Modeling Construct Change Over Time Amidst Potential Changes in Construct Measurement: A Longitudinal Moderated Factor Analysis Approach

Siyuan Marco Chen and Daniel J. Bauer

University of North Carolina at Chapel Hill

Author Note

Siyuan Marco Chen https://orcid.org/0000-0002-3346-5424 and Daniel J. Bauer, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill.

This research uses data from Add Health, funded by grant P01 HD31921 (Harris) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Add Health is currently directed by Robert A. Hummer and funded by the National Institute on Aging cooperative agreements U01 AG071448 (Hummer) and U01AG071450 (Aiello and Hummer) at the University of North Carolina at Chapel Hill. Add Health was designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill.

Analysis code used in this study is available in Supplemental Materials. An early version of this study was presented at Society of Multivariate Experimental Psychology Annual Meeting, Monterey, California, with its abstract published on *Multivariate Behavior Research*. Correspondence concerning this article should be addressed to Siyuan Marco Chen, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, 235 E. Cameron Avenue, Chapel Hill, NC 27599-3270. E-mail: chen.8358@osu.edu

Abstract

In analyzing longitudinal data with growth curve models, a critical assumption is that changes in the observed measures reflect construct change and not changes in the manifestation of the construct over time. However, growth curve models are often fit to a repeated measure constructed as a sum or mean of scale items, making an implicit assumption of constancy of measurement. This practice risks confounding actual construct change with changes in measurement (i.e., differential item functioning; DIF), threatening the validity of conclusions. An improved method that avoids such confounding is the second-order growth curve model (SGC). It specifies a measurement model at each occasion of measurement that can be evaluated for invariance over time. The applicability of SGC is hindered by key limitations: (1) SGC treats time as continuous when modeling construct growth but as discrete when modeling measurement, reducing interpretability and parsimony; (2) the evaluation of DIF becomes increasingly error-prone given multiple timepoints and groups; (3) DIF associated with continuous covariates is difficult to incorporate. Drawing on moderated nonlinear factor analysis (MNLFA), we propose an alternative approach that provides a parsimonious framework for including many timepoints and DIF from different types of covariates. We implement this model through Bayesian estimation, allowing for incorporation of regularizing priors to facilitate efficient evaluation of DIF. We demonstrate a two-step workflow of measurement evaluation and growth modeling, with an empirical example examining changes in adolescent delinquency over time.

Keywords: growth curve modeling, Bayesian regularization, measurement invariance, differential item functioning, integrative data analysis

Modeling Construct Change Over Time Amidst Potential Changes in Construct Measurement: A Longitudinal Moderated Factor Analysis Approach

The measurement of change over time has long been seen as a challenging task within psychological research settings. The difficulty is that one must assume constancy of measurement for differences over time to be interpreted as true construct change. In particular, to be able to infer increases or decreases on the construct, the measurement scale must maintain constant calibration. Golembiewski et al. (1976) refers to this as "alpha change", stating that it "involves a variation in the level of some existential state, given a constantly calibrated measuring instrument related to a constant conceptual domain." This means variation in observed scores corresponds entirely to variation in the level of a construct and none from a recalibration of the measuring instrument, i.e., when there is measurement invariance (Chan, 1998; Liu et al., 2017; Millsap & Hartog, 1988). Alpha change is implicitly invoked when applying conventional growth models, whether to single measures like cortisol levels or body mass index or, more commonly, to composite indices like a sum or mean of scale items, yet this assumption of constant measurement is seldom evaluated, leading Bereiter (1963) to raise the important question:

> When scores on a test are observed to change, how can one tell whether it is the persons who have changed or the tests? Once it is allowed that the pretest and posttest measure different things, it becomes embarrassing to talk about change. There seems no longer any answer to the question, change on *what*? (p.11)

Among other concerns, this issue led Cronbach and Furby (1970) to famously disavow the use of change scores, which implicitly assume alpha change, in favor of regressing posttest scores on pretest measures. This "residualized change" approach avoids the assumption of alpha change, but it neither provides an intuitive measure of construct change nor a good match to many theoretical models of construct change over time (Willett, 1997).

One potential solution to the problem arises when one has access to multiple measures of the construct of interest, such as a set of scale items. Then it becomes possible

to model the measurement of the construct directly and to determine if any items display changing measurement properties over time, a condition referred to as *differential item functioning* (DIF). In practice, this is done by specifying a multiple-indicator latent factor at each time point, often while also modeling changes in the latent factor over time via a growth model. McArdle (1988) referred to this as a curve-of-factors model. Others, generalizing language from hierarchical factor analysis, have referred to it as a second-order growth curve (SGC) model (Duncan & Duncan, 1996; Hancock et al., 2001; Sayer & Cumsille, 2001).

Within the SGC, it is possible to apply well-developed psychometric techniques to evaluate the assumption of measurement invariance (Chan, 1998).[1]One can test for whether the set of items shows invariance of measurement over time (or in Bereiter's words, constant calibration) and, if not, try to identify which items among the set show DIF. Even in the presence of DIF, it is often still possible to infer true construct change over time by accounting for this differential measurement in the model, so long as at least partial invariance is achieved (Byrne et al., 1989; Reise et al., 1993). To meet partial invariance, a subset of items must not show DIF. These non-DIF items are referred to as "anchors" because they equate the factor over time, preventing any drift in the scale calibration and making changes in scale scores across assessments comparable. Further, although we have so far been focused on the calibration of our measures over time, by placing this issue within the broader context of measurement invariance, we may also ask whether our measurements are equally calibrated for people of different backgrounds (Mellenbergh, 1989). Using a multiple-group framework (Jöreskog, 1971; Sörbom, 1974) with the SGC, one can also assess DIF with respect to sex, race, ethnicity, or other person-level grouping variables that might impact measurement over time (F. F. Chen et al., 2005; Kim & Willson, 2014; Leite, 2007; McArdle & Nesselroade, 2003).

---

[1] Another advantage of the SGC is that measurement error in the items can be separated from true construct variance. This feature of the model helps to overcome another critique of Cronbach and Furby

Despite their many advantages, SGCs also have several limitations. Many of these stem from the fact that the measurement model is specified in "snapshots" of discrete time, with each unique time point represented through a distinct multiple-indicator latent factor, contrasting the continuous-time representation of the growth process. This mismatch in how the temporal process is represented within the model is conceptually jarring and problematic for model estimation and the testing of DIF. As we describe in more detail later, the SGC becomes more difficult to estimate as the number of timepoints grows, the specification and testing of DIF can be cumbersome and error-prone, and the parameterization of the model does not easily allow for consideration of DIF as a function of multiple or continuous person-level characteristics (it is practically confined to a single grouping variable in most instances).

The current study therefore proposes a new longitudinal modeling framework for modeling construct change amidst potential changes in measurement over time. This new framework brings together three recent developments. First, we use a moderated nonlinear factor analysis (MNLFA; Bauer, 2017; Bauer & Hussong, 2009) specification to represent the measurement model in continuous time. This allows us to naturally model and test DIF over time and also across person-level variables. Second, we leverage advances in Bayesian estimation of latent variables models (Bainter, 2017; Depaoli & Clifton, 2015; Fox & Glas, 2001; Muthén & Asparouhov, 2012) to facilitate fitting the longitudinal MNLFA model. Previously, estimation of the MNLFA via maximum likelihood was limited to assessments made at a single time point or cross-sectionally (e.g., Bauer, 2017; Curran et al., 2014; Kolbe et al., 2022; Stevens et al., 2022). Third, we draw on recent literature proposing to use regularization methods to facilitate the identification of DIF in complex models (e.g., Bauer et al., 2019; Belzak & Bauer, 2020; Huang, 2018; Magis et al., 2015; Tutz & Schauberger, 2015). In particular, we draw on Bayesian regularization approaches (e.g., Hans, 2009; Leng et al., 2014; Lykou & Ntzoufras, 2013; Park & Casella,

(1970), that raw change scores have low reliability.

2008) proposed for MNLFA by S. M. Chen et al. (2022) and Brandt et al. (2023). Brought together, these three developments permit specification and estimation of a longitudinal MNLFA that models growth in the underlying latent construct while simultaneously evaluating and accommodating potential differences in construct measurement over time or people. It thus provides a way to answer the fundamental question posed by Bereiter (1963), "How can one tell whether it is the persons who have changed or the tests?" to which we may add the third possibility "or both?"

In the following sections we first briefly review standard first-order growth models. We then introduce longitudinal measurement models, upon which we can extend to second-order growth models. We then detail the limitations of these approaches as motivation for the longitudinal MNLFA approach. Next, we present the parameterization of the longitudinal MNLFA, showing how this approach addresses estimation and specification issues with existing second-order growth models. We then describe how Bayesian estimation and regularization is used to provide efficient tests of measurement differences (DIF over time and persons). Finally, we demonstrate how the model can be implemented with an analysis of longitudinal data on delinquent behaviors over adolescence. We provide code and instructions on model fitting in Online Supplemental Materials. Throughout, we focus on the common case of binary item-level data (extensions to ordered-categorical items are straightforward).

Growth Curve Modeling with Binary Items

Growth curve models are used to identify individual differences in trajectories of change over time (Bollen & Curran, 2006; Meredith & Tisak, 1990; Willett & Sayer, 1994). Interest may center on the average trajectory, the extent of between-person differences in trajectories of change over time, or in the prediction of these individual differences. In fitting these models, one must first specify a function to describe each individual trajectory over continuous time. Many functions can be considered, but for ease of exposition we shall focus on the linear growth model, which implies straight-line individual trajectories. We

shall also assume that the construct of interest has been measured by a series of binary items, for instance, a behavior or symptom checklist. The critical question underlying the current manuscript is how these items are used to infer construct change over time.

First-Order Growth Models

When collecting item level data, researchers often eschew a formal measurement model in favor of generating scale score via a simple algorithm like taking the sum or mean of the set of binary items (the mean score being the proportion of items scored one versus zero and the sum score being the count of items "endorsed"). A growth model is then fit to these scale scores without further attention to the individual items from which the scores were generated. Though this model is often simply called a growth model, we shall use the term "first-order growth model" to contrast it with the models to come. For mean scores, we can express a first-order linear growth model as follows, using the notation y_{tp} to indicate the scale score obtained from a set of I binary items at measurement occasion t for person p, where $t = 1 \dots T_p$ is the number of measurement occasions or timepoints assessed for person p, and $p = 1 \dots N$ is the number of subjects who provided observations of scale scores.

$$y_{tp} = \beta_{0p} + \beta_{1p} time_{tp} + e_{tp},\ e_{tp} \sim N(0, \sigma^2) \tag{1}$$

where β_{0p} and β_{1p} are the person-specific intercepts and slopes describing straight-line growth for person p, and e_{tp} is a time-specific residual that we assume for simplicity to be normally distributed, independent, and homoscedastic (though each of these assumptions can be relaxed).

The variable $time_{tp}$ indicates the value of time on the chosen time metric, such as the age of the individual. Often the values of time are the same across participants and spaced evenly across occasions of measurement (e.g., everyone is 6 years old at Wave 1, then 7 at Wave 2, 8 at Wave 3, etc.), such that the time scores for everyone can be set to consecutive integer values $0, 1, 2, \dots, T_p - 1$. With the initial time score set to zero, the intercept β_{0p} can be interpreted as the individuals' construct level at the first time point

(e.g., age 6 status). In other cases the numerical value of the time will differ across persons at a given observation number, such as when subjects enter the study with different initial ages. For example, suppose cohorts who are initially in grade 7, 8, and 9 are followed for 3 school years. Then the subscript for the observation is $t = 1, 2, 3$, but the time variable $time_{tp}$ is scored $0, 1, 2$ for the grade 7 cohort, $1, 2, 3$ for the grade 8 cohort, and $2, 3, 4$ for the grade 9 cohort (Duncan & Duncan, 2004). Alternative coding schemes are also available that yield different interpretations for the intercept (Biesanz et al., 2004). The slope β_{1p} captures the rate of change over time, which is assumed to be maintained throughout the observed time interval under the assumption of linear growth. Though we do not pursue this here, extensions to nonlinear growth trajectories are also possible, for example, by adding a quadratic effect to the model with the additional inclusion of $time_{tp}^2$.

A second set of equations decompose the individual intercepts and slopes into their average values across persons and individual-specific deviations from these averages as follows

$$\begin{aligned} \beta_{0p} &= \alpha_0 + \zeta_{0p} \\ \beta_{1p} &= \alpha_1 + \zeta_{1p} \end{aligned} \tag{2}$$

$$\begin{bmatrix} \zeta_{0p} \\ \zeta_{1p} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \psi_{00} & \\ \psi_{10} & \psi_{11} \end{bmatrix} \right) \tag{3}$$

Here α_0 and α_1 are the across-person average intercept and slope values, and ζ_{0p} and ζ_{1p} are person-specific deviations from these averages (sometimes described as *random effects*). The intercept and slope deviations are assumed to be independent across persons (so that one person's intercept or slope is independent from another person's) and identically normally distributed; these deviations have a variance of ψ_{00} for individual intercepts and ψ_{11} for individual slopes, with a covariance ψ_{10} for intercepts and slopes within each person. Substantively, the variances indicate the extent of inter-individual differences in the intercepts and slopes, while the covariance ψ_{10} describes how they are related, for instance, how rate of change is related to initial status.

This linear growth model can be viewed as a two-factor confirmatory factor analysis (CFA) model with mean structure (McArdle, 1988; Meredith, 1991; Meredith & Tisak, 1990; Willett & Sayer, 1994). In this CFA the indicator vector includes the person's observed scale score from each unique timepoint. The scale scores are loaded onto latent factors defined to represent the intercept and slope of the person's trajectory. The factor loading matrix includes a column of 1s to define the intercept factor and a column of time scores to define the slope factor. That is, the time scores from Equation (1) are encoded into the column of the loading matrix corresponding to the latent β_{1p} factor. Figure 1 shows a path diagram of the first-order growth curve model. Often, Equation (2) is

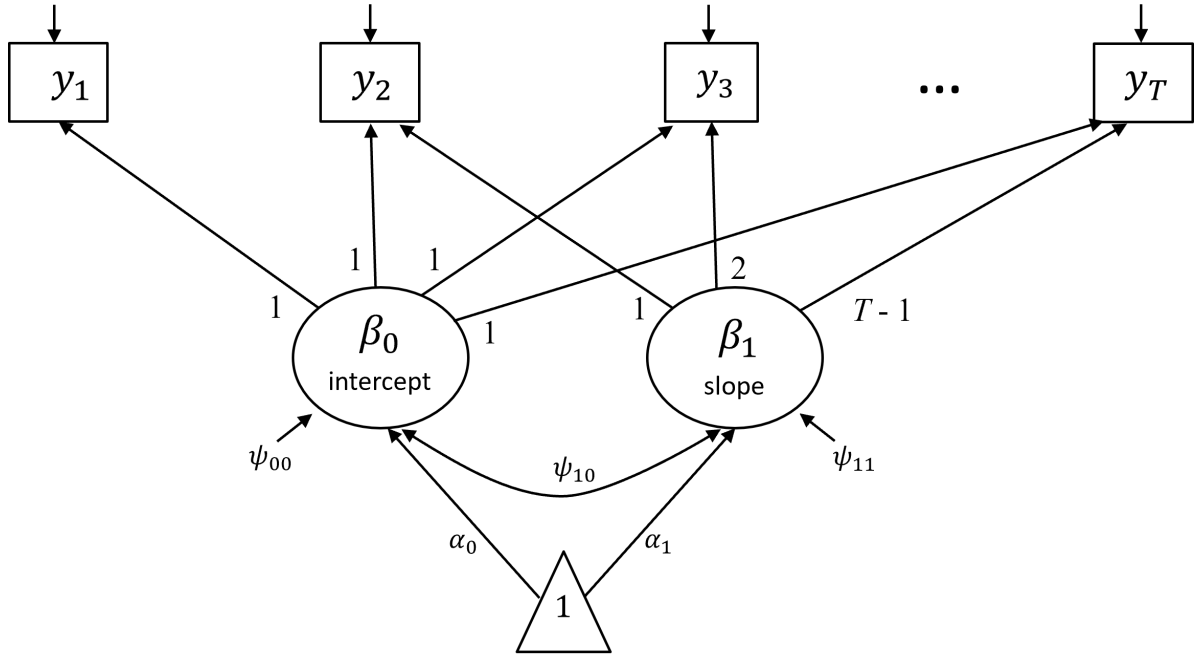


Figure 1

Path Diagram of A First-Order Growth Curve Model. Factor loadings for the slope factor go from 0 *to* $T-1$*, defining the intercept factor to be the starting point of the latent trajectory at the first time point. Person subscript p is omitted.*

augmented to include predictors of the latent factors in an effort to capture sources of individual differences in growth over time, moving from a CFA with mean structure to a

full Structural Equation Model (SEM) with mean structure. The structural model becomes

$$\begin{aligned} \beta_{0p} &= \alpha_0 + \boldsymbol{\gamma}_0' \boldsymbol{z}_p + \zeta_{0p} \\ \beta_{1p} &= \alpha_1 + \boldsymbol{\gamma}_1' \boldsymbol{z}_p + \zeta_{1p} \end{aligned} \tag{4}$$

Here $\boldsymbol{z}_p$ is a $q \times 1$ vector of time-invariant covariate (TIC) predictors, which are person-level background variables (e.g., sex, socioeconomic status) that influence the intercepts and slopes of the individual trajectories. $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ are $q \times 1$ vectors of coefficients from TICs on the average intercept and slope. Whether predictors are included or not, given the highly restricted factor loading matrix used to specify the growth function, model identification primarily depends on the number of repeated measures and the number and nature of the growth factors specified, with three repeated measures being sufficient to identify a linear growth model with intercept and slope factors (see Bollen & Curran, 2006, Ch. 2 for greater details).

Although first-order growth models are commonly fit to scale scores computed by taking the mean or sum of the item responses from each person at each time point, this practice implicitly makes several strong measurement assumptions that are often untenable in practice (Bauer & Curran, 2016; Kuhfeld & Soland, 2020). In particular, both scoring approaches start with a simple sum of items, or unit-weighted linear composite, with the mean score differing only in subsequently dividing by the number of items. This equal weighting of all items implies that they are all equivalent indicators of the underlying latent variable, failing to account for possible differences in either item severity or strength of relation to the latent construct. For instance, a scale score for depression might include the items “no longer enjoy activities I used to”, “experience sleep difficulties”, and “have thoughts about suicide”. Sum/mean scores treat these as equally severe symptoms that all equally reflect depression. Endorsing any one of these items, and no others, would lead to the same sum or mean score despite probable differences in severity between anhedonia and suicidal ideation, or the presumably weaker relationship between sleep disturbance and depression relative to either anhedonia or suicidal ideation (e.g., McNeish & Wolf, 2020).

Further, the weights are not only constant over items but also over timepoints and people, implying that the items provide constancy of measurement. Yet an item such as "cry easily" is a less severe indicator of depression for a young child than an adult, and also shows sex differences (Curran et al., 2014; Steinberg & Thissen, 2006) Violation of any of these assumptions could lead to a confounding of measurement differences with true construct change (e.g., Bollen & Curran, 2006, Ch. 8.2), resulting in biased growth estimates (Bauer & Curran, 2016; Grimm et al., 2013; Kuhfeld & Soland, 2020). Item-level missing data may further compound these problems.

One approach to address these issues is to specify a measurement model for each unique time point in the data and then use tests of factorial invariance to evaluate constancy of measurement over time. We first explicate this longitudinal measurement model and then describe how it provides the foundation for a second-order growth model that overcomes many of the limitations of the traditional first-order model.

Longitudinal Measurement Model and Invariance Testing

We start by specifying an occasion-specific factor model. We can represent the latent factor with the notation η_{tp}, which indicates the construct or trait level of person p at occasion t. η_{tp} is related to binary response y_{itp} from an item i at occasion t through a logistic response function[2]

$$P(y_{itp} = 1 \mid \eta_{tp}) = \frac{1}{1 + exp[-(\nu_{it} + \lambda_{it}\eta_{tp})]} \tag{5}$$

where ν_{it} and λ_{it} are the intercept and factor loading (or slope) parameters for item i at time t. The item intercept captures the difficulty of endorsing the item (e.g., severity) whereas the factor loading/slope captures the strength of relationship of the item to the factor. These parameters may differ between items to capture, for instance, differences in the severity or relevance of symptoms of depression like anhedonia, suicide ideation and

sleep disturbance.

At a single point in time, this item-level factor model for binary item responses is equivalent to a traditional 2-parameter logistic item response theory (IRT) model (Takane & de Leeuw, 1987; Wirth & Edwards, 2007). Here, however, this measurement model is specified at each observed time point, producing a T-dimensional item factor analysis model, where the mean and variance of the latent trait is typically allowed to vary over time and the intercepts and loadings of the items can potentially also vary over time (Andersen, 1985; Kim & Willson, 2014; von Davier et al., 2011; Wang et al., 2016). The latent factor values are assumed to be correlated over time and multivariate normal, so that $\boldsymbol{\eta}_p \sim N(\boldsymbol{\mu}_p, \boldsymbol{\Phi}_p)$, where $\boldsymbol{\mu}_p$ is a $T_p \times 1$ vector of latent trait means, and $\boldsymbol{\Phi}_p$ is a $T_p \times T_p$ factor covariance matrix with freely estimated off-diagonal elements.[3]Some constraints on these parameters are required to set the scale of the latent variable, with a common choice being to standardize the latent factor at the first time point (by setting its mean to zero and variance to one) and to assume that the intercepts and loadings are constant over time for at least a subset of items (minimally one). These items serve as the anchor items that link the scale of the latent factors at later timepoints to the scale set for the first time point.

The specification of a formal measurement model for the items over time allows for the possibility of evaluating measurement invariance, and determination of whether changes in the observed responses can be attributed to changes in the latent trait versus changes in measurement (Kim & Willson, 2014; Liu et al., 2017; Millsap, 2011). Factorial invariance over time is typically tested by evaluating whether the parameters of each item

---

[2] The model presented here assumes unidimensionality and configural invariance (Horn & Mcardle, 1992), that is, a single construct and factor structure is preserved over time.

[3] Note that a common factor mean vector and covariance matrix are assumed to hold across persons; the p subscript on the parameter matrices and T is relevant only if not all people are present at all occasions, serving solely to determine which elements come into play for the subset of time points observed for an individual.

can be constrained to be equal across occasions without significantly reducing model fit (e.g., Oort, 1998; Reise et al., 1993; Thissen et al., 1993; Woods, 2009). For example, this can be done through likelihood ratio tests comparing models with and without equality constraints on the parameters of each item (i.e., $\lambda_{i1} = \lambda_{i2} = \cdots = \lambda_{iT}$ and $\nu_{i1} = \nu_{i2} = \cdots = \nu_{iT}$ for $i = 1, \ldots, I$). Items for which the equality constraints are rejected are said to express DIF over time. Likewise, DIF due to between-subject differences on a grouping variable can be evaluated by considering across-groups equality constraints on the item parameters within a multiple-group modeling framework.

Such DIF testing methods hinge on the assumption that at least one anchor item is specified besides the target item(s) being evaluated and that the anchor(s) truly do not express DIF. Correctly identifying anchor items and accounting for DIF reduces the risk of inducing bias in the latent trait estimates (Curran et al., 2014) and enables comparison of latent trait levels across occasions and/or groups (F. F. Chen et al., 2005; Liu & West, 2018). Thus, only if one is confident in the selection of anchor items and identification of DIF should the estimated means, variances, and covariances of the latent trait be inspected for evidence of change over time or contrasted between groups. With sufficient confidence, it becomes possible to interpret and impose structure on the over-time means, variances, and covariances of the latent factor to evaluate hypotheses about trajectories of change over time. This brings us to the second-order growth model.

Second-Order Growth Models

Second-order growth curve models (SGCs) extend the longitudinal measurement model explicated above by adding a growth structure to the first-order factors η_{tp}, such that

$$\eta_{tp} = \beta_{0p} + \beta_{1p} time_{tp} + e_{tp},\ e_{tp} \sim N(0, \sigma^2) \tag{6}$$

This is similar to the first-order model in Equation (1), except that growth is modeled for the latent factor defined by the measurement model in Equation (5) rather than scores obtained as a sum or mean of item.[4]The time-specific residual e_{tp} is again assumed for

simplicity to be normally distributed, independent, and with constant variance, though it can optionally be specified as heteroscedastic to reflect changing residual variability in the latent factors across timepoints. The same structural model, given in Equations (2) and (3), describes inter-individual differences in the trajectory parameters of the SGC, and may similarly be augmented with predictors as in Equation (4). Figure 2 gives a path diagram of an unconditional SGC.

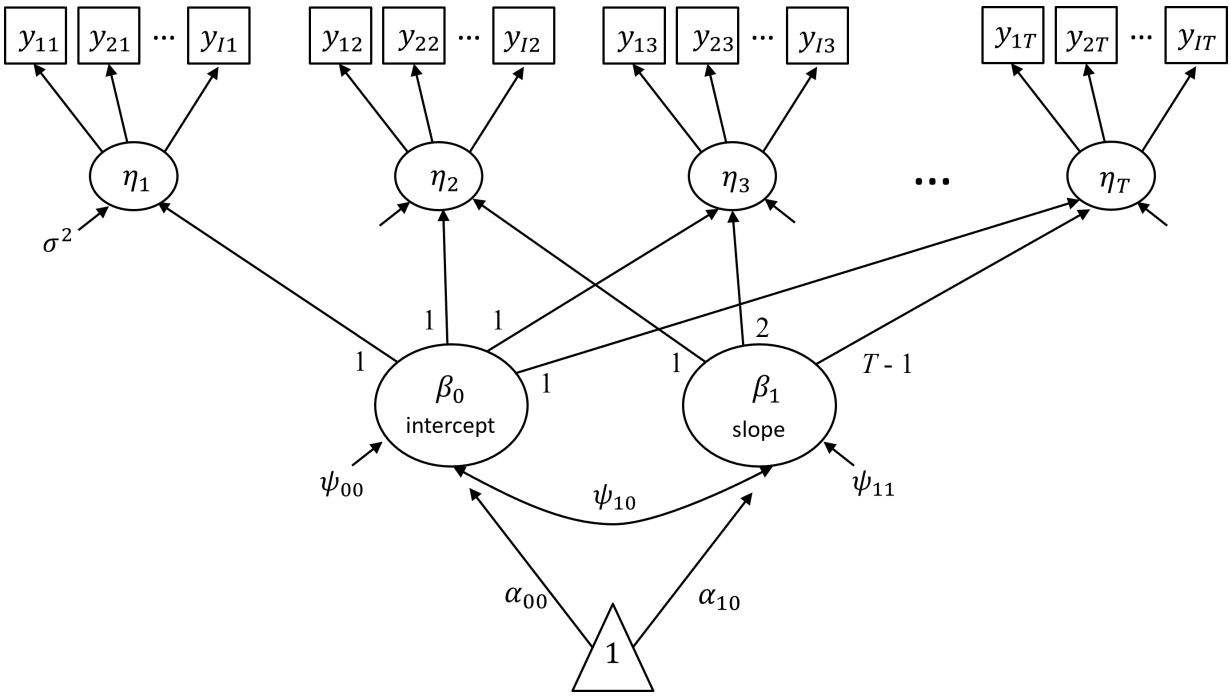


Figure 2

Path Diagram of A Second-Order Growth Curve Model. Person subscript p is omitted. To reduce clutter, item intercepts are implicit but not depicted.

In fact, a first-order growth models with scale scores can be conceptualized as an SGC with a highly constrained and somewhat ad hoc measurement model, e.g. $\eta_{tp} = \sum_{i=1}^{I} y_{itp}/I$ for mean scores. Thus, relative to first-order growth models, the principal feature of the SGC is that it models growth on latent traits defined by a formal measurement model instead of observed scale scores. Relative to the longitudinal measurement model reviewed above, however, SGC does not directly estimate latent trait means, variances, or covariances over time. These are instead implied by the growth

structure specified within the structural model (e.g., a linear growth process implies the factor mean changes linearly with time).

Model identification for the SGC follows from both the first-order growth model and the longitudinal measurement model. Like the longitudinal measurement model, we must set the scale of the latent trait and ensure this is maintained over time. Often, the latent trait at the first time point is scaled by setting α_0 to zero and ψ_{00} to one (i.e., standardizing the intercept factor of the growth process). An alternative is to set α_0 to zero and the sum of ψ_{00} and the latent trait residual variance σ^2 to one to mimic the common convention in longitudinal measurement models of standardizing the scale of the latent trait at the first timepoint. Still another option is to set the residual variance σ^2 to one. In any of these cases, mean and variance shifts at later time points are only interpretable if this scale is inherited by the latent trait at subsequent occasions of measurement, and this requires the presence of anchor items. Thus, the parameters ν_{it} and λ_{it} of at least one item (but preferably more) must be constrained to be equal over T waves. Last, to identify the structural model in Equation (2), a sufficient number of repeated measures must be obtained (e.g., three repeated measures to identify linear growth), following the same rules as first-order growth models.

Relative to the first-order growth model, the SGC has several key advantages (Kuhfeld & Soland, 2020; McArdle, 1988; Sayer & Cumsille, 2001). First, like the longitudinal measurement model, the items are allowed to differ in difficulty/severity and relatedness to the latent factors. Second, although here we have expressed the model for exclusively binary items, it extends naturally to accommodate items with different or mixed scale types. Third, mean/sum scores are not perfectly reliable, whereas we can view

[4] From a multilevel modeling perspective, the addition of an item-level measurement model for binary data also makes the SGC equivalent to a multilevel item factor model (Pastor & Beretvas, 2006; Raudenbush et al., 1991; Wang et al., 2016), consisting of a level-1 within-occasion measurement model (Equation 5), a level-2 trajectory model (Equation 6), and a level-3 person growth factor model (Equation 2).

the latent trait within the second-order model as error free. Most important, for the present discussion, is that the SGC can be used to evaluate growth while accounting for DIF on some items over time and groups. Like the longitudinal measurement model reviewed above, the item parameters can be tested for equality over time and groups (the latter by using a multiple-group framework; Kim & Willson, 2014; von Davier et al., 2011). Differences on a subset of items are permissible, provided enough anchor items remain to equate the scale of the latent trait over time and groups with confidence. We can thus use the SGC to model construct change while simultaneously testing and accounting for measurement changes, addressing Bereiter's fundamental question "how can one tell whether it is the persons who have changed or the tests?" However, as we review next, the SGC is not a fully satisfactory tool for answering this question, given limitations that curtail its applicability and usefulness in practice.

Limitations of SGC

A key feature of the SGC is that one must specify a latent factor for each unique time value observed within the data (e.g., subject age). This introduces a discrepancy in how time is treated between the measurement model in the SGC versus the growth model. In the measurement component of the SGC, time is treated as discrete or categorical with each snapshot in time represented by a time-specific measurement model. In the growth component, however, time is treated as continuous, with individual trajectories following a (usually) smooth function across measurement occasions (e.g., a straight line). In addition to being conceptually contradictory, the categorical treatment of time in the measurement model has practical implications for model estimation and invariance testing.

With respect to model estimation, the number of latent dimensions in the SGC grows with the number of unique timepoints T. As T increases, the computational cost of conventional full-information maximum likelihood (FIML) with numerical integration increases exponentially. In brief, the model likelihood is typically computed by numerically approximating the area under a multivariate latent distribution with T dimensions, which

represents probabilities of response in the data. Each dimension is approximated with "quadrature points" that are shapes with easily computable areas. To achieve a reasonable approximation, a minimum recommended number of quadrature points q is 5, but typically people prefer more (e.g., 10–15). With T dimensions and q quadrature points per dimension, FIML requires q^T quadrature points in computing the likelihood, i.e., high-dimensional quadrature, a number that increases rapidly with the number of unique repeated measures, quickly becoming computationally intractable. Alternative methods, such as diagonally weighted least squares (DWLS), are not affected by this "curse of dimensionality", but they have other drawbacks. Sparseness in the one-way or two-way frequency tables of the items can lead to estimation problems, and treatment of missing data is less optimal than under FIML (Liu et al., 2017; Wirth & Edwards, 2007). These practical implications hinder the applicability of the SGC in longitudinal studies with many timepoints or where there is considerable variability in the numbers of responses provided to items (e.g., due to missing data or when not all items are used at all timepoints).

Invariance testing also becomes progressively more inefficient and error-prone as the number of unique time values increases. Conventional tests of DIF over time, whether by likelihood ratio, Wald, or Score test, are based on testing the null hypothesis that the item parameters for a given item are equal across all timepoints versus the alternative hypothesis that they are somehow unequal. This amorphous alternative hypothesis, however, yields inefficient tests of DIF because it fails to account for the continuous nature of time. If there is DIF, it is unlikely that the item parameters change erratically over time. It is more plausible that the item parameters wax or wane systematically with time. For instance, crying becomes less normative and a progressively more severe indicator of depression as children age into adults (Curran et al., 2014). The traditional approach of treating time categorically fails to capitalize on such systematic changes and thus lacks parsimony. This lack of parsimony both decreases statistical efficiency and yields fewer practical insights about the nature of DIF. The problem is compounded if one crosses

groups with time, for instance to test whether crying is a more severe indicator of depression for boys versus girls particularly at later ages.

These conventional tests of DIF over time and/or group are also error-prone for two primary reasons. First, they rely on significance testing that iterates over items — first test DIF for Item 1, then Item 2, etc. This sequence of repeated testing is performed over time, over each grouping variable, and possibly over both group and time, etc. The high number of tests leads to accumulating risks of errors due to capitalization on chance (Draper, 1995; MacCallum et al., 1992). Second, to identify the model, at least one item must be declared as an anchor when testing another for DIF. Given the potential for DIF over time, groups, or both, selection of anchors can be difficult, and poor choices increase the likelihood of Type I errors in DIF detection (Ankenmann et al., 1999; Bauer et al., 2019; Finch, 2005; Stark et al., 2006). Together with the aforementioned issues, conventional DIF testing in SGC lacks parsimony, efficiency and interpretability and is likely to generate many errors.

A last limitation is that SGC can only consider DIF from a limited range of covariate types. The multiple-group SGC setup does not easily accommodate DIF testing from continuous covariates whose values are specific to each subject (e.g., income). Cutting up continuous covariates to produce artificial groups is poor solution (MacCallum et al., 2002). Similarly, SGC has difficulties allowing for time variables that have a high heterogeneity within waves or unevenness in the lags between assessments across persons (e.g., some are 1 year older while others are 1 year and 3 months older from the first to the second wave). Many unique values in the time metric will lead to too many time-specific factors to be practically estimable. Rounding the time variable to fewer unique values may be fine if change is gradual relative to the time scale of observation, but otherwise may occlude important aspects of the growth process such as effects of birth cohorts between subjects (Hoffman et al., 2011).

We next present a longitudinal MNLFA approach as an alternative that addresses these limitations of the classical SGC.

Longitudinal MNLFA

To address the limitations noted above, we reconceptualize longitudinal measurement within a moderated nonlinear factor analysis framework (MNLFA) framework. MNLFA was originally developed to overcome limitations of traditional models for evaluating between-person DIF in cross-sectional data (Bauer, 2017; Bauer & Hussong, 2009). To our knowledge, this is the first extension of this model to a fully longitudinal context, where it offers important advantages for the simultaneous modeling of DIF and construct change over time. As explicated below, within the longitudinal MNLFA, time is treated continuously in both the measurement model and growth model. The model allows for a more parsimonious specification of over-time DIF, by permitting the item parameter values to be moderated by the continuous value of time. The model also easily accommodates between-person predictors (categorical or continuous) as potential moderators.

Measurement Model

The measurement model for the MNLFA is similar to that presented previously in Equation (5) but with item parameters that now potentially vary both over time and persons, resulting in the addition of the subscript p to the intercept and loading:

$$P(y_{itp} = 1 \mid \eta_{tp}) = \frac{1}{1 + exp[-(\nu_{itp} + \lambda_{itp}\eta_{tp})]} \tag{7}$$

The values of the item parameters are then moderated by $\boldsymbol{x}_{tp}$, an $m \times 1$ vector of predictors, which includes time-invariant predictors $\boldsymbol{z}_p$ as a subset with constant values over t as well as the time value $time_{tp}$ for an individual p at measurement occasion t, and any other effects related to time (e.g., a quadratic effect $time_{tp}^2$ on item parameter values), as follows:

$$\begin{aligned} \nu_{itp} &= \nu_{i0} + \boldsymbol{\kappa}_i' \boldsymbol{x}_{tp} \\ \lambda_{itp} &= \lambda_{i0} + \boldsymbol{\omega}_i' \boldsymbol{x}_{tp}. \end{aligned} \tag{8}$$

Here ν_{i0} and λ_{i0} are baseline intercept and slope values for item i when $\boldsymbol{x}_{tp} = \boldsymbol{0}$. The $m \times 1$ vectors $\boldsymbol{\kappa}_i$ and $\boldsymbol{\omega}_i$ capture any DIF over time or associated with time-invariant predictors.

These predictors can be a mix of categorical and continuous variables. If $\boldsymbol{\kappa}_i$ and $\boldsymbol{\omega}_i$ are all zero, then there is no DIF for the item and it serves as an anchor over both different timepoints and people.

Growth Models

Beyond using continuous time in the measurement model, the remaining structure of the longitudinal MNLFA resembles the SGC. The functional form of growth is specified identically to Equation (6). This portion of the model uses the subset of variables within $\boldsymbol{x}_{tp}$ that reflect time scores and appear within the function describing the form of the growth trajectory, such as linearly increasing time scores for linear growth (with the addition of quadratically increasing time sores for quadratic growth). Likewise, Equations (2) and (3) continue to hold for a longitudinal MNLFA without predictors of growth, from which one can obtain estimates of the average trajectory and extent of individual differences, and Equation (4) demonstrates how the expected values for the individual intercepts and slopes can be conditioned on time-invariant predictors.

Unlike existing growth models, however, longitudinal MNLFA also allows the variances and covariances of the random effects to be conditioned on the time-invariant predictors. This is achieved by letting elements of the variance-covariance matrix vary over individuals as a deterministic functions of the time-invariant predictors, similar to cross-sectional MNLFA (Bauer, 2017), that is,

$$\begin{bmatrix} \zeta_{0p} \\ \zeta_{1p} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \psi_{00p} & \\ \psi_{10p} & \psi_{11p} \end{bmatrix} \right). \tag{9}$$

To avoid negative implied values, the variances are often specified to be a log-linear function of the predictors

$$\begin{aligned} \psi_{00p} &= \psi_{000}\ \exp(\boldsymbol{\gamma}_{r0}'\boldsymbol{z}_p) \\ \psi_{11p} &= \psi_{110}\ \exp(\boldsymbol{\gamma}_{r1}'\boldsymbol{z}_p) \end{aligned} \tag{10}$$

where ψ_{000} and ψ_{110} are the baseline intercept and slope (growth factor) variances when $\boldsymbol{z}_p = \mathbf{0}$. $\boldsymbol{\gamma}_{r0}$ and $\boldsymbol{\gamma}_{r1}$ are the $q \times 1$ effects of the predictors on the intercept and slope

variances. Similarly, to avoid implied intercept-slope factor correlations exceeding ±1, Bauer (2017) suggested use of a function based on the Fisher z-transformation, or

$$\begin{aligned} \psi_{10p} &= \psi_{00p}^{1/2}\rho_{10p}\psi_{11p}^{1/2} \\ \rho_{10p} &= 1-\frac{2}{\exp[2(\rho_{100}+\boldsymbol{\rho}_{101}'\boldsymbol{z}_p)+1]} \end{aligned} \tag{11}$$

where ρ_{10p} is the individual-specific growth factor correlation modeled by a Fisher z-transformed linear moderation function, $\rho_{100}+\boldsymbol{\rho}_{101}'\boldsymbol{z}_p$. With two growth factors, such as in a linear growth model, using log-linear functions for the variance and the Fisher z function for the correlation is sufficient to ensure that the covariance matrix of the growth factors remains positive definite (i.e., a proper covariance matrix) for all possible values of the predictors. Residual variances can also be conditioned on predictors, such that $\sigma_p^2=\sigma_0^2\exp(\boldsymbol{\gamma}_e'\boldsymbol{z}_p)$, where σ_0^2 is the baseline error variance, and $\boldsymbol{\gamma}$ are the effects of the predictors on the error variances.

In sum, the longitudinal MNLFA represents time consistently as a continuous variable in both the measurement model and growth model. Within the measurement model, it allows for DIF to be expressed as a continuous function of time as well as person-level predictors. It allows for predictor effects on the expected values of the growth factors much like other growth models. Additionally, it allows the factor variances and covariance to be conditioned on categorical and/or continuous predictors. In the next section, we compare the longitudinal MNLFA to the SGC in more detail, with a numerical illustration.

Relationship to Second-order Growth Modeling

Because the traditional SGC treats time categorically in the measurement model, it requires a time-specific factor model for every unique time value observed in the sample. This allows the item parameters to vary over time, capturing DIF as differences in item parameters between discrete timepoints. In contrast, in the longitudinal MNLFA, the item parameters are moderated as a function of continuous time and corresponding DIF

coefficients. This matches the continuous treatment of time with both models for the growth equation. To more clearly show the difference, we can mimic the SGC within a longitudinal MNLFA by treating time categorically. Let us consider a simple case in which all participants are observed three times, the time variable is age, and there is age homogeneity within timepoints (e.g., everyone is observed at ages 12, 13, and 14). We can re-express the item parameter moderation functions of the longitudinal MNLFA to be consistent with the SGC approach as follows

$$\begin{aligned} \nu_{it} &= \kappa_{i1} time_1 + \kappa_{i2} time_2 + \kappa_{i3} time_3 \\ \lambda_{it} &= \omega_{i1} time_1 + \omega_{i2} time_2 + \omega_{i3} time_3 \end{aligned} \tag{12}$$

where $time_1$–$time_3$ are indicator variables for each time point, i.e., $time_1$ is scored one for an observation at the $1st$ time point, zero otherwise; $time_2$ is scored one for an observation at the $2nd$ time point, zero otherwise, etc. The parameters $\kappa_{i1} \ldots \kappa_{i3}$ and $\omega_{i1} \ldots \omega_{i3}$ now represent time-specific item intercepts and loadings. The item intercepts and loadings can be set equal over time to construct an anchor item or allowed to differ from one if the item expresses DIF. In contrast, consider the proposed parameterization (8) under the assumption that any changes in the item parameters (i.e., DIF) occur as a linear function of age:

$$\begin{aligned} \nu_{itp} &= \nu_{i0} + \kappa_{0i} age_{tp} \\ \lambda_{itp} &= \lambda_{i0} + \omega_{0i} age_{tp}. \end{aligned} \tag{13}$$

Here the time variable age_{tp} is treated as a continuous variable, and DIF over time is represented via a single coefficient for each item parameter, regardless of the number of timepoints in the model.

Shifting the measurement model in this way overcomes many of the limitations of the SGC that we reviewed above. First, because the longitudinal MNLFA treats the time variable as continuous, it will typically provide a more parsimonious, efficient, and interpretable representation of DIF than the SGC. For instance, in the example above, DIF modeled via different item parameters at every time point requires six parameters (three

time points for both intercept and loading) whereas modeling DIF as a linear function of time requires just four (a baseline value and linear DIF effect for both intercept and loading). As the number of time points grows, this advantage becomes more substantial. Any obtained DIF in the longitudinal MNLFA is also more easily interpreted in terms of continuous change over time in the item properties. With SGC, the additional parameters needed to model DIF across time points offer the contrasting advantage that no functional form need be assumed for DIF over time, but DIF evaluation and interpretation become more difficult. For instance, one might test DIF globally across all time points or pairwise between time points (e.g., testing the equality of κ_{i2} and κ_{i3} in Equation 12) to evaluate change in item properties over time.

Second, most item-level data is discrete in nature (binary or ordinal), in which case longitudinal MNLFA will often be more computational efficient than SGC, particularly when the number of unique observed time values exceeds the number of waves of assessment (e.g., an accelerated longitudinal design). For instance, to model age-related changes in a 4-wave dataset with an age span from 12 to 17, the proposed model requires four first-order dimensions for the four assessments, with values of age varying between people within waves. In contrast, the SGC requires specification of a first-order factor for each unique value of time, requiring six dimensions for the six observed ages. When using maximum likelihood estimation with numerical integration methods (e.g., quadrature), the computational burden grows exponentially with the number of dimensions, making the lower-dimensional MNLFA specification advantageous. When the number of unique time points is large, fitting the MNLFA using Bayesian methods of estimation becomes attractive. This is because the computational burden of the Bayesian model increases only linearly with the number of dimensions (lower-order factors) in the model, helping to overcome the "curse of dimensionality" that impedes estimation with maximum likelihood. Additionally, as we explicate later, Bayesian estimation methods allow for the incorporation of priors that can be helpful for localizing DIF.

In addition to these differences with respect to the measurement model, there are also important differences with respect to the structural model. Like the traditional SGC, the second-order growth factors give structure to the first-order latent trait means, variances, and covariances. However, whereas in the SGC these implied moments vary over a limited number of fixed values of time, the MNLFA allows the implied moments to vary over any values of age, allowing for greater between-person differences in the timing of measurements. Indeed, the time scores could vary between people so much that no two individuals are measured at the same times. Finally, another important difference concerns the types of time-invariant predictors that can be accommodated. Whereas the longitudinal MNLFA allows both categorical and continuous predictors to moderate item parameters and the variances and covariances of the growth factors, the SGC is limited to considering the effects of categorical predictors (often just one) on these parameters in a multiple-groups framework.

Numerical Example

We now present a numerical example to demonstrate how the proposed longitudinal MNLFA model differs from the traditional SGC. The example uses a simulated dataset from Bauer and Curran (2016). Data is generated from a 4-wave SGC model, with 1000 observations, 8 items per wave, with age homogeneity within waves. Age is coded as -1.5, -.5, .5, and 1.5, placing the intercept of the growth function in the middle of the observed age range. There is one time-invariant predictor, a binary variable called "site of study" to mimic a situation in which data is obtained from two study sites, such as two countries. On average, individuals in the second site had higher intercepts but less steep slopes than individuals from the first site. There were no differences in the variance/covariance parameters across sites. We also included DIF over site (on Items 2, 3, and 5) and as a linear function of age (on Items 2, 3, and 4). Items 1 and 6-8 are pure anchors across both site and age; Item 4 is an anchor with respect to site but not age; and Item 5 is an anchor with respect to age but not site. Note that several restrictions on the population-generating

model make it possible to fit comparable multiple-group SGC and longitudinal MNLFA models—relatively few waves, age homogeneity within waves, and a single categorical predictor—restrictions that are needed for the SGC but not the longitudinal MNLFA. The analysis models included a traditional SGC model and a longitudinal MNLFA model. Both models were fitted to the 4 waves of data with correctly specified patterns of DIF from age and site; differences in mean and variance parameters of growth factors between sites were also allowed (though only the means actually differed over sites in the population). Model identification was achieved by constraining the sum of the residual variance and intercept factor variance in site 1 to the population true value 0.76.[5]The SGC model with discrete age DIF was estimated with maximum likelihood in *Mplus* (Version 8.3; Muthén & Muthén, 1998–2017) using a multiple-group estimation approach to allow for site differences in the item parameters and growth parameters. The longitudinal MNLFA with continuous age DIF and site differences was estimated in the *rstan* (Version 2.21.8; R Core Team, 2020; Stan Development Team, 2020) package with Bayesian estimation and generic diffuse prior distributions (greater explication of Bayesian estimation is provided in the next section).[6]Data and code used in this example are available in Supplemental Materials, which contains instructions on how to compile the data and specify the models.

Figure 3 compares the intercept and loading estimates obtained from the two models for the items across 4 waves and 2 sites. Most values are close to the 45-degree reference line. The item estimation results are close between the two models besides small

[5] This non-standard value was selected because the data-generating model standardized the over-time variance of the latent factor to facilitate selection of item parameter values, rather than standardizing the variance of the latent factor at an individual time point.

[6] In this numerical example, the Bayesian longitudinal MNLFA model took 122 minutes to converge with an Intel Core i5 CPU; the SGC model fit with maximum likelihood took 36 minutes (11 minutes if using 4 processors in Mplus.) The slower run time of the Bayesian model is expected given a relatively low-dimensional problem with only four first-order factors. With more unique time points, we would expect the run time to reverse in favor of Bayesian estimation of MNLFA.

differences due to the alternative methods of estimation, though the models arrived at these item results differently. The traditional SGC model directly estimated item parameters at each wave and site, whereas longitudinal MNLFA directly estimated DIF effects over age and site as two parameters from which the item values were implied. This contrast of model specification can be seen in Figure 4, which plotted estimates of items that contained DIF effects over age or site. As shown here, item estimates from the longitudinal MNLFA were smooth linear functions of age, whereas item estimates from the SGC model fluctuated across age via a larger number of time-specific parameters. Both were close to the true population values, but the longitudinal MNLFA provided a more parsimonious and interpretable specification.

A comparison of the structural model parameter estimates for the underlying growth process is shown in Table 1. Growth mean and variance estimates between the models appeared similar in magnitudes and were consistent with the true population values. Figure 5 compared mean trajectories plotted with the structural parameter estimates over age and by site. It can be seen that the mean trajectories closely overlapped with the population true values.

In short, the numerical example demonstrated that the proposed longitudinal MNLFA model recovered item and growth parameter results similar to traditional SGC models. The proposed model, however, did so more efficiently by representing the change in item properties over time via a parsimonious linear function. In this example we simulated data to have homogeneous age within each timepoint and DIF from only a categorical variable, so that the multiple-group SGC model could be accurately implemented. Age homogeneity within wave and DIF effects only from time and/or categorical covariates are, however, not required by the longitudinal MNLFA. By parameterizing time and other covariate effects as moderation effects, the proposed model can examine change in item measurement properties from DIF covariates that are categorical, continuous, or interacting with time.

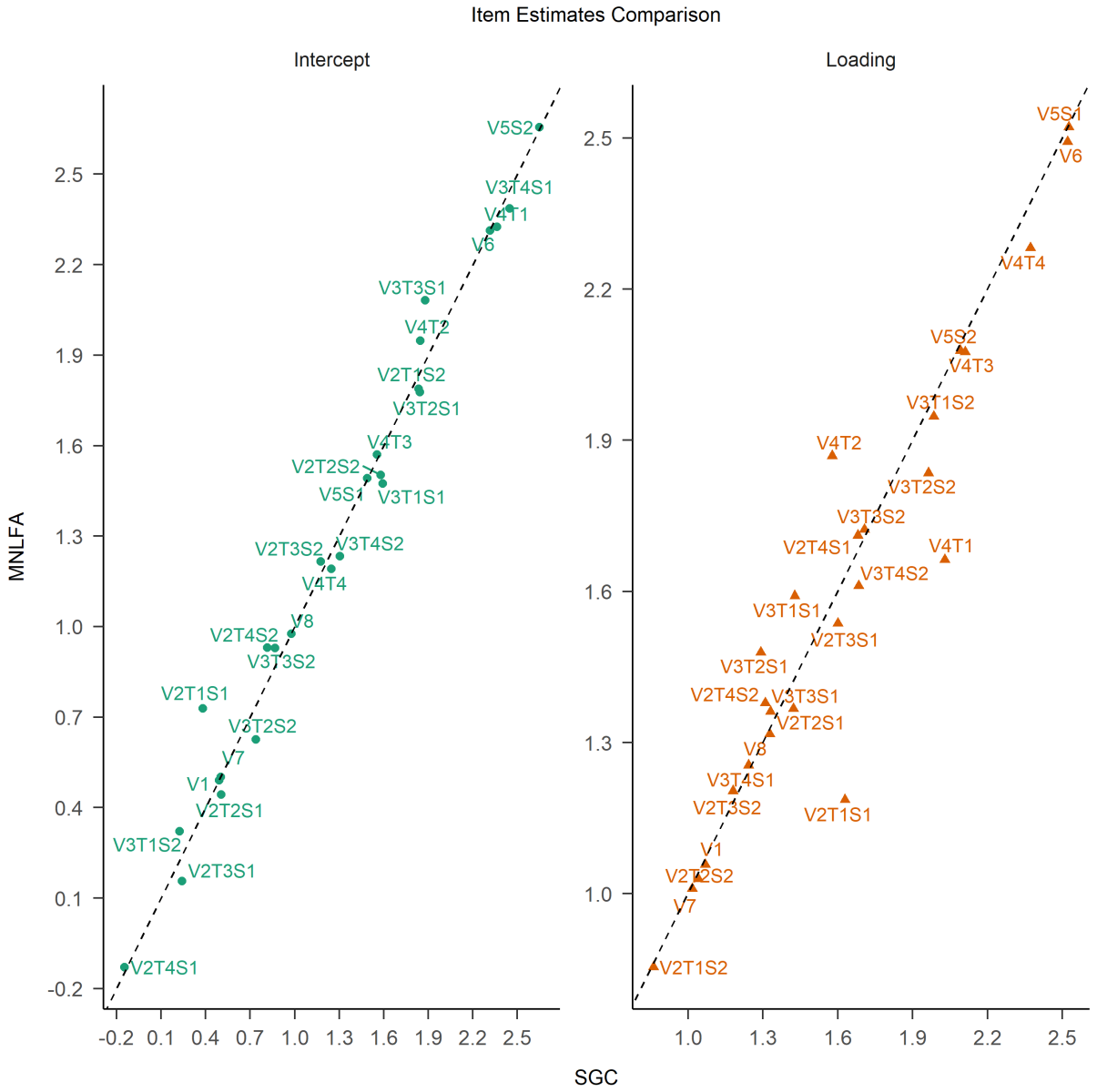
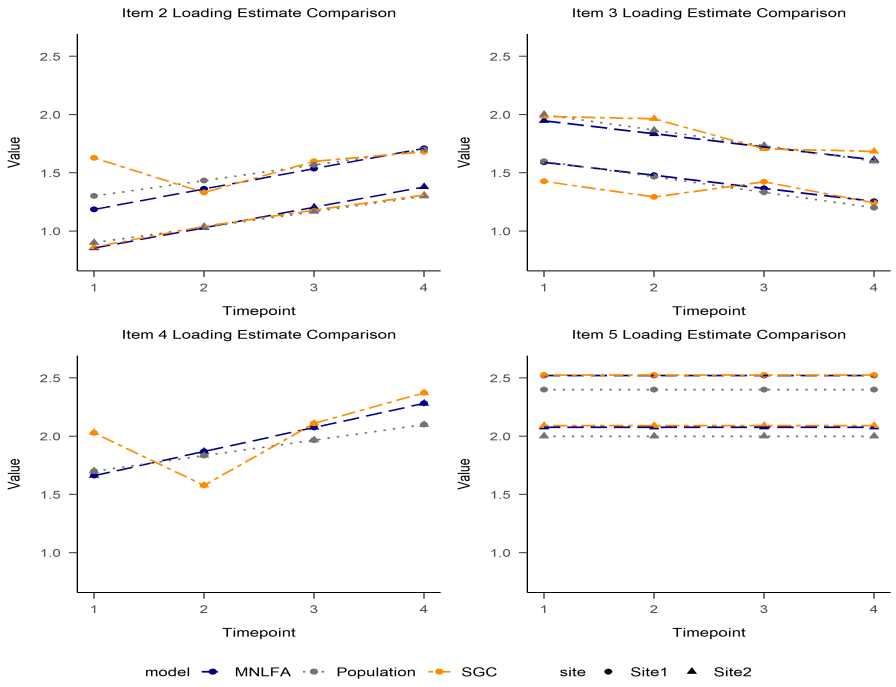
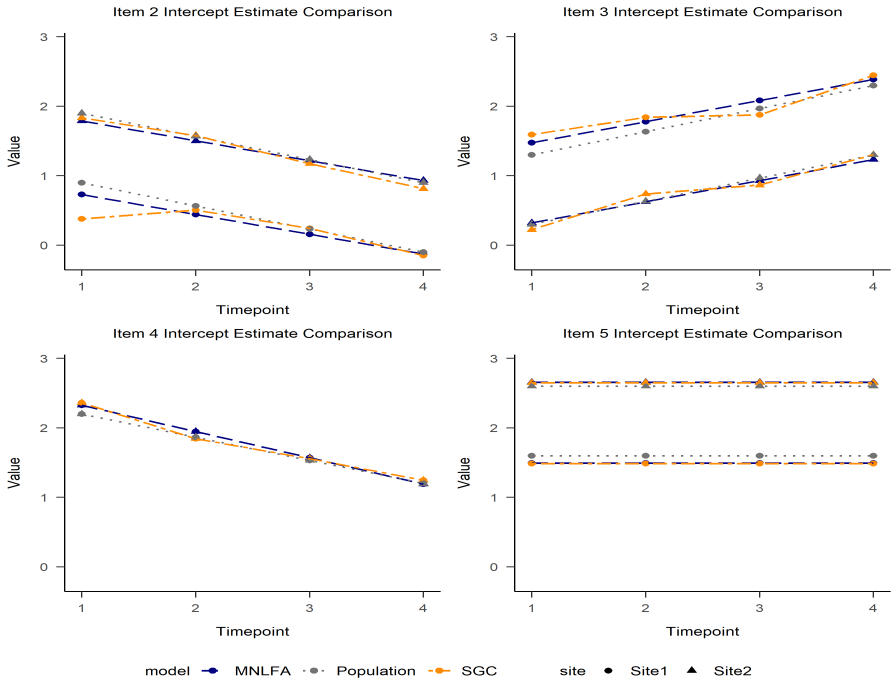


Figure 3

Aggregated Item Intercept and Loading Estimates from SGC and Longitudinal MNLFA. Intercept estimates were sign-reversed for presentation. Dashed-line is a 45-degree reference line. Letter V is the item index; T is the time index; S is the site index. T and S indices indicated the presence of DIF effects.



(a) *Item Loading Estimates*



(b) *Item Intercept Estimates*

Figure 4

Estimates and Population Values of Items with DIF Plotted Over Timepoint (Age -1.5 to 1.5) and Site. Intercept estimates were sign-reversed for presentation.

Table 1

Second-order Growth Parameter Estimate Comparison.

Growth Factor Mean	*SGC*	*MNLFA*	*Population*
Intercept Site 1	-0.060	-0.061	-0.06
Slope Site 1	0.489	0.492	0.51
Intercept Site 2	0.061	0.061	0.06
Slope Site 2	0.379	0.384	0.39
Growth Factor Variance	*SGC*	*MNLFA*	*Population*
Intercept Site 1	0.347	0.355	0.35
Slope Site 1	0.055	0.049	0.07
Covariance Site 1	0.091	0.094	0.1
Intercept Site 2	0.297	0.303	0.35
Slope Site 2	0.051	0.049	0.07
Covariance Site 2	0.097	0.097	0.1
Residual variance	0.413	0.433	0.41

When fitting the models in this numerical example we also assumed that the locations of DIF were known. In practice, under the presence of many timepoints and different types of DIF covariates, it becomes increasingly difficult to correctly identify anchor items a priori or evaluate DIF using traditional methods. To achieve proper measurement model specification and accurate estimates, we require alternative methods of DIF detection. Here, we suggest applying Bayesian regularization to the longitudinal MNLFA to achieve identification of the model without predetermined anchor items, allowing for the simultaneous assessment of DIF throughout the model. It is this topic to which we now turn.

DIF selection under Bayesian Regularized Longitudinal MNLFA

To fit the longitudinal MNLFA we use Bayesian methods of estimation, which hold two important advantages in this context. First, unlike maximum likelihood with numerical integration, the estimation time of Bayesian methods does not scale up exponentially with the number of latent dimensions. Specifically, comparing to the SGC with T unique timepoints estimated in maximum likelihood, the longitudinal MNLFA with

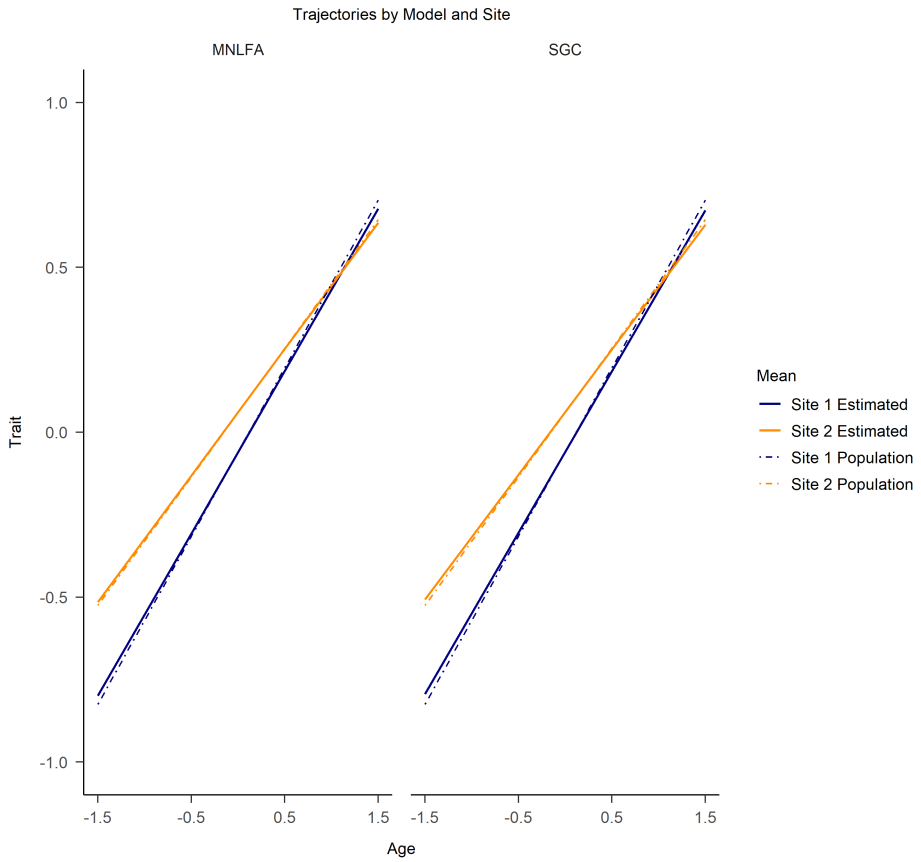


Figure 5

Estimated Mean Trajectories from SGC and Longitudinal MNLFA Compared to Population True Trajectories.

Bayesian estimation does not require evaluating q^T quadrature points to approximate the model likelihood. Instead, the Bayesian model incorporates $T_p + 2$ latent variable scores for each person, comprised of T_p scores for the time-specific lower-order factors plus two scores for the higher-order growth factors (Fox & Glas, 2001). The computational burden of Bayesian estimation can be higher than maximum likelihood when there are few latent dimensions, but the advantages of the former become substantial as T increases.[7]Bayesian methods also reduce chances of inadmissible solutions in small samples or data with sparse responses (Depaoli & Clifton, 2015; Muthén & Asparouhov, 2012). Second, a specific advantage for the current model is that we can incorporate Bayesian regularization procedures to improve the detection of DIF over time and/or persons. Here we briefly

review Bayesian estimation and regularization as these pertain to the proposed longitudinal MNLFA. More general reviews of Bayesian estimation are provided by Gelman et al. (2013) and van de Schoot et al. (2014).

Relative to frequentist methods, Bayesian estimation differs by treating population parameters as random variables rather than fixed values. That is, parameters are described via distributions rather than having single numerical values. In fitting the model, one assigns prior distributions to model parameters to describe their likely range of values for each model parameter. The parameters of these prior distributions (e.g., the mean and SD of a normal prior placed on the intercept of an item) are referred to as hyperparameters. The prior is multiplied by the likelihood of the data to obtain the posterior distribution of the model parameter. In effect, the posterior represents an update of the prior informed by the characteristics of the observed data. In practice, Bayesian estimation is conducted by approximating the joint distributions of the priors and likelihood. This is typically done with Markov Chain Monte Carlo (MCMC) sampling methods, which collects simulated random samples that are representative of the target joint distribution (e.g., Kruschke, 2010, Ch. 7). The resulting posterior distribution of each parameter of interest is summarized, for instance by its mean, standard deviation, and credible intervals, to correspond to the point estimates, standard errors, and confidence intervals usually reported with frequentist methods.

Often, Bayesian estimation is conducted using "diffuse" priors that have large scale parameters (e.g., a normal distribution with a very large standard deviation) and provide little information about the potential values of the parameters. This allows the likelihood to dominate the posterior and for Bayesian estimation to mimic the results obtained from

[7] The stochastic version of maximum likelihood accrues this same benefit by incorporating MCMC (e.g., Wirth & Edwards, 2007). Also note that this computational advantage of Bayesian estimation applies to categorical data. Though responses to scale items are often categorical, when the outcome data is continuous, maximum likelihood does not require numerical integration and is usually less computationally intensive than Bayesian estimation.

conventional maximum likelihood estimation (Browne & Draper, 2006). However, here we consider priors that can be used to inform parameter selection, sometimes referred to as Bayesian penalty methods. The idea of parameter selection is to identify which non-zero coefficients to include in the model from a set of candidate coefficients. Examples include selecting a relevant subset of predictors from a set of candidate predictors for a regression model (Bainter et al., 2020; Barbieri & Berger, 2004), or identifying a set of cross-loadings in a confirmatory factor model (Lu et al., 2016). Parameter selection using Bayesian penalty methods is achieved by giving regularizing prior distributions to all the candidate parameters. A regularizing prior has a mean of zero and a highly constrained scale parameter which influences the posterior distribution more than a diffuse prior, pulling the regularized parameter estimates towards zero. Thus unimportant effects are shrunken towards zero by this prior, while important effects with larger magnitudes still remain in the model after penalization. An example is a normal prior with a mean of zero and small variance. Relative to their frequentist counterparts (lasso, ridge, elastic net, etc.), Bayesian penalty methods are attractive because they (1) avoid intensive cross-validation to select penalty values (these are often given hyperpriors and estimated together with the model); (2) incorporate uncertainty regarding penalty parameter estimation to improve parameter selection results; (3) provide empirical standard errors for all penalized coefficients; and (4) scale efficiently to complex models with multiple latent dimensions (Kyung et al., 2010; Leng et al., 2014; Narisetty & He, 2014; Park & Casella, 2008).

Recent studies in Bayesian latent variable modeling have considered the use of penalty methods for DIF identification (Brandt et al., 2023; S. M. Chen et al., 2022) and other problems of model specification search (Feng et al., 2017; Jacobucci & Grimm, 2018; Lu et al., 2016; Muthén & Asparouhov, 2012; Pan et al., 2017; Shi et al., 2017) in single-timepoint or cross-sectional settings. A particularly useful feature of regularization is that one need not declare anchor items in advance of the analysis. DIF effects that do not manifest in the data are shrunken close to zero, which allows the estimated measurement

model to achieve "approximate model identification" without explicitly specifying anchor items. Bayesian penalty methods can thus provide lower DIF detection error rates than conventional DIF detection procedures that require potentially erroneously selected anchors (see Bauer et al., 2019; Belzak & Bauer, 2020; S. M. Chen et al., 2022, for comparisons of DIF detection results based on frequentist lasso and Bayesian penalty methods relative to traditional likelihood ratio testing).

Similarly, in the case of longitudinal MNLFA, the goal is to use Bayesian regularization to identify where there is DIF and where there is not. Here, we apply a class of Bayesian penalty priors called "spike-and-slab" priors (SSP; George & McCulloch, 1993; Ishwaran & Rao, 2005; Kuo & Mallick, 1998) to the DIF parameters within the longitudinal MNLFA. SSP expresses the probability that a parameter should be included in the model using its built-in inclusion parameters (Hans, 2010; Lykou & Ntzoufras, 2013). Thus, researchers can compare estimates of inclusion parameters to a threshold, τ, where a parameter would need an inclusion probability equal to or greater than τ to be selected for the model (Lu et al., 2016; Raftery et al., 1997). Previous simulation studies in cross-sectional MLNFA applications with binary data (S. M. Chen et al., 2022) found a threshold of 0.7 and 0.8 achieved effective DIF detection power and Type I error rate that were superior to other frequentist and Bayesian penalty methods.

In our Bayesian regularized longitudinal MNLFA, all DIF effects in the measurement model are freely estimated and evaluated under SSP. Our specification of SSP combines a penalizing Bayesian lasso prior with an inclusion parameter with a U-shaped Beta prior (Brandt et al., 2018; Lykou & Ntzoufras, 2013). For each item DIF parameter $\boldsymbol{\xi}_i^{dif}$ ($\boldsymbol{\xi}_i^{dif} = \boldsymbol{\omega}_i$ or $\boldsymbol{\kappa}_i$), the Bayesian lasso prior has a Laplace (double exponential) distribution

$$p\left(\boldsymbol{\xi}_i^{dif}|u,\phi\right) = \prod_{i=1}^{p} \frac{\phi}{2\sqrt{u^2}} \exp\left\{\frac{\phi|\boldsymbol{\xi}_i^{dif}|}{\sqrt{u^2}}\right\}, \tag{14}$$

or $\boldsymbol{\xi}_i^{dif} \sim dexp(0,(u/\phi))$. u is a known standard deviation value. ϕ is the penalty parameter used to adjust the final scale of this prior distribution. An inclusion parameter r_i is placed onto this Bayesian lasso prior. r_i is assumed to follow a U-shaped prior with

mean 0.5 and range $(0, 1)$. The full SSP is specified as a product of the inclusion parameter and the DIF parameter that is penalized by Bayesian lasso

$$\begin{aligned} \boldsymbol{\xi}_i^{dif} &= \boldsymbol{\xi}_i^{*dif} r_i \\ \boldsymbol{\xi}_i^{*dif} &\sim dexp(0, (u/\phi)) \\ r_i &\sim Beta(0.5, 0.5) \end{aligned} \tag{15}$$

where r_i conducts parameter selection, and the Laplace distribution on intermediate DIF parameters $\boldsymbol{\xi}_p^{*dif}$ applies shrinkage. Here one r_i is given to all DIF parameters for an item, though separate inclusion parameters could be given to DIF effects from different covariates. Figure 6 plots the densities of the two distributions that consist of SSP in the first row; in the second row, the figure illustrates the resulting SSP distribution that is conditional on of a range of inclusion parameter values. At a given level of lasso penalty, when the inclusion parameters are estimated with their own hyperpriors rather than given fixed values (such as a Beta prior in Equation 15), the final SSP distribution will be a mixture of densities at different values of inclusion parameters (a mixture of the densities at different heights in the bottom row of Figure 6, weighted by the Beta prior; Kaplan, 2021). Important DIF effects that were not shrunken to near zero by the lasso prior will move their inclusion parameter estimates closer to one from the prior mean 0.5, which indicates the presence of a non-zero effect. Unnecessary DIF effects, in contrast, will be shrunken close to zero by the lasso prior and cannot move the inclusion parameter away from its prior mean 0.5.[8]All penalty and inclusion estimates are estimated simultaneously, which means DIF evaluation does not require repeated model fitting across items and DIF covariates. Note that the use of a single SSP specification requires that covariates be scaled similarly, so that the penalty is equivalent in strength across covariates. We recommend rescaling time and other continuous covariates to have sample means of zero and standard deviations of 0.5 before fitting the penalized model (SD=0.5 so that transformed continuous covariates are on a similar scale as categorical variables; Gelman, 2008).

[8] When only inclusion parameters but not lasso penalties are present, such as in some regression

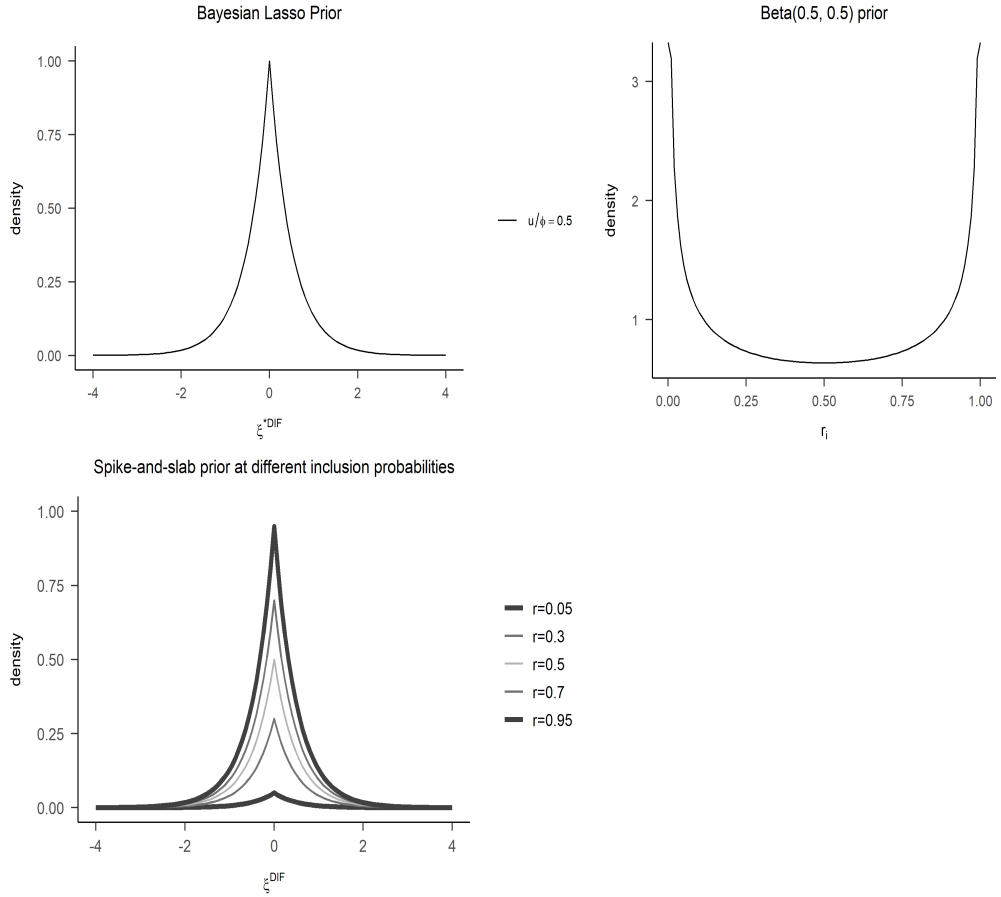


Figure 6

Density Plots of Components of Spike-and-slab Prior. Above Left: Bayesian Lasso prior conditional on a magnitude of penalty. Above Right: Beta prior for the inclusion parameter. Below: Spike-and-slab prior densities conditional on different levels of mean prior inclusion probabilities. Densities whose inclusion probabilities receive more weights from the Beta prior have thicker and darker lines; the final SSP distribution is a weighted mixture of such conditional densities.

In applying this approach, all DIF parameters are penalized with $p(\boldsymbol{\xi}^{dif}|\boldsymbol{\phi}, \boldsymbol{r})$ and estimated. Independent SSP inclusion parameters are estimated for DIF effects from different covariates. We also estimate independent lasso penalties on intercept DIF and slope DIF, $\boldsymbol{\phi} = \{\phi_{\boldsymbol{\omega}_i}, \phi_{\boldsymbol{\kappa}_i}\}$. Each of these penalty parameters has its own gamma hyperprior distribution $p(\boldsymbol{\phi})$ with a large hyperprior mean, so that the resulting SSP produces sufficient shrinkage to constrain the scale of DIF estimates and achieve approximate model identification (e.g., Brandt et al., 2018; Leng et al., 2014). The rest of the model parameters can follow generic, relatively diffuse priors.

After non-zero DIF effects have been identified with SSP inclusion estimates exceeding the designated threshold, it will often be useful to re-estimate the longitudinal MNLFA model with only the selected effects. This re-estimated model includes only the identified DIF effects, now with generic rather than penalizing priors, while excluding any DIF effects not detected by SSP. This removes bias due to SSP penalties which shrink even needed effects toward zero, so that proper growth estimates can be recovered (Belzak & Bauer, 2020; Hastie et al., 2009). We also suggest keeping the scales of the time variable and other predictors in this re-estimated model to close to those in the penalized model, to remove any extraneous influence of covariate scales on model estimation performance.

In sum, we can employ Bayesian estimation with SSP regularizing priors to efficiently estimate the model while simultaneously conducting DIF detection without the need for pre-specified anchors. After fitting the regularized longitudinal MNLFA model to reveal DIF effects in the sample, we can then re-estimate the model without penalty priors and with the identified pattern of DIF to mitigate bias due to penalization. We now turn to an empirical example to demonstrate the utility of this proposed method for describing and predicting construct change amidst potential changes in construct measurement.

---

examples (e.g., Bainter et al., 2020), the posterior estimates of inclusion parameters can move further below 0.5 for unimportant coefficients.

Empirical Example

Our example considers the development of nonviolent delinquency over adolescence. We seek to identify individual differences in change over time in nonviolent delinquency while accounting for potential changes in how this construct is measured over the 10-year age span in our analysis. We also consider potential sex differences in the expression of delinquent behavior over time. The analysis is focused on delinquency items administered from Waves 1 to 3 of the National Longitudinal Study of Adolescent to Adult Health (Add Health; Harris et al., 2019) and focuses on self-weighting core dataset, which tracked a nationally representative sample of adolescents from grade school to adulthood starting in 1994. Participants were between grades 7-12 in the first wave of assessment, the second wave of assessment came one year later, and the third wave was taken another 6 years later when subjects were young adults. Our analysis sample included $N = 6004$ individuals from the self-weighting core sample, who met the definition of adolescence (World Health Organization, 2019) by being at most 19 years of age at the first wave and provided at least partial item-level data on delinquent behaviors. Previous analyses by Bauer (2017) applied a cross-sectional MNLFA to data from the first wave of Add Health to examine age trends and sex differences. Our analysis differs by considering longitudinal data that extends over a longer age range, allowing for us to make inferences about within-person changes over time, an opportunity not available in cross-sectional data. We selected 9 measurement items on nonviolent delinquency behaviors for analysis and, similar to Bauer (2017), dichotomized these responses given low endorsement of the upper categories. Item descriptions and positive response rates are shown in Table 2 and 3 below. Four of these nine items were only administered at the first and second waves in the study.

Following the procedures described above, DIF detection was performed by fitting the Bayesian penalized longitudinal MNLFA with SSP regularizing priors on the DIF effects. Subsequently, the longitudinal MNLFA was re-estimated using an unpenalized model but including only DIF effects detected in the first step. Both models were

estimated with the *rstan* package (Stan Development Team, 2020) using an Intel Xeon E5 CPU on a remote computing cluster.[9]Online supplemental materials provide more detailed documentation of the code used to process the data, fit the models, and obtain results for the re-estimation step.

Table 2

Delinquency Behavior Items Used in Empirical Analysis

Item	Name	Descriptions
1	Graffiti	In the past 12 months, how often did you paint graffiti or signs on someone else's property or in a public place?
2	Damage property	In the past 12 months, how often did you deliberately damage property that didn't belong to you?
3	Lie	In the past 12 months, how often did you lie to your parents or guardians about where you had been or whom you were with?
4	Car	How often did you drive a car without its owner's permission?
5	Stealing>$50	In the past 12 months, how often did you steal something worth more than $50?
6	House	How often did you go into a house or building to steal something?
7	Drug	How often did you sell marijuana or other drugs?
8	Stealing<$50	How often did you steal something worth less than $50?
9	Unruly	How often were you loud, rowdy, or unruly in a public place?

Model Specification and DIF Evaluation Results

The SSP-regularized longitudinal MNLFA model used in DIF evaluation included the following potential DIF effects on items. Following Bauer (2017), an evaluation of these

[9] The regularized model took 45 hours and 42 minutes to fit; the re-estimated model took 49 hours and 21 minutes. The re-estimated model with numerous DIF effects specified tended to be slow to converge empirically.

Table 3

Percentage of Positive Endorsement and Sample Size by Age in Data.

age	N	Graffiti	Damage property	Lie	Car	Stealing>$50	House	Drug	Stealing<$50	Unruly
12	9	0	0.11	0.22	0.11	0	0.11	0	0.11	0.11
13	541	0.08	0.17	0.33	0.03	0.03	0.03	0.01	0.13	0.42
14	1323	0.09	0.19	0.42	0.06	0.04	0.06	0.04	0.2	0.45
15	1756	0.1	0.2	0.51	0.1	0.05	0.05	0.06	0.19	0.48
16	1956	0.09	0.17	0.54	0.12	0.05	0.05	0.08	0.19	0.46
17	2015	0.08	0.18	0.56	0.11	0.07	0.05	0.1	0.19	0.46
18	1995	0.06	0.13	0.52	0.08	0.04	0.04	0.09	0.15	0.42
19	1174	0.05	0.12	0.39	0.09	0.04	0.04	0.09	0.11	0.37
20	716	0.03	0.11	0.27	0.11	0.05	0.03	0.1	0.1	0.32
21	736	-	0.11	-	-	0.03	0.02	0.09	0.09	-
22	763	-	0.09	-	-	0.02	0.01	0.07	0.07	-
23	794	-	0.08	-	-	0.03	0.01	0.06	0.06	-
24	746	-	0.06	-	-	0.03	0.02	0.06	0.06	-
25	402	-	0.06	-	-	0.01	0.01	0.04	0.03	-
26	95	-	0.04	-	-	0.03	0.01	0.09	0.08	-

Note. Items not included in Wave III have empty cells.

DIF effects allowed us to examine possible linear or nonlinear changes in the manifestation of delinquent behaviors over time, as well as any sex differences in such manifest change:

$$\begin{aligned} \nu_{itp} &= \nu_{i0} + \kappa_{0i} sex_p + \kappa_{1i} age_{tp} + \kappa_{2i} {age^2}_{tp} + \kappa_{3i} age \times sex_{tp} + \kappa_{4i} age^2 \times sex_{tp} \\ \lambda_{itp} &= \lambda_{i0} + \omega_{0i} sex_p + \omega_{1i} age_{tp} + \omega_{2i} {age^2}_{tp} + \omega_{3i} age \times sex_{tp} + \omega_{4i} age^2 \times sex_{tp} \end{aligned} \tag{16}$$

where ${age^2}_{tp}$ and $age^2 \times sex_{tp}$ are DIF effects of quadratic age and its interaction with sex. All items in the measurement component (Equation 7) of the penalized model included these effects, with 4 DIF coefficients per intercept or loading. All DIF effects (but not baseline item parameters) were penalized with SSP (Equation 15). DIF effects on the intercept and the loading from each covariate were assigned one shared inclusion parameters to improve DIF evaluation accuracy on the item level (S. M. Chen et al., 2022). Lasso penalty parameters differed based on if they were given to DIF effects on intercepts

versus loadings and if they were for DIF effects from sex, age, or other higher-order effects; that is, a total of 6 lasso penalty parameters were used for each item. Together, SSP for each item consisted of the following sets of penalty and selection parameters in the regularized model

$$\begin{aligned}&\{\boldsymbol{r} : r^{age}, r^{sex}, r^{age^2}, r^{age\times sex}, r^{age^2\times sex}\}\\&\{\boldsymbol{\phi} : \phi_{\kappa}^{age}, \phi_{\omega}^{age}, \phi_{\kappa}^{sex}, \phi_{\omega}^{sex}, \phi_{\kappa}^{higher\text{-}order}, \phi_{\omega}^{higher\text{-}order}\}.\end{aligned} \tag{17}$$

A quadratic growth model allowing for sex differences in the means and (co)variances of the growth factors, the quadratic trend, and the error variance, was specified as follows:

$$\eta_{tp} = \beta_{0p} + \beta_{1p} age_{tp} + \beta_{2p} age^2{}_{tp} + \beta_{3p} age^2 \times sex_{tp} + e_{tp},\ e_{tp} \sim N(0, \sigma^2_{sex}) \tag{18}$$

$$\begin{aligned}\beta_{0p} &= \alpha_0 + \gamma_0 sex_p + \zeta_{0p}\\ \beta_{1p} &= \alpha_1 + \gamma_1 sex_p + \zeta_{1p}\\ \beta_{2p} &= \alpha_2\\ \beta_{3p} &= \alpha_3\end{aligned} \tag{19}$$

$$\begin{bmatrix}\zeta_{0p}\\ \zeta_{1p}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\psi_{00sex} & \\ \psi_{10sex} & \psi_{11sex}\end{bmatrix}\right). \tag{20}$$

where the *sex* subscript on the error variance and the growth factor covariance matrix indicated that these parameters differed between boys and girls. No growth factor was included for the quadratic trend as random quadratic effects are often negligible. We identified the model by setting the growth intercept to zero and residual error variance to one for the reference group (Male).

In this regularized model we first transformed age (Mean=17.41, SD=2.90) to have approximately a sample mean of zero and SD of 0.5, by centering age at 17.5 and dividing by 6, before calculating the quadratic and interaction terms. The penalty parameters had a $Gamma(4000, 200)$ hyperprior. At its mean, this hyperprior results in a DIF parameter

whose prior distribution has an SD of 0.05. Item baseline and growth factor mean and variance parameters were assigned normal priors with SD=2; structural effects from covariates in the growth model were given weak normal priors with SD=1.5. These priors represented the likely magnitudes that these parameters tend to have, and helped to facilitate model convergence. Baseline variances of growth factors (ψ_{000} and ψ_{110} in 10) and baseline item slope parameters were constrained positive, following typical Bayesian factor model practices to avoid sign indeterminancy of the estimates (e.g., Bainter, 2017; Fox & Glas, 2001). Full details on model priors are shown in Appendix A. The model was estimated with 4 MCMC chains and 3000 iterations per chain. Convergence was reached with all R-hat statistics below 1.01 and effective sample sizes on inclusion probability parameters above 3000.

Table 4 documents posterior estimates of inclusion probability parameters. Given the large magnitudes of some inclusion estimates, consistent with prior research by S. M. Chen et al. (2022), we chose a 0.8 inclusion threshold and highlighted estimates above this value. These highlighted effects constituted our DIF selection results. Items that exhibited DIF included *damage property* (sex, age, and age^2), *lie to parents* (sex and age^2), *driving a car without permission* (age), *stealing goods over $50* (age), *sell drugs* (sex, age, and age^2), and *being unruly* (sex and age-sex). It is also possible to vary the inclusion threshold value as a sensitivity analysis. For example, if we adopt a 0.7 threshold instead, the pattern of DIF is largely unchanged except for including age^2 DIF on *driving a car without permission* and age^2 $\times$ sex interaction DIF on *lie to parents*. Using the .8 threshold, 4 out of 9 items displayed DIF from sex, while 5 items displayed a linear or quadratic form of DIF related to age. Overall, there were sufficiently many items showing invariance over age for us to confidently infer construct change over time, and sufficiently many showing invariance to sex to confidently assess sex differences in trajectories of delinquency.

Table 4

Inclusion Parameter Estimates.

Item	Name	Sex	Age	Age^2	Age×sex	Age^2×sex
1	Graffiti	0.46	0.66	0.53	0.5	0.5
2	Damage property	0.95	0.89	0.9	0.53	0.49
3	Lie	0.97	0.47	0.98	0.64	0.72
4	Car	0.46	0.87	0.72	0.52	0.53
5	Stealing>$50	0.49	0.93	0.47	0.59	0.49
6	House	0.45	0.47	0.5	0.48	0.51
7	Drug	0.95	0.98	0.95	0.47	0.54
8	Stealing<$50	0.45	0.69	0.48	0.47	0.48
9	Unruly	0.92	0.51	0.49	0.85	0.55

Note. Values above 0.8 are highlighted.

Item and Growth Results from the Re-estimated Model

Next, we re-estimated the model without penalty but with only the detected DIF effects from the penalized model included, so that we can remove estimation bias due to the penalizing priors and obtain more accurate estimates of growth in the latent construct of delinquency. For items with only higher-order DIF effects, all relevant lower-order terms were included in the re-estimated model (e.g., treating Item 3 as having both age and age^2 DIF.) As a result, the model had 4 DIF coefficients from sex, 6 coefficients from age, 3 coefficients from age-squared, and 1 coefficient from age-sex interaction. Three items (*graffiti*, *broke into house*, and *stealing less than $50*) were anchors with respect to both age and sex. For interpretation, age in years was again centered at 17.5 and divided by 6 before fitting the model, the same as the penalized model. The trajectory intercept thus represents the expected level of delinquency at age 17.5, and the trajectory slope represents expected change in delinquency per 6-year change in age. The re-estimated model used mildly informative standard normal priors on the DIF and impact parameters, consistent with Brandt et al. (2023), to avoid non-convergence due to model complexity (see Appendix A).[10]The model used 4 MCMC chains with 2500 iterations per chain.

Convergence was reached with R-hat below 1.01 and minimum effective sample size above 600 for all model parameters.

With item and DIF parameter estimates from the re-estimated model, we can consider differences in the behavioral expressions of delinquency, by examining model-implied item estimates for individuals at different ages or between boys and girls. For example, for male individuals aged from 16 to 20, Item 7 *sell drugs* is expected to have an item intercept with increasing values (from −3.16 to −1.99) and a loading with decreasing values (from 0.97 to 0.80). This observation suggests that 1) these individuals are increasingly likely to respond positively to this item over time even when holding constant underlying general delinquency, and 2) this behavior was more relevant to the nonviolent delinquency construct at younger ages and became less so over time. At any given age, the item was a more severe and relevant indicator for girls than boys. To conserve space, a complete discussion of item parameter estimates, including DIF estimates, for this re-fitted model are provided in Appendix B. Here we instead turn to consideration of the estimates from the structural model that capture changes over time in the construct of delinquency.

Table 5 presents the estimates for the growth model parameters. Results indicated a quadratic trend in delinquency trait over time. Average delinquency behaviors peaked between 15 and 16 years old and then started to decrease through young adulthood. Additionally, the average level of delinquency observed for women was lower than that observed for men. The variance-covariance parameters describing within-group individual differences were similar between men and women, with slightly less variability observed for women in intercepts (age 17.5 levels of delinquency) and more in slopes (rates of change).

---

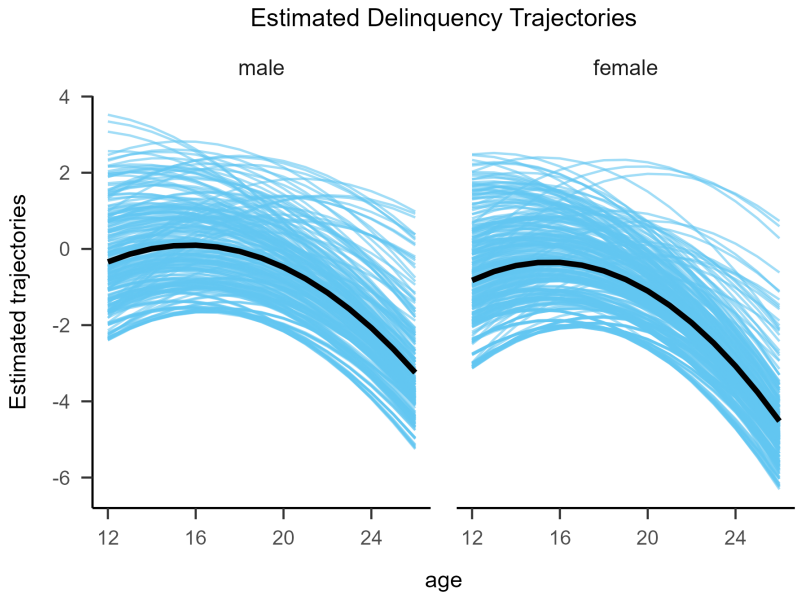
[10] Undesirable shrinkage effects on the results from such priors should be minimal given the large sample size. We conducted model re-estimation in the same data in which we evaluated DIF effects, because we assume the current dataset is representative enough to be our population of interest, rather than a sample to generalize our conclusions from.

To visualize these effects, the top panel of Figure 7 presents a sample of 500 individual predicted trajectories over the observed age range, along with the model-implied average trajectory for each group.

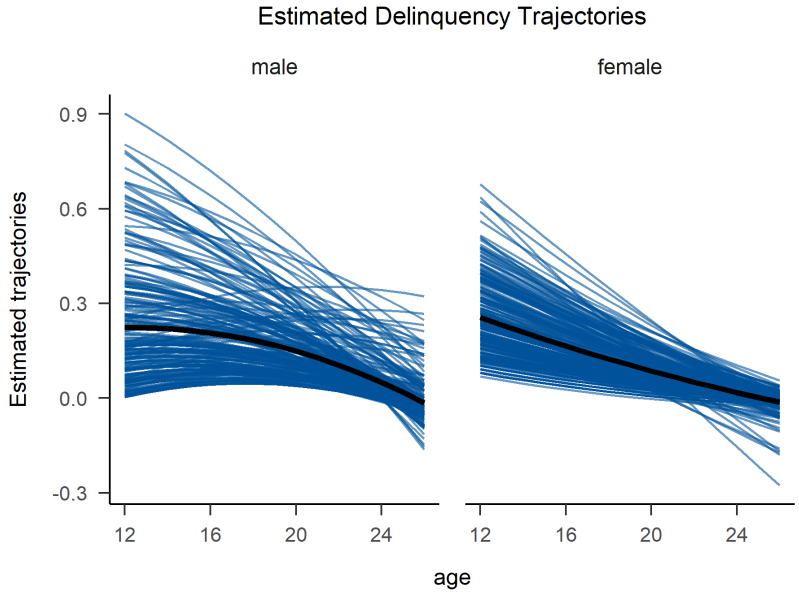
To highlight the utility of the longitudinal MNLFA relative to "business as usual" we also compared our results to those obtained by fitting a standard first-order growth model to mean score computed for each person at each time point given the available items (fit using restricted maximum likelihood estimation in SAS 9.4 PROC MIXED SAS Institute, 2018). This mean score model estimated sex-specific random effect covariance matrices but implicitly assumed constancy of measurement. Further, the availability of different items at different time points is accounted for only by dividing by the total number of available items to obtain the mean score. The lower panel of Figure 7 presents the predicted individual trajectories obtained by fitting the same growth model to the mean scores. As can be seen, by attributing all changes in item responses to latent construct change, this mean score model produced attenuated trend and variability estimates with an almost invisible quadratic age estimate (-0.046 for male and 0.013 for female). In relation to the longitudinal MNLFA, the mean-score model would result in vastly different conclusions about the developmental process under study. This attests to the importance of accounting for measurement change in longitudinal modeling and the utility of the proposed model.

Discussion

When working with observed scale scores, it is difficult, as Bereiter (1963) suggested, to validly distinguish true changes in the level of a construct from changes in the measurement/manifestation of that construct. To tease these apart requires a modeling approach that allows one to capture within-person change over time while simultaneously evaluating and accommodating potential shifts in the meaning/relevancy of items. Traditional second-order growth curve models offer this opportunity, coupling together a time-specific measurement model for the items with a model for growth and change over time for the underlying construct. However, treating time as categorical in the formulation



(a) *Longitudinal MNLFA trajectories*



(b) *Mean score model trajectories*

Figure 7

Add Health Estimated Latent Trajectories. Population mean trajectories are in black. Individual trajectories from 500 *individuals are in blue. Top: longitudinal MNLFA; bottom: mean score model.*

Table 5

Add Health Growth Parameter Estimates.

Fixed Effect	*Coefficient*	*95 % Credible Interval*
Intercept (fixed), α_0	0.000	-
Age, α_1	-0.675	(-0.90, -0.47)
Age^2, α_2	-1.141	(-1.38, -0.90)
Age$^2\times$ Sex, α_3	-0.233	(-2.30, 1.77)
Sex on Intercept, γ_0	-0.492	(-0.66, -0.33)
Sex on Slope, γ_1	-0.221	(-2.26, 1.86)
Random Effect	*Variance Component*	*95 % Credible Interval*
Intercept Male, ψ_{00}	1.992	(1.64, 2.39)
Age Male, ψ_{11}	1.813	(1.20, 2.55)
Covariance Male, ψ_{10}	-0.072	(-0.38, 0.23)
Residual variance Male (fixed), σ^2	1.000	-
Intercept Female, ψ_{00}	1.789	(1.45, 2.17)
Age Female, ψ_{11}	2.541	(1.84, 3.33)
Covariance Female, ψ_{10}	-0.088	(-0.38, 0.21)
Residual variance Female, σ^2	0.436	(0.32, 0.57)

of the SGC measurement model, with a factor specified at each unique time point, imposes a number of practical limitations on the use of these models, with deficiencies in specification (DIF parameters proliferate with time points but cannot accommodate continuous covariates), estimation (tractable only with a small number of unique time points), and interpretation (unsystematic changes in item properties over time). We proposed a longitudinal MNLFA approach to address the limitations of SGC. Our model can easily incorporate (1) more unique timepoints than measurement occasions and (2) DIF from both categorical and continuous person-level background variables, as well as their interaction with time. The consistent treatment of time as continuous both in the measurement model for the items and in the change model for the construct enhances parsimony (typically requiring fewer DIF parameters than SGC), estimation (no additional latent dimensions when people are measured at different times), and interpretation (item properties change continuously with time).

Similar to the case of SGC, to make confident inferences about growth in

longitudinal MNLFA one must identify a sufficient number of items without DIF (i.e., anchor items) to link the scale of the latent trait across timepoints. In principle, only one anchor item is required, but empirically identifying correct anchors becomes increasingly difficult as their number decreases (Yoon & Millsap, 2007). To identify and fit the longitudinal MNLFA model, we suggest the use of Bayesian estimation, which offers greater computational efficiency and enables us to implement regularizing prior distributions on the DIF parameters. We suggest a Bayesian regularization method called the SSP, which allows us to evaluate DIF simultaneously across all items without declaring anchor items and without error-prone sequential significance testing. After determining the location of DIF with the regularized model, we then suggest refitting the model with that DIF pattern but replacing the regularizing priors with weakly informative priors. This removes any shrinkage bias on the DIF parameters, allowing us to obtain more accurate estimates, with a particular interest in the structural model estimates for the growth process.

Although our emphasis has been on achieving an accurate understanding of change at the construct level while accounting for potential changes in measurement, it is important to note that the identified changes in measurement will often be of interest in their own right, and may even be the central focus of the research. In developmental psychology and related fields, for example, the term *heterotypic continuity* refers to a construct which shows continuous changes in its level over time but is expressed differently at different ages due to maturation, the development of new capabilities, differences in affordances within the social context, etc. (e.g., Kagan, 1969; Kagan & Moss, 1983). Petersen et al. (2020) points out that understanding stability versus change in the behavioral expressions of a construct helps researchers to establish when and how certain behaviors are normative versus deviant. Evaluating heterotypic continuity is essential in establishing construct validity and addressing theoretical gaps in many areas of research, such as inferring a personality attribute from diverse behaviors (Caspi & Shiner, 2007) or identifying patterns of internalizing/externalizing disorders from childhood to

adolescence (Cicchetti & Rogosch, 2002; Speranza et al., 2023). The longitudinal MNLFA model provides a means to address substantive questions like these in ways previously not possible.

It should be noted, nevertheless, that longitudinal MNLFA brings an additional assumption that the functional form of DIF is correctly specified (e.g., DIF over time can be described with a linear function). If the functional form of DIF is misspecified, this could lead to bias in the model estimates. We attempted to consider this assumption by including quadratic terms of age as a potential source of DIF in the empirical example, so that the SSP-regularized DIF evaluation model may detect nonlinear trends in item parameters. Other alternatives, such as a semiparametric approach to model functional forms in MNLFA (Molenaar, 2021), could be extended to the longitudinal context in future research.

To our knowledge, this manuscript is the first to present the longitudinal MNLFA for modeling growth over time. We believe this model holds significant advantages for addressing Bereiter (1963)'s question, how are we to separate change in measurement from true change in the underlying construct? The longitudinal MNLFA offers a new way to address this question, without the deficiencies of the traditional second-order growth model. Further, our demonstrations show that the model can be feasibly applied and yields substantively compelling results. Of course, much additional research will be needed to evaluate the performance of this new modeling approach under a variety of data analytic conditions, including comprehensive simulations examining accuracy of DIF detection and recovery of growth estimates. We hope that this manuscript provides the foundation both for this future research and for applications of the longitudinal MNLFA in practice, enhancing our ability to make valid inferences about change over time amidst changes in measurement.

References

Andersen, E. B. (1985). Estimating latent correlations between repeated testings [Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Springer-Verlag]. *Psychometrika*, *50*(1), 3–16. https://doi.org/10.1007/BF02294143

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277–300. Retrieved December 4, 2019, from www.jstor.org/stable/1435429

Bainter, S. A. (2017). Bayesian estimation for item factor analysis models with sparse categorical indicators. *Multivariate Behavioral Research*, *52*(5), 593–615. https://doi.org/10.1080/00273171.2017.1342203

Bainter, S. A., McCauley, T. G., Wager, T., & Losin, E. A. R. (2020). Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain [Publisher: SAGE Publications Inc]. *Advances in Methods and Practices in Psychological Science*, *3*(1), 66–80. https://doi.org/10.1177/2515245919885617

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*(3), 870–897. https://doi.org/10.1214/009053604000000238

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2019). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–13. https://doi.org/10.1080/10705511.2019.1642754

Bauer, D. J., & Curran, P. J. (2016). The discrepancy between measurement and modeling in longitudinal data analysis. In *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 3–38). IAP Information Age Publishing.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101–125. https://doi.org/10.1037/a0015583

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning [Publisher: American Psychological Association]. *Psychological Methods*, *25*(6), 673–690. https://doi.org/10.1037/met0000253

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. Harris (Ed.), *Problems in measuring change* (pp. 3–20). University of Wisconsin Press.

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, *9*(1), 30–52. https://doi.org/10.1037/1082-989X.9.1.30

Bollen, K., & Curran, P. (2006). *Latent curve models: A structural equation perspective*. Wiley. https://books.google.com/books?id=CDnG15sPvigC

Brandt, H., Cambria, J., & Kelava, A. (2018). An adaptive bayesian lasso approach with spike-and-slab priors to identify multiple linear and nonlinear effects in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(6), 946–960. https://doi.org/10.1080/10705511.2018.1474114

Brandt, H., Chen, S. M., & Bauer, D. J. (2023). Bayesian penalty methods for evaluating measurement invariance in moderated nonlinear factor analysis [Publisher: American Psychological Association]. *Psychological Methods*. https://doi.org/10.1037/met0000552

Browne, W. J., & Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514. https://doi.org/10.1214/06-BA117

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Caspi, A., & Shiner, R. L. (2007). Personality development [Section: 6 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470147658.chpsy0306]. In *Handbook of child psychology*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470147658.chpsy0306

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM) [Publisher: SAGE Publications Inc]. *Organizational Research Methods*, *1*(4), 421–483. https://doi.org/10.1177/109442819814004

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models [Publisher: Routledge _eprint: https://doi.org/10.1207/s15328007sem1203_7]. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 471–492. https://doi.org/10.1207/s15328007sem1203_7

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2022). Advantages of spike and slab priors for detecting differential item functioning relative to other bayesian regularizing priors and frequentist lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(1), 122–139. https://doi.org/10.1080/10705511.2021.1948335

Cicchetti, D., & Rogosch, F. A. (2002). A developmental psychopathology perspective on adolescence. *Journal of Consulting and Clinical Psychology*, *70*(1), 6–20. https://doi.org/10.1037/0022-006X.70.1.6

Cronbach, L. J., & Furby, L. (1970). How we should measure 'change': Or should we? [Publisher: American Psychological Association]. *Psychological Bulletin*, *74*(1), 68–80. https://doi.org/10.1037/h0029382

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., Sher, K., & Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate behavioral research*, *49*(3), 214–231. https://doi.org/10.1080/00273171.2014.889594

Depaoli, S., & Clifton, J. P. (2015). A bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling*, *22*(3), 327–351. https://doi.org/10.1080/10705511.2014.937849

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 45–97. Retrieved December 4, 2019, from www.jstor.org/stable/2346087

Duncan, S. C., & Duncan, T. E. (1996). A multivariate latent growth curve analysis of adolescent substance use [Publisher: Routledge _eprint: https://doi.org/10.1080/10705519609540050]. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(4), 323–347. https://doi.org/10.1080/10705519609540050

Duncan, T. E., & Duncan, S. C. (2004). An introduction to latent growth curve modeling. *Behavior Therapy*, *35*(2). https://doi.org/10.1016/S0005-7894(04)80042-X

Feng, X.-N., Wu, H.-T., & Song, X.-Y. (2017). Bayesian regularized multivariate generalized latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 341–358. https://doi.org/10.1080/10705511.2016.1257353

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with mantel-haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. https://doi.org/10.1177/0146621605275728

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika*, *66*(2), 271–288. https://doi.org/10.1007/BF02294839

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873. https://doi.org/10.1002/sim.3107

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013, November 1). *Bayesian data analysis, third edition* [Google-Books-ID: ZXL6AQAAQBAJ]. CRC Press.

George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.2307/2290777

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs [Publisher: SAGE Publications Inc]. *The Journal of Applied Behavioral Science*, *12*(2), 133–157. https://doi.org/10.1177/002188637601200201

Grimm, K. J., Kuhl, A. P., & Zhang, Z. (2013). Measurement models, estimation, and the study of change [Publisher: Routledge _eprint: https://doi.org/10.1080/10705511.2013.797837]. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 504–517. https://doi.org/10.1080/10705511.2013.797837

Hancock, G. R., Kuo, W.-L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models [Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1207/S15328007SEM0803_7]. *Structural*

Equation Modeling: A Multidisciplinary Journal, *8*(3), 470–489. https://doi.org/10.1207/S15328007SEM0803_7

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*(4), 835–845. https://doi.org/10.1093/biomet/asp047

Hans, C. (2010). Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing*, *20*(2), 221–229. https://doi.org/10.1007/s11222-009-9160-9

Harris, K. M., Halpern, C. T., Whitsel, E. A., Hussey, J. M., Killeya-Jones, L. A., Tabor, J., & Dean, S. C. (2019). Cohort profile: The national longitudinal study of adolescent to adult health (add health). *International Journal of Epidemiology*, *48*(5), 1415–1415k. https://doi.org/10.1093/ije/dyz115

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction, second edition* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/978-0-387-84858-7

Hoffman, L., Hofer, S. M., & Sliwinski, M. J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: A simulation study. *Psychology and Aging*, *26*(4), 778–791. https://doi.org/10.1037/a0023910

Horn, J. L., & Mcardle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. https://doi.org/10.1080/03610739208253916

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *71*(3), 499–522. https://doi.org/10.1111/bmsp.12130

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773. Retrieved November 10, 2019, from https://www.jstor.org/stable/3448605

Jacobucci, R., & Grimm, K. J. (2018). Comparison of frequentist and bayesian regularization in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 639–649. https://doi.org/10.1080/10705511.2017.1410822

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Kagan, J. (1969). The three faces of continuity in human development. *Handbook of socialization theory and research*, 983–1002.

Kagan, J., & Moss, H. A. (1983). *Birth to maturity: A study in psychological development*. Yale University Press. Retrieved February 16, 2024, from https://www.jstor.org/stable/j.ctt1dszxm7

Kaplan, D. (2021). On the quantification of model uncertainty: A bayesian perspective. *Psychometrika*, *86*(1), 215–238. https://doi.org/10.1007/s11336-021-09754-5

Kim, E. S., & Willson, V. L. (2014). Testing measurement invariance across groups in longitudinal data: Multigroup second-order latent growth model [Publisher: Routledge _eprint: https://doi.org/10.1080/10705511.2014.919821]. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 566–576. https://doi.org/10.1080/10705511.2014.919821

Kolbe, L., Molenaar, D., Jak, S., & Jorgensen, T. D. (2022). Assessing measurement invariance with moderated nonlinear factor analysis using the r package OpenMx. *Psychological Methods*. https://doi.org/10.1037/met0000501

Kruschke, J. (2010, November 25). *Doing bayesian data analysis: A tutorial introduction with r*. Academic Press.

Kuhfeld, M., & Soland, J. (2020). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses [Publisher: American Psychological Association]. *Psychological Methods*. https://doi.org/10.1037/met0000367

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, *60*(1), 65–81. Retrieved February 14, 2020, from https://www.jstor.org/stable/25053023

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411. https://doi.org/10.1214/10-BA607

Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items [Publisher: Routledge _eprint: https://doi.org/10.1080/10705510701575438]. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 581–610. https://doi.org/10.1080/10705510701575438

Leng, C., Tran, M.-N., & Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, *66*(2), 221–244. https://doi.org/10.1007/s10463-013-0429-6

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures [Publisher: American Psychological Association]. *Psychological Methods*, *22*(3), 486–506. https://doi.org/10.1037/met0000075

Liu, Y., & West, S. G. (2018). Longitudinal measurement non-invariance with ordered-categorical indicators: How are the parameters in second-order latent linear growth models affected? [Publisher: Routledge _eprint: https://doi.org/10.1080/10705511.2017.1419353]. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 762–777. https://doi.org/10.1080/10705511.2017.1419353

Lu, Z.-H., Chow, S.-M., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research*, *51*(4), 519–539. https://doi.org/10.1080/00273171.2016.1168279

Lykou, A., & Ntzoufras, I. (2013). On bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing*, *23*(3), 361–390. https://doi.org/10.1007/s11222-012-9316-x

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111–135. https://doi.org/10.3102/1076998614559747

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 561–614). Springer US. https://doi.org/10.1007/978-1-4613-0893-5_17

McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary psychological research. In *Handbook of psychology: Research methods in psychology, vol. 2.* (pp. 447–480). John Wiley & Sons, Inc. https://doi.org/10.1002/0471264385.wei0218

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Meredith, W. (1991). Latent variable models for studying differences and change. In *Best methods for the analysis of change: Recent advances, unanswered questions, future*

directions (pp. 149–169). American Psychological Association. https://doi.org/10.1037/10099-010

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

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis Group.

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach [Place: US Publisher: American Psychological Association]. *Journal of Applied Psychology*, *73*(3), 574–584. https://doi.org/10.1037/0021-9010.73.3.574

Molenaar, D. (2021). A flexible moderated factor analysis approach to test for measurement invariance across a continuous variable [Publisher: American Psychological Association]. *Psychological Methods*, *26*(6), 660–679. https://doi.org/10.1037/met0000360

Muthén, B. O., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*, 313–335.

Muthén, L. K., & Muthén, B. (1998–2017). *Mplus user's guide: Statistical analysis with latent variables, user's guide* (8.4). Muthén & Muthén.

Narisetty, N. N., & He, X. (2014). Bayesian variable selection with shrinking and diffusing priors [Publisher: Institute of Mathematical Statistics]. *Annals of Statistics*, *42*(2), 789–817. https://doi.org/10.1214/14-AOS1207

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(2), 107–124. https://doi.org/10.1080/10705519809540095

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The bayesian lasso. *Psychological Methods*, *22*(4), 687–704. https://doi.org/10.1037/met0000112

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal rasch modeling in the context of psychotherapy outcomes assessment [Publisher: SAGE Publications Inc]. *Applied Psychological Measurement*, *30*(2), 100–120. https://doi.org/10.1177/0146621605279761

Petersen, I. T., Choe, D. E., & LeBeau, B. (2020). Studying a moving target in development: The challenge and opportunity of heterotypic continuity. *Developmental Review*, *58*, 100935. https://doi.org/10.1016/j.dr.2020.100935

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191. https://doi.org/10.1080/01621459.1997.10473615

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to u.s. high-school data [Publisher: American Educational Research Association]. *Journal of Educational Statistics*, *16*(4), 295–330. https://doi.org/10.3102/10769986016004295

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566.

SAS Institute. (2018). Base sas 9.4 procedures guide.

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In *New methods for the analysis of change* (pp. 179–200). American Psychological Association. https://doi.org/10.1037/10409-006

Shi, D., Song, H., Liao, X., Terry, R., & Snyder, L. A. (2017). Bayesian SEM for specification search problems in testing factorial invariance. *Multivariate Behavioral Research*, *52*(4), 430–444. https://doi.org/10.1080/00273171.2017.1306432

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8317.1974.tb00543.x]. *British Journal of Mathematical and Statistical Psychology*, *27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Speranza, A. M., Liotti, M., Spoletini, I., & Fortunato, A. (2023). Heterotypic and homotypic continuity in psychopathology: A narrative review. *Frontiers in Psychology*, *14*. Retrieved February 16, 2024, from https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1194249

Stan Development Team. (2020). RStan: The R interface to Stan [R package version 2.21.8]. http://mc-stan.org/

Stark, S., Drasgow, F., & Chernyshenko, O. S. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified approach. *Journal of Applied Psychology*, 1292–1306.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402–415. https://doi.org/10.1037/1082-989X.11.4.402

Stevens, A. K., Janssen, T., Belzak, W. C. M., Treloar Padovano, H., & Jackson, K. M. (2022). Comprehensive measurement invariance of alcohol outcome expectancies among adolescents using regularized moderated nonlinear factor analysis. *Addictive Behaviors*, *124*, 107088. https://doi.org/10.1016/j.addbeh.2021.107088

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates, Inc.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in rasch models. *Psychometrika*, *80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. (2014). A gentle introduction to bayesian analysis: Applications to developmental research [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cdev.12169]. *Child Development*, *85*(3), 842–860. https://doi.org/10.1111/cdev.12169

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*(2), 318–336. https://doi.org/10.1007/s11336-011-9202-z

Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 455–465. https://doi.org/10.1080/10705511.2015.1096744

Willett, J. B. (1997). Measurement of change. In J. Keeves (Ed.), *Educational research, methodology and measurement: An international handbook* (2nd). Pergamon Press.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *116*(2), 363–381. https://doi.org/10.1037/0033-2909.116.2.363

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. https://doi.org/10.1177/0146621607314044

World Health Organization. (2019). *Adolescent health*. Retrieved December 1, 2023, from https://www.who.int/health-topics/adolescent-health

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A monte carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 435–463. https://doi.org/10.1080/10705510701301677

Appendix A

Model Priors

In our empirical analysis, the regularized DIF parameters in the SSP model were given spike-and-slab priors based on Equation (15) and (17) with $u = 1$. The sex effects on correlation and residual variance were expected to be small within the exponential functions and thus given standard normal priors.Other parameters in the SSP-regularized model had the following priors.

$$\begin{aligned}
&\nu_{i0} \sim N(-2, 2);\ \lambda_{i0} \sim N^{+}(1, 2);\\
&\alpha_1 \sim N(0, 2);\\
&\gamma_0, \gamma_1, \alpha_2, \alpha_3 \sim N(0, 1.5);\\
&(\psi_{00})^{1/2}, (\psi_{11})^{1/2} \sim N^{+}(0, 2);\\
&\gamma_{00}^{(SD)}, \gamma_{11}^{(SD)} \sim N(0, 1.5);\\
&\rho_{100}, \rho_{101}, \gamma_e \sim N(0, 1)
\end{aligned} \tag{A1}$$

where $N^{+}(\cdot)$ indicates that the parameter was given a normal prior and then constrained positive (i.e., a half-normal prior) to avoid sign indeterminancy, following typical Bayesian factor model practices.

In the re-estimated model, the DIF effect parameters were assigned $N(0, 1)$ priors. Other model parameters used the following priors.

$$\begin{aligned}
&\nu_{i0} \sim N(-2, 1.5);\ \lambda_{i0} \sim N^{+}(1, 1.5);\\
&\alpha_1, \gamma_0, \gamma_1, \alpha_2, \alpha_3 \sim N(0, 1.5);\\
&(\psi_{000})^{1/2}, (\psi_{110})^{1/2} \sim N^{+}(0, 1.5);\\
&\gamma_{00}^{(SD)}, \gamma_{11}^{(SD)} \sim N(0, 1.5);\\
&\rho_{100}, \rho_{101}, \gamma_e \sim N(0, 1)
\end{aligned} \tag{A2}$$

Appendix B

Add Health Re-estimated Item and DIF Effects

Table B1 presented estimated baseline item parameters and DIF effects. The table showed item values implied for the male group at age 16, 17, and 20, which represented age values at the 25th, 50th, and 75th percentiles, respectively. Item tracelines illustrating age DIF for item 3, 5, and 7 are presented in Figure B1.

Through these implied item values it is possible to track change in behavioral manifestations. For example, as age increased, Item 2 *damage property* was perceived as less severe/difficult (as a larger intercept means a smaller trait level needed to achieve a 50% response probability) due to extraneous reasons such as growing physical strength. It also became less discriminative among people with different delinquency levels, which indicates that damaging property is less relevant a behavioral indicator for delinquency in young adults. A similar pattern can be observed in Item 7 *sell drugs*. Using the current model we can also extrapolate item characteristics even if an item is omitted from a wave. Item 3 *lie to parents* initially became less difficult from age 16 to 17. At age 20, however, the model predicted that this item would become less relevant to delinquency behaviors as individuals no longer need to routinely seek permissions from parents; this was reflected in a lower implied loading value. A more negative implied item intercept for Item 3 at age 20 corresponded to the observation that this item already had a very low positive response rate for individuals at age 20 in Wave II (Table 3). These implied item estimates justify the decision to omit this item in Wave III.

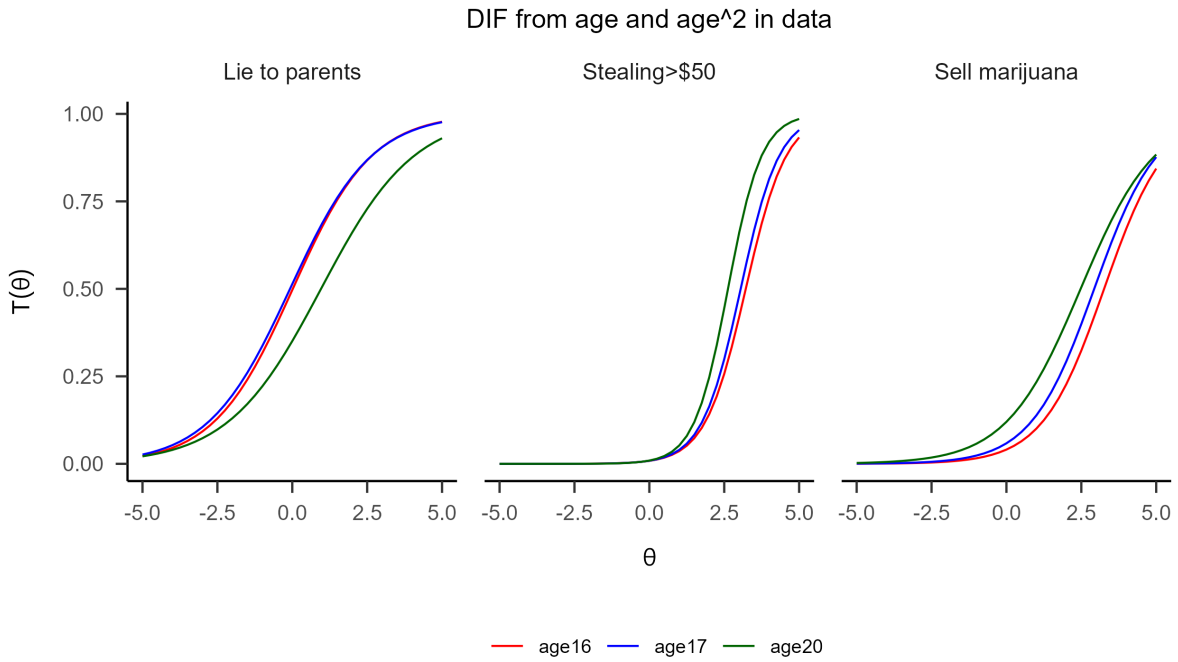


Figure B1

Add Health Implied Item Tracelines for Item 3, 5, and 7.

Table B1

Add Health Item and DIF Estimates.

Parameter	Baseline	Age DIF	Sex DIF	Age^2 DIF	Age-Sex DIF	Age 16 Implied	Age 17 Implied	Age 20 Implied
Graffiti								
int	-3.56	-	-	-	-	-3.56	-3.56	-3.56
load	1.21	-	-	-	-	1.21	1.21	1.21
Damage property								
int	-2.29	0.14	-0.53	0.50	-	-2.29	-2.30	-2.15
load	1.34	-0.26	-0.08	-0.15	-	1.40	1.36	1.21
Lie								
int	0.03	-0.47	0.64	-2.63	-	-0.01	0.05	-0.62
load	0.72	-0.16	0.12	-0.05	-	0.76	0.73	0.64
Car								
int	-2.66	0.82	-	-	-	-2.87	-2.73	-2.32
load	0.86	0.08	-	-	-	0.84	0.85	0.89
Stealing>$50								
int	-4.72	0.12	-	-	-	-4.75	-4.73	-4.67
load	1.59	0.45	-	-	-	1.48	1.55	1.78
House								
int	-5.07	-	-	-	-	-5.07	-5.07	-5.07
load	1.59	-	-	-	-	1.59	1.59	1.59
Drug								
int	-2.61	1.94	-0.82	-1.11	-	-3.16	-2.78	-1.99
load	0.93	-0.21	0.25	-0.23	-	0.97	0.95	0.80
Stealing<$50								
int	-2.38	-	-	-	-	-2.38	-2.38	-2.38
load	1.29	-	-	-	-	1.29	1.29	1.29
Unruly								
int	-0.23	0.29	0.18	-	-0.71	-0.30	-0.26	-0.11
load	0.81	0.03	0.15	-	0.03	0.80	0.81	0.82

Note. DIF estimates are on a transformed age scale. Implied item estimates are for Group 1 (male).