratic terms are fixed at zero, i.e., $\mu_{\beta^2} = Var(\zeta_{\beta^2}) = Cov(\zeta_\alpha, \zeta_{\beta^2}) = Cov(\zeta_\beta, \zeta_{\beta^2}) = 0$. A similar relation applies to the linear ALT and the quadratic ALT. As a hybrid representation, the ALT is more general framework than the LGCM or the AR, but the LGCM models are not directly nested within the ALT. In theory, we can obtain a model roughly equivalent to the LGCM from the ALT model with the same type of trajectory (i.e., linear, quadratic, or freed loadings) by setting all autoregressive coefficients $\rho_{t,t-1}$'s to zero. The main difference is that the ALT treats the first wave of data as predetermined while this first wave is treated like other repeated measures in the corresponding LGCM. In our evaluation on a LGCM and an ALT with nesting relations, we followed the simple procedure recommended by Bollen and Curran (2004, page 349), that is, to obtain a LGCM with predetermined first measure as does in a traditional ALT.

The nesting relations between an AR model and an ALT model in theory should be straightforward, that is, an AR model in Equ (1) is obtained when the growth-related terms (α_i and β_i) are omitted from Equ (4). However, the difference comes from the wave-specific intercept term for all individuals in the AR model, i.e., α_t in Equ (1), while the intercept term is typically set to zero in traditional ALT to correspond to what is done in LGCM. One can still obtain an AR from an ALT model following Bollen and Curran (2004), in which they specified a restricted version of the ALT where the intercept factor $\alpha_i = 0$ for all cases, $\mu_\beta = 1$ and $Var(\beta_i) = 0$ so that the freely estimated factor loadings function as the intercepts of the AR model. Although it reveals an equivalent relation between the AR and this restricted ALT, they do not fall under models with nesting relations. The complexity for the AR and the ALT models is that the time-specific intercepts are freed parameters in the AR, yet are typically set to zero in the ALT for the purpose of

identification. In order to test an AR model nested within an ALT, we will perform estimation on a set of more general ALT models, i.e., a LV-ALT without latent factors but one that include the intercepts as free estimates. If the LV-ALT models have convergence issues, we will adopt strategies like using the true population estimates as starting values for the free parameters. After we obtain these special forms of models with nesting relations, we can perform Likelihood Ratio Test (LRT) to evaluate 1) whether a more general model would return equal fit to data as does the data generating model, and 2) whether the fit would be poor if a simpler model than the true is chosen instead. Note that scenario 2) is more problematic than scenario 1), because a failure of 2) leads to a selection of a misspecified model, while a failure of 1) only suggests that a more general (or an unnecessarily complex) model might be chosen. For example, if the true model is AR, the LRT between the AR and the ALT with freely estimated intercepts would evaluate if the growth related parameters in the latter are indeed not significantly different from zero; vice versa, if the ALT is the true model, would the AR model return poor fit.

With the aforementioned model specification and special treatments, we have the following set of nesting models to evaluate (note that the special settings of the linear LGCM and the linear ALT are only used in LRT in each corresponding case. Results of other fit statistics in our simulation study are drawn from the models with the regular specification, respectively):

1. A linear LGCM is nested within a quadratic LGCM; similarly, a linear ALT is nested within a quadratic ALT.

2. A linear LGCM with predetermined first repeated measure is nested in a linear and quadratic ALT, and a quadratic LGCM with predetermined first repeated measure is nested in a quadratic ALT.

3. An AR is nested in a ALT (linear, quadratic, or freed loading) where time-specific intercepts are included as free parameters.

1.3 Selecting the Optimal Longitudinal Model: Current Practice and Challenges

Along with the development of longitudinal modeling techniques comes the challenge to choose the optimal model when fitting empirical data. In longitudinal data analysis, the model is selected to fulfill the goal of a study under the guidance of prior knowledge. Ideally, the hypothesized model

should optimally capture the overall shape of the growth trajectories across subjects in the sample (Blozis, 2007; Duncan et al., 1994; Hancock & Lawrence, 2006). For example, if the goal to study infant body weight gain (say, in their first year of birth) is to understand between-person variability in their individual trajectory of body weight gain over a course of time, a practice is to collect repeated measure of body weight on a sample of infants and to fit the data with a growth curve model. Since we know from previous theory that the growth in weight is the fastest in the first few months after a baby is born which is yet followed by a decreasing trend in the rate of weight gain, a researcher might thus consider a nonlinear growth curve by including in a quadratic term in the model to capture that change in growth rate.

However, two complications face the theory-driven approach. First, a theory that specific often is missing in the literature. Second, even if there are available theories, the existing ones might be controversial. In the previous example, a researcher might hypothesize that the best indicator of a babyâs current weight is the weight of the previous week, which would be represented better as an AR(1) model that is very different from an individual trajectory over time. Or there might be a third group of research who claim that the actual growth pattern is the combination of the two processes, so a hybrid model like ALT is the more reasonable model to fit. The lack of consistent knowledge on the nature of change is part of the knowledge researchers hope to gain from the information provided by the data, besides the estimated of the parameters that uniquely define the magnitude of pre-characterized change pattern.

As a consequence, researchers may use partially exploratory approach to search for the optimal model recovering the shape of growth that fit the data. Practically, an exploration can start with visual inspection of plotting individual raw scores over time. However, there are caveats about taking this approach. First, visualization is only suitable for a handful of data points, or with a subset of individuals from the sample. When the number of waves or the sample size gets larger, the observed shape of an individual curve is hardly interpretable and the variability between individuals is too massive to be informative. Second, some types of developmental process are hardly interpretable by plotting. It is very difficult to know what model to fit by plotting the means of the repeated measures or by visual observations of individual trajectories. A good example is the AR process, in which the lagged-effect does not have a consistent visual representation as in most case of other time series models. Part of the issue is that even the AR linear model can capture what

appear to be nonlinear patterns in the plots.

Additionally, visualization on raw score plotting might not be informative in the cases when multiple developmental processes coexist within a construct. Because when this is the case, the plot of individual raw scores or the mean score of a sample over time depicts an aggregated pattern that may not represent individual processes. Ideally, each trajectory needs to be modeled by a particular functional form simultaneously, the parameters of which recover the characteristics of that corresponding process. It also is worth noting that the interpretation of the parameters will depend of the set of parameters in that model. For instance, in the ALT model, we cannot interpret the random slope and random intercept in the same way that we would in the usual LGCM. The slope mean in LGCM depicts the mean rate of linear change while the slope in an ALT model refers to the mean rate of linear change above and beyond the AR process. Take the infantâs weight example again, when there exists both a rate of change and a lagged effect underlying the within-person weight growth process, the observed raw weights show the aggregate within-person weight change (ideally modeled as a hybrid model) that will not represent either the rate of change (ideally be modeled as growth curve factors) or the magnitude in lagged relation (ideally be captured by AR parameters). In fact, Bauldry and Bollen (2018) found that the ALT model was a better choice than was traditional LGCM in the weight gain example. The same value of the slopes from the two models represent vastly different linear shape, after controlling the AR values. Any effort to interpret the aggregate within-person change from a hybrid model in a comparable fashion with separate within-person change from a simpler model is misleading (e.g., Jongerling & Hamaker, 2011). Indeed, one needs to be cautious about using visual observations on the plotting of the means of the repeated measures or individual trajectories to inform what model to fit.

1.4 Selecting the Optimal Longitudinal Model: Recommendation

So what is a plausible practice of searching for the optimal model? Such practice might involve a great deal of speculation on the nature of change based on the substantive knowledge, understanding of the purpose of study, together with years of training and experience in applying the methodology. One way to proceed is to start with the most general longitudinal model (assumedly identifiable). The general form can be reduced to more parsimonious (restricted) form by adding constraints to the model and eliminate some parameters in a stepwise fashion. For example, Minjung and colleagues (2018) examined the performance of different conditional LGC starting models in terms

of the complexity of the mean and within-person variance-covariance structures in the presence of time-invariant covariates and found that only the fully saturated model (i.e., the most complex mean and within-person variance-covariance structures) recovers best the true growth trajectory using BIC and AIC fit statistics. Bianconncini and Bollen (2018) also stated that âif theory or prior work dictate the model, the LV-ALT is capable of specializing to many other simpler models and to comparing the simpler longitudinal model to other more general models. Alternatively, if there is little guidance on the best model to be selected, LV-ALT provides a way to empirically compare a wide variety of models to determine which is the most appropriate for the data (pp. 792)â.

Failing to use a general model as a starting model might result in a false model with misleading conclusion. This is particular of concern if researchers rely solely on fit statistics to pick the right model, that is, to stop the model search once observing a good model fit. A misspecified model might fit the data well under particular scenarios, when the model-implied statistics given the wrong set of parameters approximates the sample statistics. Neither fit statistics or likelihood ratio tests will be useful to identify such issues when the starting model is not general enough to include all necessary components (or a saturated model with all parameters as recommended by Minjung et al., 2018). An example of this problematic practice is given in Voelkle (2008), in which the true model is a growth curve with a quadratic term with unusual parameter values (e.g., very weak factor correlation). The model search started with one that omits the nonlinear component at the presence of a nonlinear trajectory. Although misspecified, the author found a linear ALT model, with additional (false) AR parameters accounting for omitted quadratic shape, achieved a sufficient fit. Of course the AR parameters would be misleading and the correct parameters are biased, as the consequence of not starting the model search with a general model (in this case, a nonlinear ALT with a quadratic component in the growth curve component). Though not mentioned by Voelkle (2008), the reverse situation also could be true, that is, the ALT could be the true generating model and a researcher might mistakenly fit a quadratic model to it.

Examples such as these raise the issue of whether starting with the most general model would be able to recover the true model that is more restrictive. If the general model permits us to distinguish between similarly appearing structures, then this encourages us to check our results with the general model. Alternatively, if we cannot easily distinguish between models, it suggests that we take a more cautious stance when presenting our longitudinal models.

Limited analytic research has provided examples of special cases where the models are statistically equivalent. For example, Hamaker (2005) showed that when the AR parameter lies between â1 and 1 and is invariant over time, the ALT model and the LGC model with AR relationships between the disturbances are algebraically equivalent. The equivalency no longer holds when the AR parameters in the ALT model vary over time. More recently, Bianconcini (2012) provided conditions under which a quadratic LGC could be made equivalent to an ALT model. However, the equivalency assumes regression coefficients of one and some mean and variance parameters set to zero, conditions that would be rare in practice. As interesting as these special cases are, statistical equivalence is not the same as substantive equivalence. These different models imply different processes that would have implications for theory testing or policy interventions. On the other hand, if the purpose is to just approximate the trajectories, either model with a good fit would be sufficient. Or they might be equally qualified for predicting future trends. However, model selection is hardly justifiable when the purpose is to draw substantive inference and to inform theory, as is common in social science research.

Beyond these special cases of equivalence, little is known about the ease of empirically distinguishing between the different longitudinal models. A limitation of past research is that the typical argument has generated data with say, Model A the true model and found that Model B provided a close approximation. But they did not take the other step of generating data with Model B to see whether Model A was a close approximation. What is missing in literature is a symmetric comparison on the consequences of fitting the data with different longitudinal models under a range of realistic conditions to see whether empirical methods can help us to determine the model that underlies the data.

CHAPTER 2

THE CURRENT STUDY

2.1 Goal of Study

Considering the existing ambiguity, the purpose of the present study is to investigate to what extent these different longitudinal models can distinguish the optimal model from competing alternative ones, that is, the degree to which researchers can trust model fit in finding the optimal model for the data. Relatedly, which fit indexes, if any, are most useful to accomplish this goal. We used a simulation approach so that the ground truth is known, the goodness-of-fit of models under investigation can be compared to that of the true data generating model. The research also will determine those conditions where it is most and least difficult to correctly empirically select the model.

2.2 Simulation Design and Simulation Factors

The goal of this simulation is to assess whether empirical tests and indexes can help distinguish the true data generating model from false longitudinal models. We propose that the investigation be performed via an empirical simulation design. We consider the following models: first-order autoregressive model, linear latent growth curve, latent growth curve with a quadratic term, latent growth curve with freed loadings, linear autoregressive latent trajectory, autoregressive latent trajectory with a quadratic term, and autoregressive latent trajectory with freed loadings.

Specifically, we will use three data generating models: an AR, a nonlinear LGCM with a quadratic term, and a linear ALT. All seven models discussed in the introduction, including a linear LGCM, a nonlinear LGCM with freed loadings as well as nonlinear ALT (either with quadratic term or freed loadings), will be used to fit the data. To keep the simulation of a manageable size, we will not use the last four models to generate the data.

To make the simulation conditions more realistic, we use as population parameters the parameter estimates from empirical studies that used the same type of longitudinal models (descriptions given in the next section). In order to introduce some randomness to the model and

increase generalizability, we also tested another set of the data generating model (DGM) with half sized original parameter values (except for residual variances). Three sample sizes are considered, with N = 300, 500, and 2000, to match the range of sample sizes from the three empirical studies (303, 536, and 2,061, respectively). The levels of the sample sizes also represent a moderate, medium, and large samples of longitudinal studies in practice. Each of the six DGMs will be evaluated with 1,000 replications. To best reflect difficult yet common missing data situation in longitudinal studies, we simulated each dataset with a moderate proportion of missingness (i.e.,15%) that is following the missing-at-random (MAR) mechanism, i.e., the missingness depends on the values of the observed variables or the variables that remain complete. We adopted the multivariate amputation procedure to relate the missingness on one variable to the missingness on another variable, as implemented in the function âamputeâ built in âmiceâ package mice. Under the MAR, the probability of becoming incomplete for each candidate is a weighted sum score that is calculated from a linear combination of the variables (Schouten, Lugtig, & Vink, 2018).

2.2.1 Analytical Procedure

Next, we fit each simulated data set with the seven models previously discussed. To select the model, we compared the goodness-of-fit statistics of the seven models, and perform likelihood ratio test statistics between the nested models. Additionally, we examined the implied growth form for each model (with their estimated parameters), the goal is to find how far off the predicted shape of growth trend is for each false model. Relatedly, we observed potential consequences of fitting the data by a false model. Lastly, we made recommendations for model selection and testing procedures. All simulation and analysis are conducted in Mplus Version 7, data and simulation management is done in R studio, and the package âMplusAutomationâ is used to automate the simulation runs. The code for the DGM is avaiable on Open Science Framework (OSF).

2.2.2 Outcome Measures

We will be using standard fit statistics in SEM literature to compare the goodness-of-fit of all the models fitted to the three sets of DGMs. In addition, we evaluate the recovery rate of the DGM as the best fitting model by each fit measure. For goodness of fit as well as the recovery of the DGM as the best fitting model, we compare performance across the fit statistics. The fit statistics we include in the current investigation are: chi square statistic, BIC (Schwarz et al., 1978), AIC (Akaike, 1987), CFI (Bentler, 1990), TLI (Bentler & Bonett, 1980), RMSEA (Steiger, 1990), and

SRMR (JÃ¶reskog & SÃ¶rbom, 1981).

We know that goodness-of-fit statistics are a measure of how well the model fit the data by evaluating the model implied covariance against the sample covariance (Bollen, 1989). However, there are different ways of calculating the residuals, hence the different format of fit statistics. First, the likelihood ratio chi square is a test of whether the hypothesized model fits as well as the saturated model to the data. Let F be the estimation object function, we know that $T = (N-1)F$ follows an asymptotic chi square distribution with degree of freedom df (Bollen, 1989). The p value reflects the statistical significance of the discrepancy between the model implied covariance matrix and the population covariance matrix, hence, smaller p values indicate a larger discrepancy hence a poorer fit. The chi square test might be easily subject to false positive results with large samples, that is, even a very small discrepancy will be identified as significant when the sample size is sufficiently large. However, if the model is true and the distributional assumptions hold, then the chi square test statistics will follow a chi square distribution regardless of sample size.

In comparison, CFI and TLI are baseline comparison indices. The formulas are derived by Equ 2.1 and Equ 2.2, respectively.

$$CFI = 1 - max[\frac{\chi_t^2 - df_t}{\chi_i^2 - df_i}, 0]. \tag{2.1}$$

where χ_i^2 is the chi square corresponding to the baseline model, or independence model, and χ_t^2 is the chi square corresponding to the fitted model.

$$TLI = \frac{(\chi_i^2/df_i) - (\chi_t^2/df_t)}{(\chi_i^2/df_i) - 1}. \tag{2.2}$$

For both CFI and TLI, a value 1 is an ideal fit. A value larger than 1 might indicate overfit or just sampling fluactuations, so they are forced to a max of 1. We follow the general guideline where a value larger than .90 indicates an acceptable fit and one larger than .95 an excellent fit (Bentler, 1990; Tucker & Lewis, 1973).

The RMSEA is a stand alone index defined by the square root of the discrepancy per degree of freedom:

$$RMSEA = \sqrt{\max[\frac{T - df}{df(N-1)}, 0]}. \tag{2.3}$$

The SRMR is given by:

$$SRMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}[(s_{ij} - \hat{\sigma}_{ij})]^2}{m(m+1)/2}}. \tag{2.4}$$

where s_{ij} is an element of the sample covariance matrix, $\hat{\sigma}_{ij}$ is an element of the model-implied covariance matrix, m is twice the number of observed variables.

For RMSEA and SRMR, a small, close to zero value indicates a good fit, a value less than .08 is generally considered an acceptable fit and one less than .05 an excellent fit (Browne & Cudeck, 1993; Hu & Bentler, 1999). However, one needs to be more cautious on the interpretation of these fit indices. For example, the cutoffs for some indexes such as RMSEA do not work consistently with small samples and are too forgiving for large samples.

We also include information criteria typically used for model selection in practice. BIC and AIC are measures for the relative amount of information lost based on an LRT when data are represented by a given hypothesized model compared to the saturated model. Hence, less information lost (i.e., smaller value) indicates a higher quality model (i.e., better fit). Note that we used BIC expressed by the chi square test statistic, the degrees of freedom and sample size, that is easy to interpret (see Equ 2.5). The interpretation of the value of this BIC is straightforward: a value greater than zero favors the saturated model over the estimated (hypothesized) model while a negative value favors the latter model (Raftery, 1995). Similar to BIC, AIC rewards goodness of fit as assessed by the likelihood function, with a penalty that is an increasing function of the number of estimated parameters. The difference between BIC and AIC when used for model selection is that BIC also penalize the incorrect model more with increased sample size, although the model's absolute fit looks better with a bigger the sample size is (other things equal). Sometimes the penalty on large sample is too strict so that BIC tends to favor the simplest model when candidate models present satisfying, comparable fits. Given a set of candidate models for the data, the preferred model is the one with the minimum BIC or AIC value.

Table 2.1: Population Parameters In The Three DGMs

Model Estimates	AR(1)	Quadratic LCM	Linear ALT
	Hansen et al. (2017)	Sterba (2014)	Zyphur et al. (2008)
Mean(α)	n/a	3.216	104.098
Mean(β_1)	n/a	9.923	10.492
Mean(β_2)	n/a	-5.384	n/a
Var(α)	n/a	.107	217.997
Var(β_1)	n/a	4.221	8.326
Var(β_2)	n/a	2.288	n/a
Cov(α, β_1)	n/a	.258	32.034
Cov(α, β_2)	n/a	-.211	n/a
Cov(β_1, β_2)	n/a	-2.821	n/a
ρ_{21}	.833	n/a	.363
ρ_{32}	.808	n/a	.287
ρ_{43}	.861	n/a	.274
ρ_{54}	.714	n/a	.249
ρ_{65}	n/a	n/a	.203
ρ_{76}	n/a	n/a	.166
ρ_{87}	n/a	n/a	.119
σ_1^2	.489	.194	172.98
σ_2^2	.258	.132	219.71
σ_3^2	.346	.121	218.67
σ_4^2	.306	.052	217.83
σ_5^2	n/a	.102	223.58
σ_6^2	n/a	.089	230.65
σ_7^2	n/a	.155	268.62

Note: ân/aâ suggests that parameter not estimated in the corresponding model

$$BIC = \chi_t^2 - df_t * \log(N). \tag{2.5}$$

$$AIC = \chi_t^2 - 2df_t. \tag{2.6}$$

2.3 Empirical Studies for Data Generation

In this section, we briefly introduce the background of the three empirical studies that we selected to be used in our empirical simulation.

2.3.1 Study I: Hansen et al. (2017) Fraction Magnitude Understanding using AR(1)

This study examined the co-development of early fraction magnitude understanding and mathematical achievement, as part of a larger longitudinal study aimed at understanding

mathematical development among elementary students. Third-graders (N = 746) in nine public schools from two adjacent districts initially participated in the study, 536 of which stayed throughout the entire assessment from their fourth to sixth grades. The outcome measure being repeatedly measured is fraction number line estimation, in which students estimated the location of 28 fractions and mixed numbers on 0-1 and 0-2 number lines. To score the measure, the percent absolute error (PAE) was calculated by dividing the absolute difference between the estimated and actual magnitudes by the numerical range of the number line (1 or 2), and then multiplying by one hundred for each estimate. Each student was assigned a single score that represented his or her mean PAE on the task. Scores were then multiplied by -1 so that higher scores indicated higher achievement. As an intermediate step toward the ultimate multivariate model of the cross-lagged effects between FNLE and a general mathematical outcome, the authors fit a AR(1) model of the FNLE. The AR(1) model indicates that development of FNLE follows an autoregressive process, specifically, a unit increase in FNLE score at one semester is expected to be associated with a .714 - .861 increase in the following semester from the fourth to sixth grades (see the first column in Table 2.1 for the lagged coefficients).

2.3.2 Study II: Sterba (2014) Infantâs Weight Growth using Nonlinear LGCM

This example uses panel data from the Longitudinal Health and Nutrition Survey (N = 2,631, of which 2,061 infants had complete data to track the functional form of infant growth in the Philippines. The purpose is to identify heterogeneity in growth that is associated with social and health behaviors. Pregnant mothers in one province in the Philippines were enrolled in 1983 and 1984. As a starting point, a linear LGM is fit, followed by quadratic, shape-factor, piecewise linear, piecewise quadratic, and exponential LGMs. Variances were estimated for all growth coefficients, and heteroscedastic residual variances across time were observed. The intercept was coded at birth (0 years) and the unit change in time was coded to be 1 year. The quadratic LGM implies that, on average, infants weigh 3.22 kg at birth, with an instantaneous velocity of 9.92 kg per year, that decelerates by 2 x â5.38 kg/year; there is significant individual variability in all aspects of change. Parameter estimates of the nonlinear LGCM are given in the second column of Table 2.1.

2.3.3 Study III: Zyphur et al. (2008) Job Performance using Linear ALT

This study aims to understand the longitudinal processes in job performance. The authors hypothesized both within-person and between-person change in job performance. They adopted the

ALT model to address job performance over time as a process such that past performance can determine future performance directly (AR component) and that individuals often have distinct latent performance trajectories due to individual-difference factors (i.e., LGC component). They used existing longitudinal data from Ployhart and Hakel (1998), who investigated the performance of 303 securities brokers. Performance was operationalized as the square root of gross sales commissions averaged across a 3-month period, making each performance observation equivalent to an annual quarter over eight quarters (i.e., 24 total months).

To test their hypothesis, the authors followed a stepwise model selection procedure to validate the co-existence of an autoregressive process and a linear growth curve (and excluded the quadratic shape), and thus obtained a final model of a linear ALT model. They concluded that the growth of job performance follows a moderate AR process and a linear, significant increase trend net of the AR process. This sample showed large variance in terms of their initial status (large intercept variance), and a moderate amount of variability in the rate of change net of AR process. There is also a positive relation between the intercept and slope factors, that is, those who start with a better status also tend to show a slightly faster improvement in their job performance. The final parameter estimates are presented in the third column of Table 2.1. We use these three models as data generating models in the simulation study, and the corresponding parameter estimates as the population parameters, respectively.

CHAPTER 3

RESULTS

In this section, we report our simulation results for the three data generating models (DGM) described above. The results have three components: rate of convergence, goodness of fit, and the recovery of the DGM as the best fitting model. For rate of convergence, we present the percent of converged models for the seven candidate models when fitting to the three DGM, respectively. For goodness of fit as well as the recovery of the DGM as the best fitting model, we compare performance across the fit statistics. In addition, we performed likelihood ratio chi square test (LRT) on models that have nesting relations with the DGM to evaluate whether they were statistically distinguishable.

3.1 Convergence Results

In our simulation, models such as the latent basis (or freed loadings) had difficulty reaching a final solution, a phenomenon called nonconvergence in the SEM literature (e.g., Anderson & Gerbing, 1984). Nonconvergence occurs when the estimation algorithm fails to arrive at values which meet prescribed minimum criteria within a set number of iterations (Anderson & Gerbing, 1984). It means that the set of parameter estimates in these models cannot meet the criteria under the default setting of Mplus (i.e., the value of the derivative convergence criterion used in the Quasi-Newton algorithm for ML estimator of continuous outcomes is set at .00005, with the maximum number of iterations at 1,000). Research has found that a nonconvergent solution is often associated with sampling fluctuation or model misspecification (Anderson & Gerbing, 1984). The corresponding guidance from Mplus is that âconvergence problems are often related to variables in the model being measured on very different scales, poor starting values, and/or a model being estimated that is not appropriate for the data. In addition, certain models are more likely to have convergence problems (Mplus Userâs Guide, pp. 415). Following their recommendation that âconvergence may be reached by increasing the number of iterations or using the preliminary parameter estimates as starting valuesâ, we implemented these three remedies: increased the number

Table 3.1: Rate of Convergence (%)

Simulation Factors	**Large Parameter**						**Small Parameter**					
	10TP			**6TP**			**10TP**			**6TP**		
	300	500	2000	300	500	2000	300	500	2000	300	500	2000
					DGM = Quadratic LGCM							
AR	100	100	100	100	100	100	100	100	100	100	100	100
lALT	100	100	100	100	100	100	100	100	100	100	100	100
qALT	100	100	99.8	100	100	100	100	100	100	100	100	100
lbALT	38	34	29	50	48	39	98	99	100	88	92	99
lLGCM	93.4	92.2	94.8	99.7	99.4	99.5	99.8	99.9	93.4	100	100	99.7
qLGCM	100	100	100	99.1	99.6	100	100	100	100	99.4	99.6	99.1
lbLGCM	14	11	2	100	100	100	100	0	0	0	0	0
					DGM = Linear ALT							
AR	100	100	100	100	100	100	100	100	100	100	100	100
lALT	100	100	100	100	100	100	100	100	100	100	100	100
qALT	100	100	100	98	100	100	100	100	99	100	100	
lbALT	65	59	62	77	81	68	93	97	100	71	77	94
lLGCM	100	100	100	100	100	100	100	100	100	100	100	100
qLGCM	100	100	100	100	100	100	100	100	100	100	100	100
lbLGCM	100	100	100	100	100	100	98	100	100	99	100	100
					DGM = AR					AR	100	100
100	100	100	100	100	100	100	100	100	100			
lALT	100	100	100	100	100	100	100	100	100	100	100	100
qALT	99	100	100	99	100	100	100	100	100	91	95	100
lbALT	0	66	86	12	12	10	0	0	0	85	91	99
lLGCM	100	100	100	100	100	100	100	100	100	100	100	100
qLGCM	100	99	100	100	100	100	100	100	100	100	100	100
lbLGCM	100	100	100	100	100	100	100	100	100	100	100	100

Note: the first column is the names for fitting models, in which AR = Autoregressive Model, lALT = Linear Autoregressive Latent Trajectory, qALT = Quadratic Autoregressive Latent Trajectory, lbALT = Latent Basis Autoregressive Latent Trajectory; lLGCM = Linear Latent Growth Curve Model, qLGCM = quadratic Latent Growth Curve Model; lbLGCM = latent basis Latent Growth Curve Model

of iterations (from 1,000 to 10,000), lowered the convergence criteria (from .00005 to .001), and manually inputted starting values. The starting values were corresponding parameter estimates we obtained by fitting the model to population covariance matrix (rather than to the sample covariance matrix). Only the last strategy helped some (but not all) problematic models reach convergence. For example, when fitting the ALT models to data generated by the AR model, the ALT models were not converged in the beginning. We then used estimates of the nonconverged models as starting values for the intercepts. Table 3.1 presents the percentage of convergence after implementing these strategies for each DGM (fitted model by row and simulation conditions by column).

We found that models with freed loadings were most likely to have convergence problems. For example, their convergence rates were particularly low when fitted to data generated by a quadratic LGCM in some conditions: the freed loadings LGCM only converged consistently to data with the 6-timepoint, the original or large parameter setting; for the data with long wave (10-timepoint), the rate ranged 2% to 14% in the large parameter conditions while no model converged in the half-sized or small parameter conditions, respectively. Most of the freed loadings ALT models converged when fitted to data generated by the small parameter values; in comparison, the rate ranged from 29% to

50% under the large-parameter conditions. Fitting the freed loadings ALT model to data generated by the linear ALT also had some trouble converging to a unique solution particularly under the large-parameter condition (ranging from 62% to 77%), although they are not necessarily the most parameter-heavy model (e.g., a 6-timepoint quadratic ALT has more parameters to estimate than does a 6-timepoint latent basis ALT).

Model misspecification is a possible source of nonconvergence. Since none of the DGM has a freed loadings trajectory, it is possible that nonconvergence results from misspecification in growth-related (slope) parameters, besides the omission of the AR parameters when the DGM is a linear ALT or an AR model. A model that omits important parameters (e.g., AR-related parameters) or includes wrong parameters (e.g., freed loadings slope parameters rather than a specified shape) might fail to return a function of estimable parameters that reasonably approximates the variance-covariance structure of the variables provided by the data. We also hoped to identify other factors that contributed to this problem. However, our results across simulation conditions did not clearly reveal the sources of nonconvergence. For example, we know that the larger sample sizes reduce sampling fluctuations and thus should relieve the issue of nonconvergence when the true model is fitted. While it is the case for the small-parameter conditions, we found that some models with larger sample sizes have lower chance of reaching convergence under certain situations, e.g., the freed loadings models across the 10-timepoint large-parameter conditions. Another observation is that the larger number of timepoints seemed to be accompanied by more frequent nonconvergent models, although one would intuitively think that complex model should be easier to converge with more waves of data. One reason might come from the difficulty of computing a larger covariance matrix in the maximum-likelihood estimation. Another guess is that the it is more difficult to approximate with alternative functions when the shape of the true trajectory or the AR processes get more precise with more waves of data, because any misspecification in the trajectory or patterns become increasingly more prominent. Further, the discrepancies between the small parameter versus and large parameter sets perhaps point to another scenario that is common in practice: some specific combinations of growth-related or AR-related parameters generate trajectories that might be more easily approximated by other functions.

Indeed, nonconvergence is a complex issue and should be interpreted with care. Based on the results of the simulation and prior literature we hypothesize that the nonconvergence in the

simulation is a function of some combination of model structural misspecifications of either the wrong model structure or including unnecessary parameters, smaller sample sizes, or higher error variances. A deeper examination on the causes of nonconvergence is outside the scope of the current study. For more details, we defer readers to the literature such as Anderson and Gerbing (1984), Bentler and Chou (1988), and JÃ¶reskog (1967).

3.2 Goodness-of-Fit Results

In this section, we present the model fit results from all the converged models by each DGM. We are aware that the results were affected by the omission of models that failed to reach a convergence. We will further discuss this in the limitation section. Comparisons are made across candidate fit measures on the goodness of fit for the true model versus the other estimated models. A well-behaved fit statistic is defined as one that consistently shows better fit for the true model than the misspecified models. We present the results according to the different DGMs.

3.2.1 Quadratic LGCM as the DGM

Chi square test. We found that p values of the DGM and quadratic ALT models were almost equivalent across conditions (ranging from .47 to .52), indicating an equally impressive fit. In other words, the inclusion of additional, unnecessary AR parameters did not hurt the model fit as reflected in the chi square p-value. The discrepancies between the data and other models resulted in large and statistically significant chi square test statistics. One exception was the freed loading ALT model that had acceptable fit with the 6-timepoint model and a moderate sample size (N = 300). Data with more repeated measures or larger sample sizes did not fit as well with the latent-basis ALT model. This suggests that a quadratic growth trajectory with a few waves of data and a moderate sample might be hard to empirically distinguish from the freed loadings ALT model. A large sample size or more waves of data made it easier to distinguish between these two models.

BIC and AIC. We found that BIC favored the DGM and the quadratic ALT models over the saturated model. Different from the equal chi square test p-value results, however, both BIC and AIC further distinguishes the two: the DGM had a significant lower BIC or AIC value than that of the corresponding quadratic ALT model across conditions (where a moderate or larger difference is one that is larger than three; Raftery, 1995). Interestingly, the 6-timepoint linear or latent basis ALT model has a good BIC fit (although poorer than that of the DGM) when sample size is moderate. It seemed that a large sample size is necessary for BIC statistics to distinguish some

short wave ALT model from the saturated model. It is also seen that BIC performed better in distinguishing the true model from the saturated model or in detecting misspecification in the model (e.g., the omitting a quadratic term) with increased sample size. AIC results showed an overall similar trend to that of BIC; however, the absolute AIC values were less directly interpretable. For example, although it looked like the DGM had the best AIC fit, values of the corresponding quadratic and latent-basis ALT were not different. It might be practically challenging to distinguish models that have close AIC values.

CFI and TLI. We found that fit results of CFI and TLI in the same condition were almost identical, and that they both had difficulty distinguishing the DGM from the other models. In addition, the DGM and the quadratic ALT model had identical CFI or TLI results (to two-digits) across all conditions, the fits of other ALT models were also near perfect. The performance showed not much variation among varying sample sizes or parameter sizes. This means it would be difficult to determine the DGM by relying only on the CFI and TLI.

RMSEA and SRMR. RMSEA behaved somewhat similar to the chi square p-value in that no distinction was obvious for the DGM and quadratic ALT models, and that the latent-basis ALT fit acceptably well to data generated by the 6-timepoint DGM. SRMR also showed a similar but slightly poorer performance than other statistics in distinguishing the DGM from alternative models such as linear or latent-basis ALT, particularly in the small-parameter setting or with fewer waves of data. SRMR had the weakest ability to reliably distinguish the true quadratic trajectory from alternative function of trajectories, linear or nonlinear alike.

LRT. By setting the mean, the variance and covariances of the quadratic term with the linear slope and intercept to zero, the quadratic LGCM is equivalent to the linear LGCM and hence these models are nested. Recall that we also obtained a set of LGCM models nested within the quadratic ALT model by treating the first wave of data as predetermined for the purpose of the LRT test. We found that while the LRT test between the DGM with predetermined first measure and the quadratic ALT model was always not significant, the LR difference between the linear LGCM model and the DGM was likely significant, suggesting that the omission of the quadratic term in the LGCM was detectable by the LRT. It also seemed that such differences were more likely to be detected by the LRT in the small parameter set and with fewer the number of waves condition, despite the sample size. It is possible that the quadratic trend would be more easily approximated

by a linear trend when the trajectory is characterized by a slight curvilinearity that is hard to be modeled when stretches out over many time points.

To summarize, we found that if the data were generated by a quadratic LGCM, most fit indexes would indicate an excellent fit for the true model as well as the quadratic ALT model where the former is a special case of the latter model. In comparison, neither the AR, linear LGCM, or latent basis/freed loadings LGCM fit well. Results on the other ALT models were more divergent amongst fit statistics. Only the chi square, BIC, and RMSEA indicated a poor fit for the linear ALT most of the time. The discrepancy between the freed loadings ALT and data generated by the Quadratic LGCM was difficult to detect when data were generated by the small number of waves or with a moderate sample size (N = 500).

3.2.2 Linear ALT as the DGM

Chi square test. According to chi square statistics, all ALT models fit excellently to the data generated in all conditions. As expected, a linear ALT showed no chi square difference from a nonlinear ALT in most cases, except that under the 10-timepoint, large parameter set conditions, the chi square slightly favored the DGM over the latent basis ALT. AR and most LGCM models did not fit well. Interestingly, the p value of the quadratic LGCM indicated an insignificant discrepancy to the data generated by a small-parameter, 6-timepoint ALT model with a moderate sample size. However, the discrepancy was significant once sample size increased to 2000. Apparently, omitting either a strong AR (which reduced to a linear LGCM) or a strong linear growth curve component of a linear ALT (which reduced to an AR model) will end up with a poor chi square fit. However, without sufficient observations (e.g., sample size, number of waves), chi square might not detect the discrepancy between the true linear ALT with a weak, short AR process and a quadratic LGCM where the AR effects was approximated by a quadratic trajectory.

BIC and AIC. As expected, all ALT models showed a better fit than the saturated model, but the nonlinear ALT models fit slightly worse than did the DGM. However, BIC underperformed chi square in distinguishing the DGM from the LGCMs. Specifically, it is only when sample size was as large as 2,000, the DGM showed uniformly the best fit. Otherwise, the nonlinear LGCMs fit well to data generated under the large-parameter conditions (while AR or linear LGCM fit poorly); all LGCMs showed comparable BIC fit to those of the DGM under the small-parameter conditions, linear LGCM had even a slightly better BIC fit than did the DGM when sample size was medium.

This observation that BIC selected a simpler model than the DGM when sample size is not sufficiently large is a little surprising. But given that BIC often favor a more parsimonious model with fewer parameters when fitted to two unsatisfactory models (Raftery, 1995), the bias toward the simpler model over the true model might result from the lack of power (sample size). AIC, in comparison, favored the incorrect AR model as the best fitting model at all times. All ALT models had very similar and worse AIC fit than did the incorrect AR model.

CFI, TLI, RMSEA, and SRMR. The DGM and nonlinear ALT do not differ in terms of their CFI, TLI, RMSEA, or SRMR fit. Nonlinear LGCMs also exhibited good CFI or TLI fit, although slightly worse than did the DGM. Additionally, under the small-parameter situations, all LGCM models including the linear ones fit the data well according to CFI, TLI, RMSEA, and SRMR. The performance of CFI and TLI was invariant across parameter sizes, sample sizes, or number of timepoints, while parameter sizes seemed to affect that of the RMSEA and all of the factors seemed to slightly affect the SRMR.

LRT. The LRT results showed no distinction between the DGM and the quadratic ALT model, while 100% of the time AR or linear LGCM with the predetermined first measure had a significant worse fit than the DGM. In other words, LRT can perfectly detect the omitting of either the linear growth curve beyond the AR process or the AR beyond growth curve. As expected, the inclusion of additional quadratic term of the growth curve did not lead to a meaningful likelihood difference for the more complex model.

As discussed in the introduction, since hybrid types of change processes coexist in a linear ALT, misspecification can happen either in the AR component (e.g., the LGCM models), the growth curve (e.g., AR), or both (freed loadings LGCM). Results were diverse with respect to the type of misspecification being involved. We found that the AR model, which completely omitted the linear trajectory of the ALT, did not fit the data by all measures. As for the linear LGCM, while chi square p values, BIC, and RMSEA indicated a poor fit when fitted to data generated by the large parameter set (suggesting strong AR processes), only a poor chi square fit was observed when under the small parameter set (suggesting moderate AR effects); surprisingly, the other measures returned an acceptable to good fit under such conditions. In terms of the quadratic or freed loadings LGCM, the models where AR was omitted and the trajectory was not linear, chi square was again the only measure that can detect the misspecifications given sufficient sample size; other measures, by

contrast, indicated somewhat acceptable to good fit for the quadratic and freed loadings LGCM.

3.2.3 AR as the DGM

Chi square test. We found all the ALT models showed a good fit to the data generated by the AR model according to the chi square test, with zero estimates of the growth related parameters. Amongst the three, the latent-basis ALT (when it converged) fit the data the best and equally well as did the DGM. In comparison, none of the LGCMs showed an acceptable chi square fit to the data.

BIC and AIC. The DGM AR model had the best BIC fit at all times. The ALT models also returned good BIC or AIC fit. Between the ALT models, the BIC fit of the linear ALT model was better than those of the quadratic and the latent basis ALT models when fitted to the data generated with AR models using moderate lagged parameters. These results suggested that when AR process was strong, the chi square p-value and BIC tend to select the DGM over the more general ALT models; while when the true AR process was weak, there is not a clear preference for the DGM over the ALT model unless a very large sample size was given. However, the misspecification in LGCM, i.e., an omission of AR processes, was easily detected by either chi square or BIC statistics.

CFI, TLI, RMSEA, and SRMR. CFI, TLI, and RMSEA returned very good fit for all ALT models except that the latent-basis ALT did not fit well under one condition, i.e., large-parameter, 10-timepoint, N = 300 (oddly, the corresponding chi square, BIC, AIC suggested a good fit). This suggests that CFI, TLI, and RMSEA did not distinguish an AR from the ALT models. In addition, sample size did not affect the performance of CFI, TLI, or RMSEA as much as it did for chi square, BIC, or AIC. SRMR results differed amongst the ALT models: linear or freed loading ALT models did not fit well across conditions, quadratic ALT did not fit only to data generated by the small-parameter, 10-timepoint model.

LRT. The LRT test between the DGM (AR) and the ALT models were not significant most of the times, suggesting that the discrepancy between the AR and the ALT models were negligible as would be expected given the relationship between these two models. When the DGM is the AR model, our results showed that the fit statistics under evaluation behave similarly in distinguishing the DGM from alternative ones. And we found that such ability is associated with the strength of the AR effects: when strong AR effects were present, the DGM can be distinguished by most fit indexes from other models. When the AR effects were moderate, results on the ALT models were

mixed while the LGCMs led to a poor fit according to the majority of fit statistics.

3.3 Recovery of DGM as the Best Fitting Model

The recovery of the DGM as the best fitting model was defined as the proportion of replications in which the DGM has the optimal fit indicated by each fit statistic. The nonconverged models reported above were excluded from the calculation. To account for trivial, practically meaningless differences, we counted best-fitting models as tied if the top-ranking models had a difference that is less than +/- 2 for BIC or AIC, or less than .01 for other fit statistics. Further, for the cases where the DGM is nested within the fitting model (i.e., models that have all the correct free parameters in the DGM plus some additional free parameters), the selection between the DGM and the more general model does not matter much, as opposed to a misspecified model where either wrong parameters were included or true parameters in the DGM were missing. Indeed, one could argue that in practice, the DGM is equivalent to the more general model with the additional free parameters equal to zero. It is for this reason that we present the results in three categories: 1) the DGM is the only best fitting model (presented as blue bars in the graphs); 2) the DGM only ties with the more general models that contain all the correct free parameters and additional parameters in the situations listed by Table 3 (presented as yellow bars in the graphs) and 3) the DGM ties with models that contain âwrongâ free parameters in which the wrong parameters are defined by those not in the DGM (presented as red bars in the graphs). We perceive that the cumulative percent summed up by the first two categories represents the ability of a fit statistic to select the true model, though ties complicate the selection process.

Follow these guidelines, Figures 1-3 show the results for the best fitting model according to each fit statistics across the 12 simulation conditions (one graph per condition) for the three DGM, respectively. In each figure, the upper two rows are the large-parameter sets and the lower two rows are the small-parameter sets, within which the first row being the 10-timepoint model and the second row the 6-timepoint model, respectively. Each column represents a different sample size condition (from left to right: N = 300, N = 500, N = 2,000). In each graph, the x-axis represents fit statistics, while the y-axis is the cumulative percentage summed from the three categories.

Under these criteria, we found that BIC had the highest probability to select the DGM as the unique best fitting model among all the measures. When the DGM is quadratic LGCM or AR(1), statistics other than BIC also showed a high probability of selecting as the best model either the

DGM or both the DGM together with a model in which it is nested. Chi square p-value and SRMR statistics had the weakest ability for selecting the true quadratic LGCM. When the DGM is linear ALT, however, most statistics besides the BIC had trouble selecting the DGM from other models. AIC performed the worse in this case. Below we present more details for each DGM.

3.3.1 Quadratic LGCM as the DGM

Overall, BIC performed the best among all fit statistics in identifying the DGM as the unique best fitting model; other statistics besides the chi square p-value showed comparable cumulative probability, i.e., in the majority of time they select either only the DGM or both the DGM as well as the quadratic ALT as the best fit (Figure 7). When data were generated with longer waves (10-timepoint vs. 6-timepoint), SRMR always selected the DGM together with wrong models as equal best fits, while when generated with shorter waves, CFI also showed this tendency. Chi square p-value recovered the DGM as the best model only half the time, while the quadratic ALT model was selected as the best fit for the remainder. In practice, this would not necessarily result in a wrong model selection (for the reason we stated before, i.e., the two models can be seen as equivalent). But the fact that the DGM was not selected at the same time indicate the tendency to select an over-parameterized model.

3.3.2 Linear ALT as the DGM

Overall, BIC had the highest unique recovery for the DGM as the best fitting model under the large-parameter conditions, although the performance under the small-parameter conditions varied according to sample size (Figure 8). Chi square p-value underperformed BIC, selecting the DGM only half the time across conditions. Other fit statistics almost never identified the DGM as the sole best fitting model, among which not only the more general quadratic ALT but also other models had equal fit with the DGM. This was particularly likely in the small-parameter condition. Among these remaining fit indexes, the CFI showed the highest recovery rate with ties, followed by RMSEA and TLI. The cumulative recovery rates of these statistics were higher than that of BIC in the small-parameter condition when sample size is moderate (N = 300) or medium (N = 500). The performance of SRMR was poor overall, but the most surprising result was that AIC had zero chance of recovering the DGM as the best fit.

We further examined what alternative models were chosen as the best fit model by different fit statistics if not the DGM. We found that BIC selected the linear LGCM or quadratic LGCM over

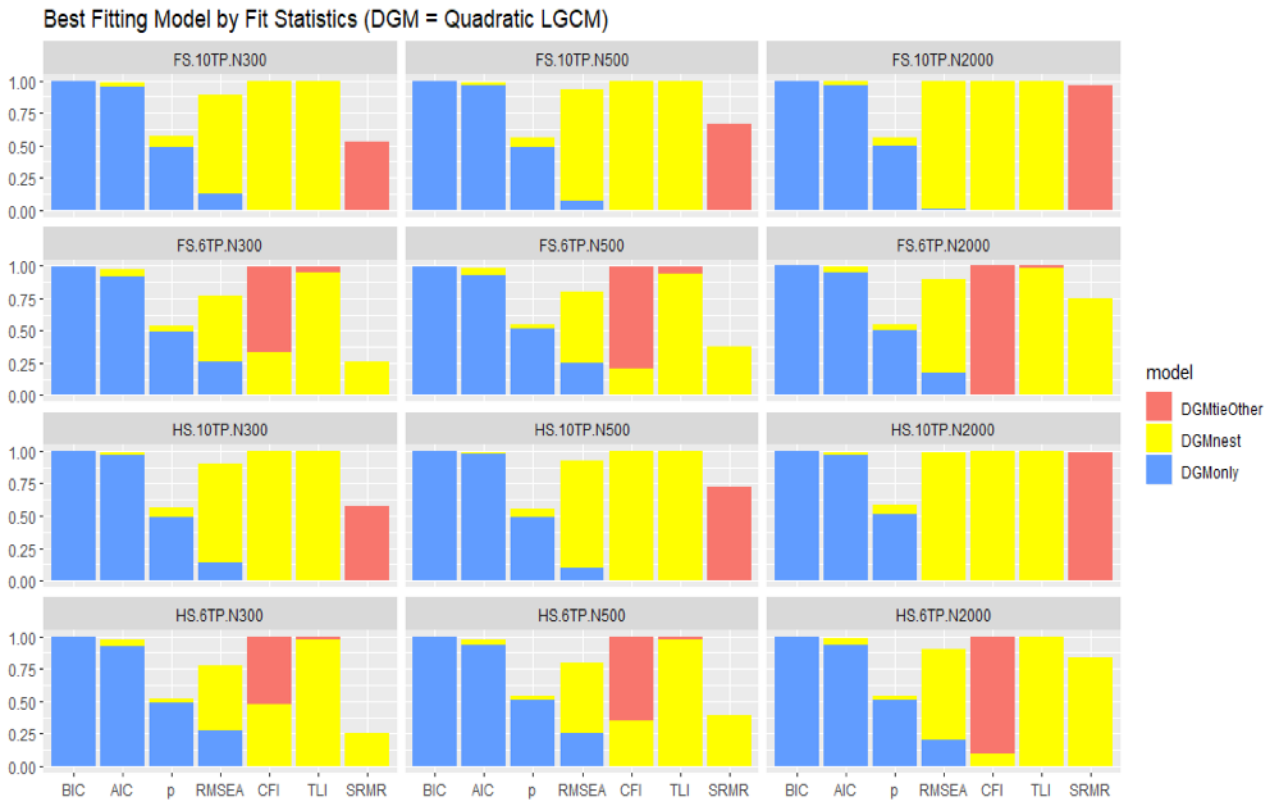

Figure 3.1: Quadratic LGCM: Recovery of Data Generating Model by Fit Statistics

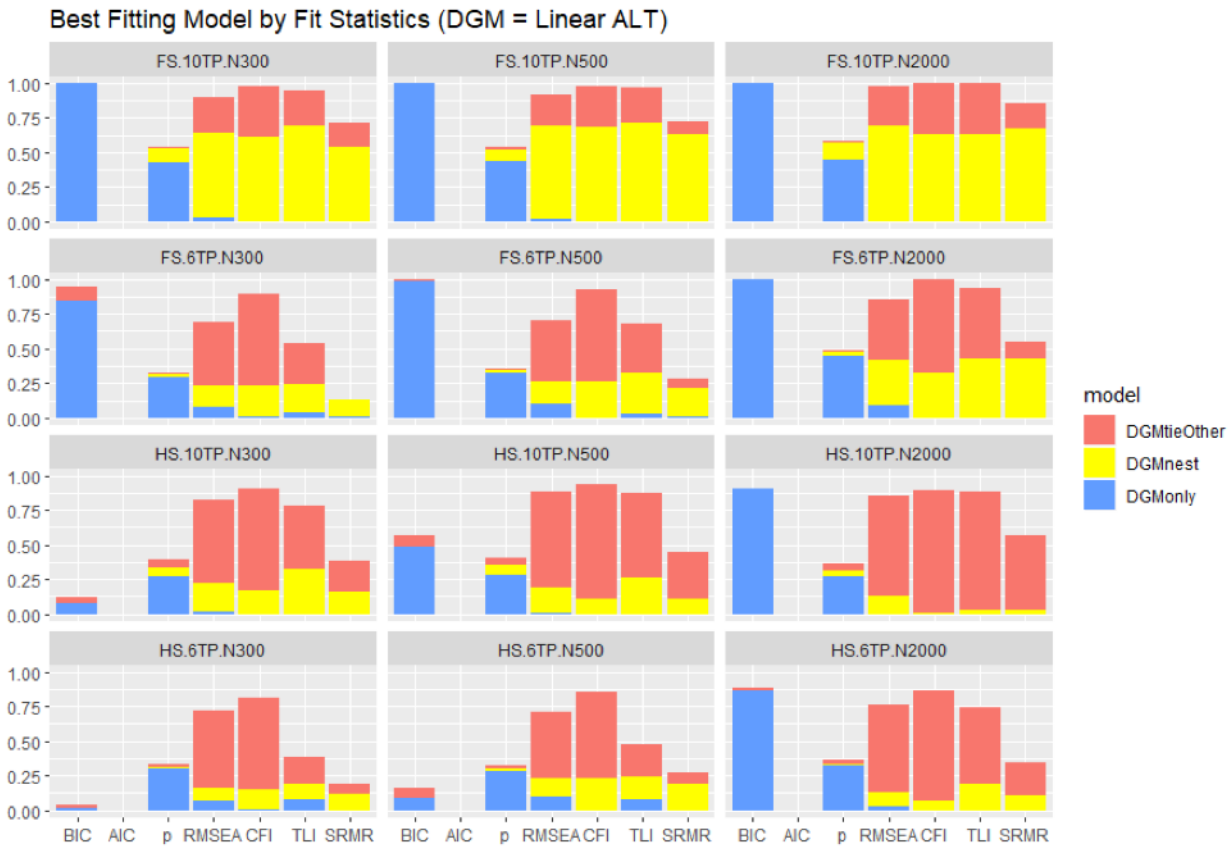

Figure 3.2: Quadratic LGCM: Recovery of Data Generating Model by Fit Statistics

the DGM, while chi square, RMSEA, TLI, CFI most likely chose one of the nonlinear ALT models instead. Interestingly, the AR model had the best AIC fit all the time, which was why we observed zero recovery rate of the DGM by AIC. Results of SRMR was mixed: AR had the best SRMR fit to the 10-timepoint data, while nonlinear ALT had the best SRMR fit to the 6-timepoint data.

3.3.3 AR as the DGM

The overall recovery rate for the AR DGM as the best fitting model was excellent across fit statistics. BIC, AIC can almost perfectly identify the true model as the unique best fit, except in the small-parameter, 10-timepoint data condition where the sample size was moderate or medium. Here BIC tended to select the linear ALT as the best-fit instead. Chi square p-value had the lowest cumulative recovery and performed the worst in the large-parameter, 10-timepoint condition with medium to large sample sizes or in the small-parameter, 6-timepoint condition, in which cases the latent-basis ALT model was mostly likely to be chosen as the best fit, followed by quadratic ALT model. Other fit statistics, particularly CFI and followed by TLI, tended to select AR and the more general model, i.e., one or more ALT model(s), as the equally best fits. A good thing is that no other misspecified models were chosen to have equal best fit. Cumulative recovery rates were very satisfactory. The overall lowest cumulative recovery rates across RMSEA, CFI, TLI, and SRMR were observed under the small-parameter, 6-timepoint condition with moderate or medium sample sizes, when the best fitting model was latent-basis ALT instead. An implication of these finding is that when an AR process is weak and short-spanned, it is more likely to be well resembled by some alternative nonlinear trend so that the discrepancy is hard to be detected by fit indexes.

Figure 3.3: Quadratic LGCM: Recovery of Data Generating Model by Fit Statistics

CHAPTER 4

DISCUSSION

The purpose of this study was to examine the degree to which researchers can use empirical means to choose among competing longitudinal models. Our study bridged a crucial gap in the literature, that little is known about the ease of empirically distinguishing between the different models in analyzing longitudinal data with insufficient theoretical guidance. Indeed, it has been a difficult task for many applied researchers to choose an appropriate longitudinal model to represent the change trajectory over time, because often there are uncertainties in the change pattern and yet a lack of consistent substantive theory or empirical evidence to guide the selection. On the one hand, exploratory approaches are often adopted in the search for an appropriate model, such as visual inspection on the plot of the raw data. However, visualization is informative only under limited circumstances and interpretation is sometimes misleading, particularly when hybrid developmental processes are involved. On the other hand, while analytic research has provided special cases where the models are statistically equivalent, the conditions upon which the equivalency hold are rare in practice, not to mention that statistical equivalence is not the same as substantive equivalence. Therefore, we aimed to provide some empirical guidance to longitudinal research facing these model choice challenges. Particularly, we focused on the examination of several widely used longitudinal models to dictate patterns of change with common functional forms. Our simulation investigation is characterized by two features. To keep our models realistic, we used data generating models that were based on published empirical examples and their estimates. We also varied the type of longitudinal model that was the corresponding true data generating model, but compared among the same set of possible models to determine how well we could distinguish the true from the false models.

Four major research questions that we sought to answer were:

1. Which model fit indices are most reliable in identifying the true DGM?

2. Are there some DGM that are harder to distinguish as true relative to other DGMs?

3. To what degree do characteristics of the parameters of the DGM such as rate of change, curvilinearity, strength of autoregression, the amount of random errors, number of time points, sample size of observations, etc. make it more difficult to determine the correct DGM?

4. Should the search for the DGM begin with a model with too many parameters and trim back or with too few parameters and build up to more complex DGMs?

4.0.1 Most Reliable Fit Statistic(s)

First, we observed some general patterns of the fit statistics with respect to their capability to distinguish the DGM model from models. Overall, BIC is most accurate in selecting the DGM as the best fitting model, regardless of the DGM or the type of misspecification. That is, whether the fitting model is a restrictive form of or nested within the DGM (under-parameterized), or a more general model than the DGM (over-parameterized), or mis-parameterized from the DGM (some wrong parameters are included instead of correct ones). In addition, the chi square test is reliable under most scenarios in identifying significant discrepancies between the data and misspecified models, particularly when trajectory-related parameters are eliminated. In contrast, fit measures such as CFI, TLI, and RMSEA are less helpful in selecting the correct model. They often have a tendency to favor a more complex model than necessary, e.g., select quadratic ALT when the DGM is a quadratic LGCM, and what is more concerning is that they sometimes show equal fit for the DGM and models with some type of misspecification. For example, when the DGM is a hybrid model, i.e., a linear ALT, it is possible they select some LGCMs as the best fitting model, particularly given certain conditions. In general, SRMR has the weakest ability to select the true model from alternatives. Last but not least, LRT is particularly reliable in detecting under-parameterized misspecifications (i.e., the omitting of correct parameters) and eliminating models that are nested within the true model. Therefore, when it comes to which fit index to aid for the longitudinal model selection, we recommended to use a combination of fit statistics with primary attention given to BIC, chi square test, as well as LRT when there exists a nesting relation among the candidate models.

4.0.2 Most Detectable DGM

The AR model is the most easily detected longitudinal model across all fit measures. Information criteria (i.e., AIC and BIC) consistently detect the AR when it is the true DGM.

Although CFI, TLI, RMSEA, and SRMR might identify both the true AR model and some more general (over-parameterized) ALT models as the equal best fits, this is less concerning because the latter is not a âwrongâ model in that the unnecessary parameters are not significantly different from zero.

Similar behaviors amongst fit statistics are observed for the identification of a true quadratic LGCM. However, we found that challenge arises for all fit measures in the distinction of the DGM from the freed loadings ALT or the freed loadings LGCM, particularly under those challenging situations. This suggested that the freed loadings models can provide a satisfactorily approximation to a short-wave quadratic LGCM so closely that the discrepancy between the substitute trajectory and the true model-implied trajectory is not detectable. The inclusion of the incorrect parameters (i.e., the AR parameters from the freed loadings ALT) in substitute for the missing slope parameters makes the distinction between the true and alternative models particularly difficult. Statistically, part of the residual variances in the incorrect model might be âaccidentallyâ explained by the additional âwrongâ parameters, this is particularly possible when the sample size is small and sampling fluctuation is high. We shall discuss what other characteristics of the model are involved with greater details next.

It is increasingly difficult to select the correct model from the alternatives using fit index when data was simulated from more complex DGM model. We found that results amongst the fit statistics were most divergent when the DGM is a linear ALT, a complex model where both autoregressive process and growth trajectory are present. In a hybrid model like the ALT, there could be multiple sources of misspecification, not all of them are equally difficult to be detected. For example, the AR models showed a consistent poor fit due to the omission of the true linear growth trend. In comparison, even though the LGCMs lack the AR parameters, it is not always easily distinguished from the true ALT. This is particularly the case when quadratic LGCM or the freed loadings LGCM were chosen. In these cases, chi square is the most sensitive statistics to detect the misspecification in the alternative models. However, (with a little surprise) we found that BIC sometimes was in favor of a model simpler than the DGM, i.e., the linear LGCM. In comparison, other measures opt for the LGCMs with a quadratic term or freed loadings. The selection of wrong model(s) is most likely observed under challenging situations in the DGM such as a large random error and/or a small sample size.

4.0.3 Important Conditions Affecting Detection of DGM

Turning next to the characteristics of model and data that might be associated with the distinction, we found that each of the factors under investigation plays an important role. For instance, larger sample size or more waves of observation increased the likelihood of the chi square test or BIC detecting the discrepancy between the DGM and alternative models. A combination of difficult situations might largely decrease the chance of selecting the correct model from some alternatives. For example, the discrepancy between a true quadratic LGCM is difficult to be distinguished from a freed loadings ALT or a freed loadings LGCM using the chi square test or BIC under conditions such as a small number of waves, or a moderate sample size, or large amount of random error in the DGM. Similarly, without sufficient observations (e.g., sample size, number of waves), chi square might not distinguish a true short-wave linear ALT with a moderate AR process from a quadratic LGCM. The presence of nontrivial random error from the DGM, and the sampling error from the data due to small sample size or short waves, seem to be a deciding factor in the difficulty of identify the true model, or the ease to approximate it with alternative functional forms. For example, when one or both are met, it is very likely to find a quadratic trajectory that can provide a close approximation to the combined pattern of the AR processes and the linear trajectory from an ALT model. Indeed, on the one hand, some models that have complex structures might be empirically hard to distinguish or be easily resembled by alternative forms. On the other hand, increased observations â a sufficiently large sample size and more waves of data â will facilitate the distinction between these complicated models.

4.0.4 Search Strategies

Altogether, our result suggest that when theory is lacking it is a good practice to start with the most general (identifiable) model (e.g., the quadratic ATL model). To reduce it to a more parsimonious model, sets of parameters can be eliminated progressively as long as the goodness of fit does not drop substantially. It is likely that one might end up choosing a more general model than necessary, because our findings indicate that the correct model is equivalent to the restricted form of the general model in which the additional or unnecessary parameters should have insignificant estimates. Hence, the interpretation would essentially be the same whichever model is chosen. However, if one is concerned about obtaining an overly complex model, it is suggested that BIC result should be given special attention because our results show that it is the least likely to

favor an over-parameterized model.

Nevertheless, challenges might appear at different stages of this stepwise model selection fashion regarding the following aspects: 1) the most general model might not converge; 2) if the optimal model is not the most general one, what would be an alternative option for the starting model; and 3) how can we increase our chances of stopping at the correct model instead of a close approximation that carries a wrong substantive interpretation. Answers to these questions may differ on a case-by-case basis, but here we offer some general suggestions based on our findings.

First, when nonconvergence occurs when fitting the most general model, it could result from too few waves of data or insufficient sample size. Researchers can implement common practices to enhance convergence such as inputting starting values, increasing the number of iterations, or switching to less conservative convergence criteria. However, if convergence remains a problem, it is possible that the model is more complex than necessary. With careful examination, one might then move forward with simpler models such as the AR or LGC models as a starting point. In the meantime, we recommend the BIC and Chi square test to compare the fit of the more restricted model with the previous model at each step. When we are reducing from a more general model to its restrictive form (e.g., from quadratic LGCM to linear LGCM), LRT should also be adopted. We proceed until most of the fit index indicate a poor fit to a model, and we return to the previous well-fit model as the optimal model.

One of the most challenging cases is when more than one model with no nesting relations indicate comparable good fit, e.g., the LGCMs with the AR model, or quadratic LGCM with the linear ALT. Because these models are nonnested, we cannot compare the relative fit using LRT. From our results, it seemed like the distinction between the LGCMs with the AR model should be easier if we look at all the fit indexes. However, for the distinction of quadratic LGCM with the linear ALT, we might be better off to look closely at BIC and the chi square. As shown from our study, while BIC might tend to favor a simpler model, chi square could be very sensitive to misspecifications given sufficient sample sizes. It is therefore a good practice to find a tradeoff of the two or more fit references whenever necessary. Admittedly, we imagine there might be some other unforeseeable empirical challenges that adds to the complexity. Model selection without substantive theory is always a dauting task. But our results at least suggest that even without any hypothesis, some fit indexes including BIC and the chi square test are useful more often than not.

To sum up, our evaluation emphasizes the following highlights: BIC is the most reliable measure in selecting the optimal longitudinal model. The chi square test is also sensitive to detect misspecifications and works particularly well to identify a true linear ALT model. LRT is most helpful only when the candidate models have some nesting relations. Goodness-of-fit results work most consistently when distinguishing a true AR model from false models. The most challenging distinctions are between short-wave nonlinear LGCMs and ALT models when sample size is moderate. Larger sample size or more waves of observation increased the likelihood to detect the discrepancy between the DGM and the alternative models. It is always a good empirical practice to start with the most general model (among those that can be identified and converged), and reduce it to more restrictive forms until one reach the most parsimonious representation with roughly or very consistent goodness fit results.

4.1 Limitation and Future Directions

We note that there are several limitations of our study. As with all simulations, the research design can affect the results. First, we only used a subset of longitudinal models as the DGMs instead of all. While there might be some value to include the additional sets of the DGMs using the remaining four candidate models, we did not given the time constraints of this project. Second, the list of the seven candidate models does not exhaust all longitudinal models. For instance, we did not use alternative hybrid models such as the Latent Change Score model or the State Trait models, as it has been shown that they are special cases of the more general Latent Variable ALT model (Bianconcini & Bollen, 2018). We suspect that some of our findings of how fit statistics work can be generalized given the same type of longitudinal models. The major difference lies in the parameter estimates and interpretations, which might differ as a result of how components of change are specified in the models. Although less common, LGCMs or ALT models with even higher order polynomial terms have been documented in literature. Researchers might opt for these models if the hypothesis on the shape of curvilinearity suggested so, and that the number of waves and sample size of the data were sufficient enough to allow for the inclusion of complicated nonlinear terms. While the general message to start with the most flexible (identifiable) model and proceed to simpler ones by adding restrictions, the ease of distinguishing the true from false using fit indexes is unknown particularly amongst those with no nesting relations.

Another limitation is that our simulation covers only a certain levels of random factors. For

example, the number of timepoints and sample size in the moderate conditions and the portion of missingness (i.e., 6 waves, N = 300, missing = 15%, respectively) were chosen to meet the minimum requirement of model identification and convergence in the most complex models (i.e., the quadratic ALT or the freed loadings ALT). However, empirical longitudinal studies sometimes face even tougher situations such as only a handful waves of data less than 6, a smaller number of observations than 300 or with larger or even increasing portion of missingness (i.e., due to attrition). If the data were indeed even more varied than the moderate conditions considered in the simulation, generalizing from our results to these conditions might be unwarranted. Some models under investigation might not be applicable or identifiable in these scenarios, and the starting model has to be downsized to a more restrictive representation. Even with identified models, sampling fluctuations are likely larger so that the distinction between candidate models is made more difficult. That is, finding more than one close approximations to the true underlying change patterns could be more likely. As it does in practice, nonconvergence is another issue to consider in our study. The fact that not all the models with simulated data converged raises issues about how best to report the results. Following a common practice, we eliminate all replications that did not converge and our success rate was based on using only replications that converged. If we counted nonconverged replications as failures, then the success rates would be lower, especially under those conditions when nonconvergence was relatively frequent.

Finally, another compromise comes from the adoption of an empirical simulation approach. We are aware that with a selective set of the parameters, the model implied trajectories represent specific change patterns. However, many characteristics of the trajectory might vary, sometimes largely, across different substantive contexts, e.g., rate of change, size of curvilinearity, magnitude of association between the two, etc. The uncertainty is that we do not know to what extent these characteristic affect the distinguishability of the fit statistics. The data generated by the DGMs have certain structures such as the scales of the measures, strength in the correlations, magnitude of random errors, etc. It is difficult to know how much our results generalize under new conditions. However, we did vary several factors including incorporating missingness. Particularly, to avoid the limitation to specific change patterns or trajectories, we repeated each DGM by using the original parameter set from the selected studies as well as the half sized parameter set (except for random errors). We believe this strategy help to make the results more applicable. Last but not least, some

topics are related but are out of the scope of the current investigation. For instance, the present study does not consider the choice of the starting model and model selection with or without covariates. Further, our study does not account for measurement error. For the simplicity, all the evaluations are conducted on observed variables, assuming no measurement error for each variable as well as measurement invariance in the validity of repeated measures over time. Future studies are needed to examine these areas.

BIBLIOGRAPHY

Akaike, H. (1987). Factor analysis and aic. *Psychometrika*, *52*(3), 317–332. https://doi.org/10.1007/BF02294359

Anderson, J., & Gerbing, D. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*(2), 155–173. https://doi.org/10.1007/BF02294170

Anderson, T. W. (1960). Some stochastic process models for intelligence test scores. In S. K. K. J. Arrow & P. Suppes (Eds.), *Mathematical thinking in the social sciences* (pp. 205–20). Stanford University Press.

Azzalini, A. (1987). Growth curves analysis for patterned covariance matrices. In M. L. Puri & J. P. Vilaplana (Eds.), *New perspectives in theoretical and applied statistics*. John Wiley & Sons Inc.

Bauldry, S., & Bollen, K. A. (2018). Nonlinear autoregressive latent trajectory models. *Sociological Methodology*, *48*(1), 269–302. https://doi.org/10.1177/0081175018789441 doi: 10.1177/0081175018789441

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, *88*(3), 588–606.

Bianconcini, S. (2012). Nonlinear and quasi-simplex patterns in latent growth models. *Multivariate Behavioral Research*, *47*(1), 88–114. https://doi.org/10.1080/00273171.2012.640598

Bianconcini, S., & Bollen, K. A. (2018). The latent variable-autoregressive latent trajectory model: A general framework for longitudinal data analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 791–808. https://doi.org/10.1080/10705511.2018.1426467 doi: 10.1080/10705511.2018.1426467

Blozis, S. A. (2007). On fitting nonlinear latent curve models to multiple variables measured longitudinally. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(2), 179–201.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social forces; a scientific medium of social study and interpretation*, *89*(1), 1–34. https://doi.org/10.1353/sof.2010.0072

Bollen, K. A. (1989). Structural equations with latent variables wiley. *New York*.

Bollen, K. A., & Curran, P. J. (1999). An autoregressive latent trajectory (alt) model: A synthesis of two traditions. *Paper presented at the 1999 Meeting of the Psychometric Society June, 1999 Lawrence, KS*.

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (alt) models: A synthesis of two traditions. *Sociological Methods & Research*, *32*(3), 336–383.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). John Wiley & Sons.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing illode1 fit. In K. A. Bollen & J. S. Long (Eds.), *Testingstructural equation models* (pp. 132–162). Sage.

Chi, E. M., & Reinsel, G. C. (1989). Models of longitudinal data with random effects and ar(1) errors. *Journal of the American Statistical Association*, *84*(406), 452–459.

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods*, *10*(1), 3–20.

Cudeck, R., & Harring, J. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual review of psychology*, *58*, 615–37. https://doi.org/10.1146/annurev.psych.58.110405.085520

Curran, P. J., & Bollen, K. A. (1999). Extensions of the autoregressive latent trajectory model: Explanatory variables and multiple group analysis. *Meeting of the Society for Prevention Research June*.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 105–136). American Psychological Association.

Curran, P. J., Howard, A. L., Bainter, S. A., 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.

Diggle, P., Liang, K.-Y., & Zeger, S. L. (1994). Longitudinal data analysis. *New York: Oxford University Press*, *5*.

Duncan, O. D. (1969). Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin*, *72*(3), 177–182.

Duncan, T. E., Duncan, S. C., & Stoolmiller, M. (1994). Modeling developmental processes using latent growth structural equation methodology. *Applied Psychological Measurement*, *18*(4), 343–354. https://doi.org/10.1177/014662169401800405

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. Routledge Academic.

du Toit, S. H. C., & Browne, M. W. (2007). Structural equation modeling of multivariate time series [PMID: 26821077]. *Multivariate Behavioral Research*, *42*(1), 67–101. https://doi.org/10.1080/00273170701340953

Ghisletta, P., & McArdle, J. J. (2001). Latent growth curve analyses of the development of height. *Structural Equation Modeling*, *8*(4), 531–555.

Goldstein, H., Healy, M. J., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in medicine*, *13*(16), 1643–1655.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 197–218). Columbia University Press.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton university press Princeton, NJ.

Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course*. Charlotte, NC: Information Age Publishing.

Hansen, N., Rinne, L., Jordan, N., Ye, A., Resnick, I., & Rodrigues, J. (2017). Co-development of fraction magnitude knowledge and mathematics achievement from fourth through sixth grade. *Learning and Individual Differences*, *60*, 18. https://doi.org/10.1016/j.lindif.2017.10.005

Harvey, A. (1991). *Forecasting, structural time series models and the kalman filter*. Cambridge University Press. https://EconPapers.repec.org/RePEc:cup:cbooks:9780521405737

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

Humphreys, L. G. (1960). Investigations of the simplex. *Psychometrika*, *25*(4), 313–23.

Jong, P. D. (1991). The diffuse kalman filter. *Annals of Statistics*, *19*(2), 1073–1083. https://doi.org/10.1214/aos/1176348139

Jongerling, J., & Hamaker, E. L. (2011). On the trajectories of the predetermined alt model: What are we really modeling? *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 370–382.

Kenny, D., & Zautra, A. (2001). Trait-state models for longitudinal data. In L. Collins & J. L. Horn (Eds.), *New methods for the analysis of change* (pp. 243–263). American Psychological Association.

Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis*. Academic Press.

Mandys, F., Dolan, C. V., & Molenaar, P. C. M. (1994). Two aspects of the simplex model: Goodness of fit to linear growth curve structures and the analysis of mean trends. *Journal of Educational and Behavioral Statistics*, *19*(3), 201–215. http://www.jstor.org/stable/1165294

McArdle, J. (2001). A latent difference score approach to longitudinal dynamic analysis. In R. Cudeck, K. G. Jöreskog, D. Sörbom, & S. Du Toit (Eds.), *Structural equation modeling: Present and future: A festschrift in honor of karl jöreskog* (pp. 342–380). Scientific Software International.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual review of psychology*, *60*, 577–605.

Meredith, W. (1984). On" tuckerizing" curves. *the annual meeting of the Psychometric Society, Santa Barbara, CA, 1984*.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122.

Ou, L., Chow, S.-M., Ji, L., & Molenaar, P. C. (2017). (re) evaluating the implications of the autoregressive latent trajectory model through likelihood ratio tests of its initial conditions. *Multivariate behavioral research*, *52*(2), 178–199.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Sage.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111–163.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Rogosa, D., & Willett, J. B. (1985). Satisfying a simplex structure is simpler than it should be. *Journal of Educational Statistics*, *10*(2), 99–107.

Selig, J., & Little, T. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach [PMID: 26794479]. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Sterba, S. K. (2014). Fitting nonlinear latent growth curve models with individually varying time points. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 630–647. https://doi.org/10.1080/10705511.2014.919828 doi: 10.1080/10705511.2014.919828

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*.

Steyer, R., & Schmitt, T. (1994). The theory of confounding and its application in causal modeling with latent variables. In A. E. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 36–67). Sage Publications, Inc.

Tucker, L., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://EconPapers.repec.org/RePEc:spr:psycho:v:38:y:1973:i:1:p:1-10

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT press.

Zyphur, M. J., Chaturvedi, S., & Arvey, R. D. (2008). Job performance over time is a function of latent trajectories and previous performance. *Journal of Applied Psychology*, *93*(1), 217–24.