Adjusting for Measurement Noninvariance With Alignment in Growth Modeling

Mark H. C. Lai[1]

[1] Department of Psychology, University of Southern California

Author Note

Mark H. C. Lai https://orcid.org/0000-0002-9196-7406

This work was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) and was accomplished under Grant #W911NF-20-1-0282. The views, opinions, and/or findings contained in this paper are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents. I would like to thank Yichi Zhang for helping with the literature search and Hailin Yue for preparing the data for the Applied Example.

This is an Accepted Manuscript of an article published by Taylor & Francis in Multivariate Behavioral Research on July 8, 2021, available online: http://www.tandfonline.com/10.1080/00273171.2021.1941730.

Correspondence concerning this article should be addressed to Mark H. C. Lai, 3620 S. McClintock Ave, Los Angeles, CA 90089. E-mail: hokchiol@usc.edu

Abstract

Longitudinal measurement invariance—the consistency of measurement in data collected over time—is a prerequisite for any meaningful inferences of growth patterns. When one or more items measuring the construct of interest show noninvariant measurement properties over time, it leads to biased parameter estimates and inferences on the growth parameters. In this paper, I extend the recently developed alignment-within-confirmatory factor analysis (AwC) technique to adjust for measurement biases for growth models. The proposed AwC method does not require a priori knowledge of noninvariant items and the iterative searching of noninvariant items in typical longitudinal measurement invariance research. Results of a Monte Carlo simulation study comparing AwC with the partial invariance modeling method show that AwC largely reduces biases in growth parameter estimates and gives good control of Type I error rates, especially when the sample size is at least 1,000. It also outperforms the partial invariance method in conditions when all items are noninvariant. However, all methods give biased growth parameter estimates when the proportion of noninvariant parameters is over 25%. Based on the simulation results, I conclude that AO is a viable alternative to the partial invariance method in growth modeling when it is not clear whether longitudinal measurement invariance holds. The current paper also demonstrates AwC in an example modeling neuroticism over three time points using a public data set, which shows how researchers can compute effect size indices for noninvariance in AwC to assess to what degree invariance holds and whether AwC results are trustworthy.

Keywords: measurement invariance, factorial invariance, longitudinal, alignment optimization, growth model

Word count: 7,403

Adjusting for Measurement Noninvariance With Alignment in Growth Modeling

Longitudinal data allow researchers to make inferences on changes across time due to natural events, developmental maturation, or carefully designed interventions. In social and behavioral sciences, researchers have used growth modeling to examine changes across multiple waves of data in alcohol misuse in adolescence (Barnes et al., 2000), correlates of growth of vocabulary production during toddlerhood (Pan et al., 2005), and the role of age stereotypes on memory performance over time in late adulthood (Levy et al., 2012), to name just a few examples. However, for the results of growth modeling to be valid, the operationalization of constructs should remain the same across waves; otherwise, any observed differences across time can be confounded by incompatible measurements (e.g., Shadish et al., 2001).[1] Even when the same instrument is being used across time, in the presence of various developmental and cultural changes, the measurement properties of an instrument may shift over time, introducing bias to the analyses. Therefore, *longitudinal measurement invariance*, the condition that an instrument measures one or more constructs in the same way across time, is required for growth modeling results to be meaningful (Grimm et al., 2016; Horn & McArdle, 1992; Widaman et al., 2010).

Given that measurement in behavioral sciences is usually imprecise, it is not uncommon to find violations of longitudinal measurement invariance for psychological instruments. For example, Obradović et al. (2007) found that an instrument measuring interpersonal callousness did not maintain its measurement properties after four years in a 9-year longitudinal study with a group of boys considered "antisocial." Wu et al. (2009) found that two items in a scale measuring life satisfaction did not satisfy longitudinal invariance over six months in two samples of university students in Taiwan. Blankson and McArdle (2013) tested longitudinal invariance of six cognitive tests in a representative longitudinal study of U.S. participants in their 50s, and their results failed to support longitudinal invariance for the mental status factor across a period of 18 years. Finally, Lommen et al. (2014) found that a posttraumatic stress scale did not maintain the same measurement properties before and after deployment in two groups of Dutch soldiers,

[1] See, for example, Curran and Hussong (2009), Petersen et al. (2020), Tyrell et al. (2019), for alternative approaches to harmonize different instruments intended to measure the same construct across time.

leading the authors to question whether the same construct was measured before and after deployment using the same scale.

Violations of longitudinal measurement invariance, which I also simply refer to as *noninvariance*, do not mean that research questions on change cannot be answered. At least when the degree of violation is mild to moderate, one established strategy is to estimate the degree of bias by identifying a partial invariance model, and adjust that bias in a second-order growth model that specifies the relations between observed indicators and the latent construct at each wave (to be discussed later in this paper; see Ferrer et al., 2008; Widaman et al., 2010). However, the identification of a partial invariance model usually requires many iterations of model fitting and modifications, which potentially capitalizes on chance (MacCallum et al., 1992) and requires substantially more efforts than the growth model itself.

On the other hand, an alternative approach is to use the newly developed alignment optimization (AO) technique (Asparouhov & Muthén, 2014) in multiple-group analysis to come up with an approximate invariance model (to be discussed later), which requires fitting only one measurement model. More recently, Marsh et al. (2018) extended the alignment method to an approach called alignment-within-confirmatory factor analysis (AwC), which incorporates AO into a multiple-group regression model to obtain estimations of latent regression parameters adjusted for violations of invariance. However, to my knowledge, there has been no previous research extending the AO procedure in the context of longitudinal measurement invariance as it is not currently implemented in major structural equation modeling (SEM) software.

The purpose of the current paper is four-folded. First, I propose a simple solution to extend AO to longitudinal invariance. Second, I extend the AwC approach to growth modeling to obtain adjusted inferences on the growth parameters when there are violations of measurement invariance. Third, I report the results of a Monte Carlo simulation study to evaluate the proposed method across conditions of sample size, degree of noninvariance, average growth rates, and model specification. Finally, I illustrate with an applied example how my proposed method can be easily implemented in the R software.

Longitudinal Factorial Invariance

I first define the longitudinal factor model used for the current discussion, which is based on the discussion of Meredith and Horn (2001). Specifically, for a study with T waves with one construct η measured by p indicators $\mathbf{y} = [y_1, \ldots, y_p]'$, there are pT manifest variables, and the longitudinal factor model can be defined as

$$\mathbf{y}_t = \boldsymbol{\nu}_t + \boldsymbol{\lambda}_t \eta_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $t = 1, \ldots, T$ indexes waves, $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ contains the factor loadings (regression weights; also called pattern coefficients) and measurement intercepts of the linear prediction from η, and $\boldsymbol{\varepsilon}$ contains both the stable, construct-irrelevant specific factors and the random measurement error; I denote εs as unique factors in the current study following Grimm et al. (2016).

It is assumed that $\boldsymbol{\varepsilon}$ is independent to η as it does not capture the construct of interest, and the components of $\boldsymbol{\varepsilon}$ are jointly normal with expected values of 0.[2] In addition, researchers usually make the local independence assumption so that $\text{Var}(\boldsymbol{\varepsilon}_t) = \Theta_{\varepsilon t}$ at a given wave t is a diagonal matrix of uniqueness with elements $\theta_{\varepsilon 1}, \ldots, \theta_{\varepsilon p}$. On the other hand, because some determinants of unique factors are stable across time for the same item, it is common to allow unique factor covariances across waves such that $\text{Cov}(\varepsilon_{jt}, \varepsilon_{jt'}) \neq 0$ for $t \neq t'$ and all $j = 1, \ldots, p$.

Under the above factor model, the measurement parameters linking $\mathbf{y}$ and η are $\boldsymbol{\lambda}_t$s, $\boldsymbol{\nu}_t$s, and $\boldsymbol{\Theta}_t$s. Therefore, strict *factorial invariance*, meaning measurement invariance under the factor model, requires that $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}$, $\boldsymbol{\nu}_t = \boldsymbol{\nu}$, and $\boldsymbol{\Theta}_t = \boldsymbol{\Theta}$ for all ts (Meredith, 1993). In practice, however, such a condition rarely holds, and so researchers commonly follow the popular approach by Widaman and Reise (1997) to test four stages of factorial invariance:

1. Configural invariance (Horn & McArdle, 1992; Horn et al., 1983), where $\boldsymbol{\lambda}_t$ contains the same zero elements across waves; this is automatically satisfied when dealing with a

[2] Whereas it is reasonable to assume that the random measurement error has an expected value of 0, the same assumption is less reasonable for the specific factors as they may change across time in a developmental process. However, the means of the specific factors can be absorbed into the measurement intercepts so that the model can still hold. This is a potential source of intercept noninvariance.

unidimensional construct;

2. Weak invariance (also metric/pattern invariance; Millsap, 2011), where $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}$ for all ts;
3. Strong invariance (also scalar invariance), where $\boldsymbol{\nu}_t = \boldsymbol{\nu}$ for all ts in addition to weak invariance; and
4. Strict invariance, where $\boldsymbol{\Theta}_t = \boldsymbol{\Theta}$ for all ts in addition to strong invariance.

As shown in Ferrer et al. (2008), at least strong invariance is required to assure that observed changes in the means of the manifest variables, which is usually the focus in growth modeling, are not confounded with changes in measurement properties of the instrument (i.e., noninvariance). Otherwise, researchers may wrongly conclude that there are meaningful changes in the target construct over time, when indeed the changes in observed scores are driven by noninvariant loadings and/or intercepts of a few items. Therefore, many scholars (e.g., Grimm et al., 2016; Horn & McArdle, 1992; Widaman et al., 2010) have suggested that researchers establish factorial invariance of their measurement before performing growth modeling. As previously discussed, however, strong invariance generally does not hold, at least not exactly, so methods to adjust for noninvariance are needed.

Partial invariance—traditional method to adjust for noninvariance. The traditional method to adjust for noninvariance is to search for a partial strong invariance model (e.g., Byrne et al., 1989; Yoon & Millsap, 2007), where invariant parameters are constrained to be equal across time while noninvariant parameters are freely estimated. As previously demonstrated (e.g., Lai et al., 2021), as long as the proportion of actual noninvariant parameters is not large, this approach would work reasonably well. Besides, it provides valuable information regarding which items on a scale showed large violations of invariance. Such an approach, however, has several drawbacks. First, there is a risk of capitalization on chance as it requires iteratively testing parameter constraints, which may lead to an unstable solution (e.g., MacCallum et al., 1992). Second, it requires a lot of effort in locating noninvariant parameters. When the number of time points, indicators, and/or constructs is large, researchers may need to manually fit tens or hundreds of models to arrive at a partial invariance model. This may also lead to lots of researcher degrees of freedom that make results potentially not replicable (Chambers, 2019).

Third, as demonstrated in Marsh et al. (2018), this specification search approach may lead to large bias and imprecision in parameter estimates and inferences when the proportion of noninvariant parameters is relatively large (e.g., more than 1/3 or half).

The above-listed drawbacks are potential reasons that the partial longitudinal invariance model is not commonly used in the literature. A quick search of articles published in *Child Development* in 2018–2019 showed 21 articles that used growth modeling in the SEM framework, but only one (4.7%) used a second-order growth model that potentially adjusted for measurement errors and biases.

AO and AwC. An alternative approach to the factorial invariance problem is the alignment optimization (AO) method proposed by Asparouhov and Muthén (2014) for multiple-group structural equation modeling. To understand AO, first note that due to factor indeterminacy (Kline, 2016), in the configural invariance model with one latent variable per wave, each wave requires one constraint to identify the variance-covariance structure and one constraint to identify the mean structure. There are infinitely many possible sets of identification constraints, such as (a) fix the latent means and variances to 0 and 1, respectively, for all waves; (b) fix the latent mean and variance to 0 and 1, respectively, for the first wave, and constrain the loadings and intercepts of the first indicator to be invariant across waves. Both (a) and (b) place $2 \times T$ identification constraints to the model and give the same model fit and the same expectation and covariances of $\mathbf{y}$, as do infinitely many other possible sets of constraints. However, they correspond to different latent means and variances, factor loadings, and intercepts values, and have different implications of factorial invariance.

AO aims to achieve a set of measurement parameter estimates that retain large noninvariances while keeping other parameters approximately invariant across groups. It uses a component loss function to “align” the parameters so that the latent variables are on similar metrics and are thus comparable. Such an optimization problem is similar to the rotation problem in exploratory factor analysis (EFA) aiming to achieve a simple structure that retains large loadings while minimizing small loadings. An additional set of constraints to scaling the latent variables is to fix the mean and variance of the latent variable to 0 and 1, respectively, for the first group.[3]

With T sets of measurement parameters and assuming equal sample sizes across waves, the component loss function with respect to the parameter differences of a set of aligned loadings and intercepts, $\lambda_{t,a}$ and $\nu_{t,a}$, is defined as

$$F = \sum_{j=1}^{p} \sum_{t_1 < t_2} f(\lambda_{jt_1,a} - \lambda_{jt_2,a}) + \sum_{j=1}^{p} \sum_{t_1 < t_2} f(\nu_{jt_1,a} - \nu_{jt_2,a}). \tag{2}$$

While there could be many options for the loss function f for the differences in individual parameters across waves, Asparouhov and Muthén (2014) proposed the use of

$$f(x) = \sqrt{\sqrt{x^2 + \epsilon}}, \tag{3}$$

which has been found to work very well for multiple-group analyses with many groups (e.g., 15–60 groups in Marsh et al., 2018) and with a few groups (e.g., 2–4 groups in Lai et al., 2021), using a small ϵ such as 0.01 or 0.001. Readers can find a numerical example in the Appendix, which further illustrates the component loss function.

Much like in EFA where different rotation methods give the same model-implied correlation/covariance matrix, when using AO, traditional fit indices in CFA, like the root-mean-square error of approximation (RMSEA) and the comparative fit index (CFI), are not sensitive to different alignment solutions, as the aligned loadings and intercepts have exactly the same fit as the configural model. That said, fit indices should still be informative to other aspects of model misspecification in the AwC growth model, such as unique covariances or nonlinear growth shape. However, researchers should supplement fit indices with effect size indices for noninvariance, such as the d_{MACS} index discussed in the Simulation Study section, to assess to what degree longitudinal factorial invariance holds.

Currently, AO can only be applied in confirmatory factor analytic (CFA) models without any imposed structures on the latent variables or any external covariates or outcome variables.

[3] This was denoted as "fixed" alignment in Asparouhov and Muthén (2014), which is suitable in the current paper as then the growth parameters can be interpreted as standard deviation unit of the first occasion. Another option is "random" alignment which sets the average of the means across groups/occasions to zero.

However, Marsh et al. (2018) proposed a two-step alignment-within-CFA (AwC) procedure that greatly enhanced the usefulness of AO. After obtaining the aligned measurement parameters using AO in the first step, in the second step of structural modeling, AwC requires fixing one loading and one intercept for each latent variable to be equal to the solution in AO so that the metric of the latent variables will be similar to that from the AO solution. Thus, parameters found to have large differences across groups in AO are kept as such, so that theoretically the resulting structural parameter estimates (e.g., latent means and variances) will be less confounded with measurement bias. Lai et al. (2021) conducted a simulation study and found that AwC performs well in terms of precision and confidence interval (CI) coverage rates for latent path coefficients across sample size and degree of noninvariance conditions.

To my knowledge, however, until now the applications of AO has been limited to multiple-group analyses, as the software Mplus (L. K. Muthén & Muthén, 2017), which first implemented AO, does not support AO with longitudinal factorial invariance at the time of writing. On the other hand, it is straightforward to extend AO to longitudinal measurement models by applying the same optimization algorithm on the loadings and intercepts obtained from a longitudinal configural invariance model with longitudinal data. After that, AwC can be used to adjust for noninvariance using a second-order growth model, as reviewed below.

Second-Order Growth Model

Growth modeling aims to model the trajectory of one or more constructs over time. In a commonly adopted linear growth model, each individual's trajectory is described by two person-specific parameters: level (initial status) and slope (growth rate). Traditionally, and still a popular practice, researchers use first-order growth models (Ferrer et al., 2008), meaning that the construct is represented by a single composite score in each wave. Such an approach, however, can lead to erroneous estimations and inferences in the presence of (a) measurement unreliability, and (b) measurement noninvariance. For (a), it is well known that failure to account for unreliability biases structural coefficient estimates (Cole & Preacher, 2014; Kenny, 1979). For (b), as demonstrated in Ferrer et al. (2008), strong invariance is needed to establish a meaningful comparison of construct means across time, which is a prerequisite to interpret the growth parameters meaningfully.

Unfortunately, a first-order growth model does not allow for the evaluation and adjustment of (a) and (b). To fully capitalize on the capability of structural equation modeling, a second-order growth model can instead be used by replacing the single composites with longitudinal factor models of multiple indicators. Such a model imposes a growth structure on the η variables in (1). Specifically, under the linear SEM framework,

$$\boldsymbol{\eta}_i = \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i, \tag{4}$$

where $\boldsymbol{\xi}_i$ contains r person-specific growth parameters for the ith person, $\boldsymbol{\Gamma}$ is a $T \times r$ matrix specifying the contrast codes for modeling time trend, and is usually fixed, and $\boldsymbol{\zeta}_i$ contains latent disturbances of deviations from the predicted trajectory with $\mathrm{E}(\zeta) = 0$ for all persons and waves. For example, with a linear growth model, there are $r = 2$ person-specific growth parameters, and usually

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & T-1 \end{bmatrix}$$

Figure 1 shows a path diagram for a linear growth model with four waves. It is commonly assumed that conditioning on the growth parameters $\boldsymbol{\xi}$, ηs are normally and independently distributed so that $\mathrm{Var}(\boldsymbol{\eta}|\boldsymbol{\xi}) = \boldsymbol{\Psi} = \mathrm{diag}(\psi_{11}, \ldots, \psi_{TT})$. The growth parameters $\boldsymbol{\xi}$ are assumed multivariate-normally distributed with $\mathrm{E}(\boldsymbol{\xi}) = \mathbf{K}$ and $\mathrm{Var}(\boldsymbol{\xi}) = \boldsymbol{\Phi}$.

The benefits of a second-order growth model are that it takes into account measurement unreliability (Hancock et al., 2001) and, through modeling of partial strong invariance, adjusts for violations of longitudinal noninvariance, so that the resulting growth parameter estimates are less biased (E. S. Kim & Willson, 2014; Leite, 2007). However, as previously pointed out, the use of a partial strong invariance model is only valid when researchers do not mistakenly constrain any noninvariant parameters, and in practice, it may not work when the proportion of noninvariance is large (Marsh et al., 2018).

On the other hand, using AwC, one can model growth while adjusting for noninvariance using alignment, which does not require a priori knowledge of noninvariant parameters and an iterative process of searching for them. In the following, I first report results from a Monte Carlo simulation study evaluating the performance of the AwC approach in terms of parameter bias, efficiency, and confidence interval coverage. A step-by-step example of applying AwC using real data is then provided.

Simulation Study

I report how I determined my design conditions, the number of replications, and all evaluative measures of the simulation results. I used the `SimDesign` package (Chalmers, 2020; Chalmers & Adkins, 2020) in R (Version 4.0.3; R Core Team, 2020) to structure the simulation studies. The full simulation code can be found in the supplemental materials.

In the present simulation study, I evaluated the performance of the AwC approach for estimating a linear growth model with potential violations of factorial invariance. Based on previous simulations (e.g., E. S. Kim & Willson, 2014; M. Kim et al., 2016; Kwok et al., 2007; Liu & West, 2018), a typical latent growth model fitted in the literature has four waves, so I set $T = 4$ in my simulation. The data generating model is shown in Figure 1, following a linear growth pattern. Each latent response η is measured by five indicators (not shown in the Figure), which is similar to the design in Liu and West (2018) and E. S. Kim and Willson (2014). The measurement parameter values used to generate the data are shown in Table 1, and the growth parameter values are discussed in the design conditions. I kept the measurement parameters at Wave 1 the same across all simulations so that the scale remains constant. At Wave 1, the composite reliability is .806. Following Liu and West (2018), I also added a lag 1 autoregressive structure for each unique factor across waves with a lag 1 autocorrelation of .20, a lag 2 autocorrelation of $.20^2$, and so forth (i.e., $\text{Corr}[\varepsilon_{jt}, \varepsilon_{jt'}|\boldsymbol{\eta}] = .20^{|t-t'|}$).

Design Conditions

The current simulation has a 3 (sample size) × 3 (proportion of noninvariance) × 2 (average growth rate) × 2 (model misspecification) design, as described below.

Sample size (N). From the review by Kwok et al. (2007), the mean sample size of longitudinal studies published in *Developmental Psychology* was 210 (SD = 180), whereas from the meta-analysis by Huang (2011) on the relationship between self-concept and academic achievement in 39 longitudinal studies, the median sample size was 267. Therefore, I chose 100, 250, and 1,000 for our sample size conditions for small, medium, and large samples, which was similar to the conditions in E. S. Kim and Willson (2014).

Proportion of noninvariant parameters/items ($r_{\text{ni}}/p_{\text{ni}}$). I generated data with various $r_{\text{ni}}/p_{\text{ni}}$ conditions, where r_{ni} was defined as the proportion of noninvariant loadings and intercepts out of 40 parameters (i.e., 20 loadings + 20 intercepts), and p_{ni} was the proportion out of the five items that were invariant over time. Specifically, I manipulated r_{ni} to be 0%, 25%, and 55%, and the corresponding p_{ni} to be 0, 40%, and 100%. For conditions with $r_{\text{ni}} = 25\%/p_{\text{ni}} = 40\%$, I simulated item 5 to have large biases in loadings across all four waves (based on the criterion from Nye et al., 2018) and have small biases in intercepts for Waves 2 and 3, and item 4 to have large biases in intercepts across all four waves (see Table 1). For conditions with $r_{\text{ni}} = 55\%/p_{\text{ni}} = 100\%$, there was a mix of small, medium, and large biases in the intercepts and loadings, but more importantly, none of the five items were fully invariant across waves, which allows an examination of whether AwC can be a viable option with no invariant items.

Growth rate (κ_2). I set the average growth rate per wave, which is the mean of the linear slope factor, to be either 0 or 0.25. The level $\kappa_2 = 0$ was chosen to evaluate Type I error rates of the AwC procedure, while $\kappa_2 = 0.25$ corresponds to a medium growth rate.

The mean of the intercept factor was set to 0 without loss of generality. The variances of the intercept and the slope factor were 0.5 and 0.1, respectively, and the covariance between them was set to 0.089, which was consistent with E. S. Kim and Willson (2014). The error variances of η_1 to η_4 were equally set to 0.5. Therefore, at Wave 1, the intraclass correlation—the proportion of variance the intercept factor accounted for—was 0.5. When $\kappa_2 = 0.25$, the marginal R^2 effect size was 0.38 (Johnson, 2014).

Model misspecification. In practice, researchers rarely have data that perfectly fit the data well. Therefore, I had two sets of conditions for model misspecification, where the generated data either followed exactly or deviated slightly from the model in equations (1) and (4). For conditions with model misspecification, after generating the η values based on equation (4), I added a small quadratic trend such that

$$\eta^*_{it} = \eta_{it} + (t - 2.5)^2 \xi_{3i}, \tag{5}$$

where ξ_3 is the quadratic growth factor with mean = -0.01 and variance = 0.004. Besides, to resemble minor misspecification in the measurement model, I used a procedure similar to MacCallum and Tucker (1991) by adding minor unique factor covariances with magnitudes between -0.1 and 0.1 to the generated y values. The R code for generating the unique factor covariances can be found in the supplemental materials. Overall, the misspecification corresponds to a population RMSEA of .057.

Data Generation

For each simulation condition, I used R to simulate 2,500 data sets, which was sufficient to keep the Monte Carlo error to ±2% of the parameter and *SE* estimates. It was also sufficient to keep the margin of error for empirical Type I error rates to 5% ± 0.5%, which satisfied the stringent criterion defined by Bradley (1978). For each condition I used the model defined in equations (1) and (4) (and equation 5 for conditions with misspecifications) to compute the marginal mean vector and covariance matrix of the 20 manifest variables, and used the `rmvn()` function from the *mvnfast* package (Fasiolo, 2016) in R to simulate multivariate normal data.

Data analysis. I used *lavaan* (Version 0.6.7; Rosseel, 2012) for all my analyses. For each simulated data set, I fitted (a) AwC, an AwC-growth model, (b) FI, a second-order growth model assuming full strong invariance that constrains all loadings and intercepts to be equal across waves, and (c) PI, a second-order growth model assuming partial strong invariance with equality constraints only on the unbiased items (except when $r_{\text{ni}} = .55$). When $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, all intercepts were noninvariant, so I placed the intercept equality constraints on the first item, which resembled the usual practice (see Shi et al., 2017) while allowing the loadings and the intercepts of the other items to be freely estimated without cross-wave constraints. For both FI and PI, the models were identified by fixing the loadings of the first item (which is assumed invariant) to 0.8 and the intercepts of that item to 0 across all waves, so that the scales of the latent variables are the same across replications. For AwC, I fixed the loadings and the intercepts of the first item in each wave to the values based on the alignment solution. For all methods, I constrained the error variance associated with the ηs to be equal (i.e., $\psi_{11} = \ldots = \psi_{44}$). Maximum likelihood estimation for multivariate normal data was used for all methods.

For each method I obtained point and *SE* estimates and the 95% Wald CI reported from *lavaan* for the means and variances of the level and slope growth factors. For each parameter θ (i.e., means and variances of levels and slopes), the evaluative measures were described below.

Evaluative Measures

Bias. The bias was computed as $\bar{\hat{\theta}} - \theta$, where $\bar{\hat{\theta}} = \frac{\sum_{r=1}^{R} \hat{\theta}_r}{R}$ is the mean of the $\hat{\theta}_r$ estimates across 2,500 replications and θ is the population parameter value.

Root mean squared error (RMSE). Considering the bias-variance trade-off (e.g., Ledgerwood & Shrout, 2011), a slightly biased estimator may be preferred over a biased estimator if the former has a smaller sampling variance. Thus, for each method M I computed the RMSE of the parameters of interest, defined as

$$\text{RMSE}(\hat{\theta}^M) = \sqrt{\frac{\sum_{r=1}^{R} (\hat{\theta}_r^M - \theta)^2}{R}}.$$

A method with a smaller RMSE should be preferred.

Error rates of 95% CI. To evaluate the CIs based on AwC and PI, I computed the 95% Wald CI as $\hat{\theta} \pm z_{.975} \hat{SE}(\theta)$, where $z_{.975}$ is the quantile in a standard normal distribution corresponding to a probability of .975. For each parameter, the empirical error rates were calculated as the proportion of times the constructed CI failed to contain the population parameter value (i.e., 1 − coverage rates). A valid 95% CI should have an error rate of 5%. Note that when the population value of a parameter is zero, the 95% CI error rate is also the empirical Type I error rate of a Wald test with a 5% nominal significance level.

In addition, to estimate the proportion of parameters that are substantially noninvariant for each simulated data set, I computed the d_{MACS} effect size proposed by Nye and Drasgow (2011). The d_{MACS} effect size represents the standardized mean difference of each item across two groups or waves due to differences in loadings and intercepts. When an item is invariant across all groups/waves, $d_{\text{MACS}} = 0$, which is the minimum value. For example, if $d_{\text{MACS}} = 0.5$ for item 1 between Wave 1 and Wave 3, it means that the noninvariance in loadings and intercepts of item 1 across these two waves results in a mean difference of half a standard deviation. In the simulation, after obtaining the aligned loadings and intercepts, I computed d_{MACS} for each of the 30 pairwise comparisons (5 items, each with 6 contrasts of time points). As Nye et al. (2018) suggested a cutoff of $d_{\text{MACS}} < .20$ for negligible noninvariance, for each simulated sample I computed (a) the proportion of pairwise comparisons with $d_{\text{MACS}} > .20$ and (b) the proportion of items (out of five) with at least one $d_{\text{MACS}} > .20$. Sample R codes for computing d_{MACS} statistics after alignment can be found in the supplemental materials.

Results

In some replications there were warnings from *lavaan* that some estimated variances were negative or that the estimates resulted in non-positive definite covariance matrices of the latent or observed variables; however, for all replications in all simulation conditions, the fitted models converged, so we used all 2,500 replications to summarize the results.

Mean level (κ_1). As shown in Figure 2, when all items were invariant (i.e., $r_{\text{ni}} = 0$), all three methods (PI, FI, and AwC) were unbiased when there were no misspecifications, but had a small downward bias of about -0.01 when there were misspecifications (i.e., unmodelled quadratic trend and unique covariances). The biases were relatively stable across sample size conditions. The estimates under PI and AwC were not affected when $r_{\text{ni}} = .25/p_{\text{ni}} = .40$, but the estimates under FI started to show upward bias as it falsely constrained some noninvariant parameters to be equal. When $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, PI showed the worst bias ($M_{\text{bias}} = 0.21$) as it anchored on a noninvariant item; FI showed large biases ($M_{\text{bias}} = 0.15$), whereas AwC was also biased but to a much lesser degree ($M_{\text{bias}} = 0.04$).

The RMSEs and error rates of 95% CIs for estimating the mean level were shown in Table 2. When $r_{\text{ni}} = 0$ or 0.25, PI was generally more efficient than AwC, but the difference was not large. On the other hand, when $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, PI was the least efficient. When $r_{\text{ni}} > 0$, AwC generally had a smaller RMSE than FI, which was largely driven by the bias of FI. For CI error rates, AwC generally maintained error rates < 5% when $r_{\text{ni}} \leq .25$ even in the presence of misspecification, and its error rates ($M_{\text{err}} = 3.38\%$) tended to be lower than those based on PI ($M_{\text{err}} = 5.12\%$); FI had large CI error rates when r_{ni} and sample size increased as it did not yield a consistent estimate. When $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, AwC also had increased CI error rates when sample size increased, indicating that it also did not provide a consistent estimate, but the error rates were much smaller than those under PI and FI.

In addition, there were some counterintuitive results as the CI error rates were smaller when model misspecification was present and when $r_{\text{ni}} = .55/p_{\text{ni}} = 1$. Upon further investigation, such results were likely due to wider sample CIs when data were simulated with misspecification (8% wider for PI and 12% wider for AwC).[4]

Mean slope (κ_2). The bias of estimating κ_2 is summarized in Figure 3, which depends on its population value. For conditions with $r_{\text{ni}} = 0$ and $r_{\text{ni}} = .25/p_{\text{ni}} = .40$, PI yielded unbiased estimates, whereas AwC estimates showed small positive biases when $r_{\text{ni}} = .25/p_{\text{ni}} = .40$ and $\kappa_2 = .25$ (up to 0.03, or 12.56%), and FI estimates showed stronger biases (up to 0.07, or 28.54%). When $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, both PI (up to 37.03%) and FI (up to 49.68%) showed strong bias regardless of sample size; AwC was still biased but to a much lesser degree, with the largest bias of 28.21% when $N = 100$, but reduced to 14.80% when $N = 1{,}000$.

As shown in Table 3, like κ_1, the RMSE pattern was largely driven by the bias pattern. Similarly, in terms of CI error rates, AwC yielded CIs with the lowest error rates when $r_{\text{ni}} \leq .25$,

[4] This explanation was suggested by an anonymous reviewer.

regardless of model misspecifications. On the other hand, FI yielded highly inflated error rates when $r_{\text{ni}} > 0$. When $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, AwC had inflated error rates when $\kappa_2 = 0$ (i.e., Type I error rates) of up to 15.04%, but it was much better than FI, which had error rates of up to 98.44%, and PI, which had error rates of up to 98.76%. When $\kappa_1 = .25$, the CI error rates for all methods were much higher due to the larger biases in the estimates.

Level variance (ϕ_1). Figure 4 shows the relative bias (i.e., bias / ϕ_1) when estimating ϕ_1. Similar to the results for κ_1, when there were no misspecification both PI and AwC yielded estimates with little bias for conditions with $r_{\text{ni}} \leq .25$, but the misspecification led to an underestimation of about 10%. For FI, the underestimation was bigger. When $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, all methods suffered larger biases, but AwC yielded estimates with smaller bias (relative bias between -21.10% to -8.90%) than FI (relative bias between -28.40% to -11.29%) and PI (relative bias between -23.43% to -10.91%).

The RMSE patterns (Table 2) were similar to the ones for estimating mean level, with AwC generally performing better than falsely assuming invariance. For CI error rates, when there were no misspecifications, AwC had rates $< 5\%$ for all conditions with $r_{\text{ni}} \leq .25$, but increased to up to 12.16% when $r_{\text{ni}} = .55/p_{\text{ni}} = 1$. The error rates of FI increased as a function of r_{ni} and N and were much higher than AwC. The error rates of PI increased as a function of N when $r_{\text{ni}} = .55/p_{\text{ni}} = 1$ and were higher than AwC. When there were misspecifications, all methods in all conditions had increased error rates, but the error rates were lowest with AwC.

Slope variance (ϕ_2). Figure 4 shows the relative bias (i.e., bias / ϕ_2) when estimating ϕ_2. With the current simulation set up, the misspecifications generally resulted in downward biases for ϕ_2, whereas increasing r_{ni} resulted in upward biases for FI and AwC and downward biases for PI (when $r_{\text{ni}} = .55/p_{\text{ni}} = 1$), and the relative bias for ϕ_2 was larger than for ϕ_1. It was found that AwC had a larger bias than FI when $r_{\text{ni}} = .55/p_{\text{ni}} = 1$ and $N \leq 250$.

The RMSE patterns (Table 3) were similar to the ones for other parameters. The CI error rates tended to be above 5% for all methods even without misspecifications, and with misspecifications, AwC resulted in better control of error rates for all conditions with $r_{\text{ni}} \leq .25$. When $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, AwC actually had better error rates when there were misspecifications, mainly due to the compensatory effects of misspecifications and noninvariance resulting in smaller biases.

d_{MACS} effect size. It was found that when using the AwC method, the κ_1 and κ_2 estimates were acceptable when (a) less than 30% of the pairwise d_{MACS} was larger than .20 *AND* (b) less than 50% of the items had at least one $d_{\text{MACS}} > .20$. More details about the analyses with the d_{MACS} effect size can be found in the supplemental material.

Summary and Remarks

From the simulation, I found that AwC generally worked well in reducing bias on growth parameter estimates due to noninvariance, and performed best in terms of bias when the sample size is large (e.g., 1,000). It produces a slight loss of efficiency compared to the correctly specified partial invariance model when the proportion of noninvariant parameters is small but performs better than picking the wrong anchor item in a partial invariance model when the proportion of noninvariant parameters is large. It also generally shows better control on Type I error rates and CI coverage rates. Therefore, the proposed AwC growth method is a viable alternative to the traditional partial invariance approach. On the other hand, while using a correctly specified partial invariance model works well, it leads to the highest bias when it anchors on items with large noninvariance; in the current simulation, it performs worse than the strong invariance model, as in the latter noninvariance in different directions partially cancels out (see Horn & McArdle, 1992).

One limitation of the simulation is that it does not inform whether the magnitude of noninvariance, which was not a manipulated factor, would affect the results.[5] A supplemental simulation was conducted for the simulation conditions with $r_{\text{ni}} = .55/p_{\text{ni}} = 1$ but with the magnitude of noninvariance reduced by half, and the results can be found in the supplemental materials (https://github.com/marklhc/awc-growth-supp). In summary, the parameter bias was smaller for all methods with a smaller magnitude of noninvariance, but the overall pattern of the results was similar. The AwC method still performed better than PI and FI, but all methods

[5] An anonymous reviewer brought up this excellent point.

showed non-negligible biases when estimating the level and slope parameters.

Another issue of interest is whether the results of AwC depend on which indicator has the identification constraints,[6] which I here refer to as the reference indicator. Theoretically, because the loadings and the intercepts of the reference indicator are fixed to the corresponding values of the AO solution, the metric of the latent variables will remain similar, so the latent parameter estimates should be the same. However, if one chooses an indicator with weak loadings (i.e., close to 0), the metric of the latent variables will be only weakly identified, leading to larger standard errors of the latent parameters. In the simulation study, I followed Marsh et al. (2018) to use the first indicator as the reference indicator for AwC, which happened to be one with the largest loadings. To evaluate the sensitivity of AwC growth model results to choices of reference indicator, I reran the simulations using the second indicator (with loadings = .50) as the reference indicator. As expected, the parameter bias of AwC remained similar, but the constructed 95% CI was generally wider due to larger standard error estimates, leading to lower statistical power. Based on these results, a tentative recommendation is to choose an item with large loadings as the reference indicator, but future research is needed to determine the optimal choice of reference indicator.

Finally, based on the associations between the d_{MACS} effect size statistics and the estimated mean intercept and slope, I suggest a 50/30/20 rule of thumb for using d_{MACS} effect size statistics to gauge the appropriateness of the AwC method: AwC is trustworthy when (a) no more than 50% of items have one or more $d_{\text{MACS}} > .20$ and (b) no more than 30% of the pairwise $d_{\text{MACS}} > .20$.

Applied Example

The illustrative data come from Waves I (1995–1996), II (2004–2006), and III (2013–2014) of the Midlife in the United States project (MIDUS; Brim et al., 2020; Ryff, Almeida, Ayanian, Carr, et al., 2017; Ryff, Almeida, Ayanian, Binkley, et al., 2019). For the demonstration, I investigated how neuroticism changed over time. At each wave, participants indicated how each of the four words: "moody," "worry," "nervous," and "calm," described them on a 4-point scale (1 =

[6] Both the Associate Editor (Keith Widaman) and an anonymous reviewer brought up this excellent point.

A lot, 2 = *Some*, 3 = *A little*, 4 = *Not at all*). To make interpretations easier, I reverse-coded the first three items so that for all items, a higher score indicated higher neuroticism. For illustration, I used a subsample of participants who were 40 years old or below at Wave I; also, to simplify the illustration I included only participants with no missing data on all four items across all three waves, resulting in a subsample of 833 participants ($M_{\text{age}} = 33.79$). The descriptive statistics of each item can be found in the supplemental materials (https://github.com/marklhc/awc-growth-supp).

A longitudinal configural invariance model with three factors was first fitted to the 12 observed variables (four items across three waves), and unique covariances of the items across waves were allowed. To use AO, it is easiest to identify the configural model by fixing the latent factor variances to 1 and the latent means to 0.[7] The overall χ^2 test for this model was statistically significant, χ^2 (N = 833, df = 39) = 74.49, p < .001, indicating lack of exact fit. However, the model fit was acceptable using common standards, with CFI = .991, RMSEA = .033, 90% CI [.021, .044], and SRMR = .038. The factor loading and intercept estimates before alignment are shown in Table 4, which is not very meaningful as the model does not place the latent variables on similar metrics across waves. It should be emphasized that like other methods for evaluating measurement invariance, one needs to make sure the configural model demonstrates acceptable model fit before performing AO.

Using the loading and intercept estimates from the configural model, I obtained the aligned loadings and intercept estimates that minimized the component loss function, using the `invariance.alignment()` function from the *sirt* package (Robitzsch, 2020b) in R. The aligned solutions are also shown in Table 4, together with the aligned factor means and variances. From the aligned solution, the latent factor means were estimated to be -0.31 in Wave II and -0.27 in Wave III.

As shown in the simulation results, AwC may result in biased latent parameter estimates when the proportion of noninvariant parameters/items is large, as indicated by the d_{MACS}

[7] Other ways of identifying the model, such as fixing the latent factor variances to 1 and the latent means to 0 for Wave 1 while constraining the loadings and intercepts of the first item to be equal across waves, give identical model fit and lead to the same aligned solution.

statistics. In the neuroticism example, there were 12 pairwise comparisons, and two of them (16.7%) showed non-negligible d_{MACS}: "moody" for Wave 1 vs. Wave 3 (0.23), and "calm" for Wave 1 vs. Wave 2 (0.21); 50% of the items showed at least one $d_{\text{MACS}} > .20$. Based on the suggested 50/30/20 rule of thumb, I continue with the AwC growth model.

As shown in the supplemental materials, I fixed the loadings and intercepts of the second indicator ("worry"), which had the largest loadings overall, to the values from the AO solution for each wave (e.g., 0.79, 0.78, and 0.80 for loadings; 2.62, 2.63, and 2.62 for intercepts), and the resulting model fit was exactly the same as the unaligned configural model. I then fit a second-order linear latent growth model with the same minimum identification constraints. Based on the mean pattern, a linear growth model is probably not a good fit for the data, but I keep it for my illustration as the linear growth model as it is widely used. It should also be pointed out that other growth shapes can be easily applied, and readers can check out excellent resources by Grimm et al. (2016) and Newsom (2015), for example. The AwC growth model had an acceptable fit, χ^2 (N = 833, df = 40) = 101.70, $p < .001$, CFI = .985, RMSEA = .043, 90% CI [.033, .053], and SRMR = .041. Based on the parameter estimates, after adjusting for potential violations of factorial invariance, the mean slope estimate was -0.124, 95% CI [-0.166, -0.082], indicating an overall decreasing trend of about 0.124 *SD* in neuroticism per wave.

To illustrate the sensitivity to different reference indicators, I also fit an AwC growth model using "calm" as the reference indicator, which had the lowest loadings (0.32 to 0.36). This AwC growth model with alignment had a similar fit, χ^2 (N = 833, df = 40) = 81.03, $p < .001$, CFI = .990, RMSEA = .035, 90% CI [.024, .046], and SRMR = .039. The mean slope estimate was -0.127, which was similar to the estimate when using "worry" as the reference indicator, but the 95% CI [-0.215, -0.038] was wider.

The full R code for this example can be found in the supplemental materials (https://github.com/marklhc/awc-growth-supp).

Discussion

In growth models, for growth parameters to be meaningful, the quantification of the target construct must be consistent across time. Under the common factor model with continuous and normally distributed indicators, this means that strong factorial invariance needs to hold. When strong invariance is violated for some but not all items, falsely assuming invariance and using a full strong invariance model results in biased growth parameter (i.e., level and slope) estimates and the corresponding between-person variance estimates, as demonstrated in previous studies (Ferrer et al., 2008; Liu & West, 2018) and the current simulation. The empirical Type I error rates for the mean slope (i.e., CI error rates when the true slope is zero) increase as sample size and proportion of noninvariant parameters increase and approach 100% when N = 1,000. In other words, if noninvariance is not correctly accounted for, researchers are almost guaranteed to falsely detect significant growth or changes, when none exists.

One can also use a second-order growth model with a partial strong invariance model to adjust for the noninvariance, which performed well in my simulation when the proportion of noninvariant parameters is relatively small (e.g., $<$ 25%) and there are at least some truly invariant items. However, two major limitations of this approach is that (a) it requires either prior knowledge or an intensive iterative specification search process, which may capitalize on chance (MacCallum et al., 1992; Marsh et al., 2018), (b) it may lead to even worse bias when it anchors on the wrong item(s) (see Ferrer et al., 2008; Shi et al., 2017), and (c) it cannot be used when all items are noninvariant, based on our simulation results. All of them are potential reasons that the second-order growth model with adjustment of partial invariance has not been widely adopted.

In the current paper, I propose adapting the alignment optimization (AO) and the alignment-within-CFA (AwC) techniques, originally developed in multiple-group analyses, to growth modeling to adjust for longitudinal noninvariance. To my knowledge, the current paper is the first in demonstrating how AwC can be applied to longitudinal factor models.

The AwC growth method has several advantages. First, compared to searching for a partial invariance model, which usually requires many iterations of adding/relaxing constraints and examining modification indices or other fit indices, AwC only requires fitting a longitudinal configural model, performing alignment optimization, and fitting a second-order growth model. Therefore, it presents less burden for applied researchers and avoids problems that different researchers may use different cutoffs for freeing invariance constraints. Second, unlike the partial invariance approach, the AwC approach does not require identifying anchoring item(s). As demonstrated in Marsh et al. (2018) and Shi et al. (2017), and also in my simulation, using noninvariant items as anchors can lead to severe bias in structural parameters; by not depending on any anchoring items, AwC thus eliminates one potential source of error.

Researchers should use AwC with caution, however. As the current study show, when the proportion of substantially noninvariant parameters (with $d_{\text{MACS}} \geq .20$) is large (e.g., > 30%; see also B. Muthén & Asparouhov, 2014) or when the proportion of noninvariant items is large (e.g., > 50%), AwC still leads to biased parameter estimates, even though the bias may be smaller than using a noninvariant anchor item with a partial invariance model. The observed bias in AwC was consistent with Asparouhov and Muthén (2014)'s suggestion that the alignment method may fail when the "assumption of approximate measurement invariance is violated" (p. 506), meaning a substantial proportion of parameters with medium-to-large noninvariance. Therefore, when using AwC, I recommend researchers to report the range of d_{MACS} values, the proportion of $d_{\text{MACS}} > .20$, and the proportion of items with at least one $d_{\text{MACS}} > .20$, and be skeptical of parameter estimates when more than 30% of d_{MACS} are > .20 or when more than 50% of the items have one or more $d_{\text{MACS}} > .20$. Furthermore, a large proportion of noninvariance may suggest that an instrument does not measure constructs that are comparable over time.[8] Instead of merely applying AwC or partial invariance for statistical adjustment, researchers should carefully consider the developmental nature of the target constructs and the content of the items to decide whether the instrument can still be meaningfully compared over the time span of the research; it is possible that the instrument does not allow for meaningful comparisons over certain period of time, and refinement of the instrument or development of a new instrument will be needed.

[8] Both the Associate Editor and an anonymous reviewer brought up this excellent point.

Limitations and Future Directions

The current simulation study is not without limitations. First, I only evaluated the linear growth models as my goal was mainly to introduce how AwC can work for longitudinal data and provide the first piece of evidence of its performance; future research can thus examine alternative growth models, such as polynomial growth, piecewise growth, and latent change score models (McArdle & Grimm, 2010; McArdle & Hamagami, 2001). Second, it is possible to apply AwC to designs with more time points and potentially with intensive longitudinal data with many time points (Bolger & Laurenceau, 2013; Hamaker & Wichers, 2017), in which case the advantage of AwC may be even bigger as identifying an appropriate partial invariance model is hard with many time points. However, the results by Asparouhov and Muthén (2014) and Marsh et al. (2018) on independent groups suggested that AO/AwC may produce biased latent parameter estimates when the ratio between group sample size and the number of groups is less than 6 (e.g., 90 individuals per group with 15 groups), so future studies are needed to examine the sample size requirement for using AwC with a larger number of time points. Third, as AO can also be applied to ordered categorical data (B. Muthén & Asparouhov, 2014), future research can explore whether my findings on AwC hold for such data.

In addition, my simulation only focused on violations of factorial invariance with respect to time, but in real research, noninvariance can happen with respect to a combination of time and demographic variables (e.g., gender, age; Horn & McArdle, 1992; E. S. Kim & Willson, 2014), which has been an important but understudied area of research. The AwC approach is potentially useful by considering simultaneous invariance across combinations of time points and demographic subgroups, and future research is needed to formalize how AwC can work in such designs and evaluate its performance and efficiency. Finally, the current study assumes that the sample size is constant across time points, meaning that data are complete or listwise deletion has been used; when there is missing at random attrition that can be handled by full-information maximum likelihood, one can include weights in the component loss function for alignment to reflect different sample sizes across time (see Asparouhov & Muthén, 2014), but future research is needed to evaluate the use of such weights in the AwC growth modeling method.

Given that the AwC method is relatively new, there are also a lot of research opportunities to further optimize it. For example, the component loss function proposed by Asparouhov and Muthén (2014) was chosen mostly because of its empirical performance, and alternative functions or family of functions may perform better in some models and may have better theoretical justifications (see Robitzsch, 2020a). Another direction that can greatly benefit the research community is to automate the steps for fitting second-order growth models with AwC so that users can just specify one second-order growth model; programs can then automatically provide fit indices of both the configural model and the final growth model and the growth parameter estimates after adjustment with AwC, as well as effect size indices indicating the degree of noninvariance.

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Barnes, G. M., Reifman, A. S., Farrell, M. P., & Dintcheff, B. A. (2000). The effects of parenting on the development of adolescent alcohol misuse: A six-wave latent growth model. *Journal of Marriage and Family*, *62*(1), 175–186. https://doi.org/10.1111/j.1741-3737.2000.00175.x

Blankson, A. N., & McArdle, J. J. (2013). Measurement invariance of cognitive abilities across ethnicity, gender, and time among older americans. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *70*(3), 386–397. https://doi.org/10.1093/geronb/gbt106

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

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

Brim, O. G., Baltes, P. B., Bumpass, L. L., Cleary, P. D., Featherman, D. L., Hazzard, W. R., Kessler, R. C., Lachman, M. E., Markus, H. R., Marmot, M. G., Rossi, A. S., Ryff, C. D., & Shweder, R. A. (2020). Midlife in the United States (MIDUS 1), 1995-1996. https://doi.org/10.3886/ICPSR02760.v19

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

Chalmers, R. P. (2020). *SimDesign: Structure for organizing monte carlo simulation designs* [R package version 2.0.1]. https://CRAN.R-project.org/package=SimDesign

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chambers, C. (2019). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice* (2 edition). Princeton University Press.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Fasiolo, M. (2016). *An introduction to mvnfast.* [R package version 0.1.6]. https://CRAN.R-project.org/package=mvnfast

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology*, *4*(1), 22–36. https://doi.org/10.1027/1614-2241.4.1.22

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches.* The Guilford Press.

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Hancock, G., Kuo, W.-L., & Lawrence, F. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 470–489. https://doi.org/10.1207/S15328007SEM0803_7

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3-4), 117–144. https://doi.org/10.1080/03610739208253916

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invarient: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist*, *1*(4), 179–188.

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*(5), 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225

Kenny, D. A. (1979). *Correlation and causality*. Wiley.

Kim, E. S., & Willson, V. L. (2014). Measurement invariance across groups in latent growth modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 408–424. https://doi.org/10.1080/10705511.2014.915374

Kim, M., Kwok, O.-M., Yoon, M., Willson, V., & Lai, M. H. C. (2016). Specification search for identifying the correct mean trajectory in polynomial latent growth models. *The Journal of Experimental Education*, *84*(2), 307–329. https://doi.org/10.1080/00220973.2014.984831

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

Kwok, O.-m., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, *42*, 557–592. https://doi.org/10.1080/00273170701540537

Lai, M. H. C., Liu, Y., & Tse, W. W.-Y. (2021). Adjusting for partial invariance in latent parameter estimation: Comparing forward specification search and approximate invariance methods. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-021-01560-2

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, *101*(6), 1174–1188. https://doi.org/10.1037/a0024776

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

Levy, B. R., Zonderman, A. B., Slade, M. D., & Ferrucci, L. (2012). Memory shaped by age stereotypes over time. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *67*(4), 432–436. https://doi.org/10.1093/geronb/gbr120

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? *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 762–777. https://doi.org/10.1080/10705511.2017.1419353

Lommen, M. J. J., van de Schoot, R., & Engelhard, I. M. (2014). The experience of traumatic events disrupts the measurement invariance of a posttraumatic stress scale. *Frontiers in Psychology*, *5*, 1–7. https://doi.org/10.3389/fpsyg.2014.01304

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., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109*(3), 502–511. https://doi.org/10.1037/0033-2909.109.3.502

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*, *23*(3), 524–545. https://doi.org/10.1037/met0000113

McArdle, J. J., & Grimm, K. J. (2010). Five steps in latent curve and latent change score modeling with longitudinal data. In K. van Montfort, J. H. Oud, & A. Satorra (Eds.), *Longitudinal Research with Latent Variables* (pp. 245–273). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11760-2_8

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change.* (pp. 139–175). American Psychological Association. https://doi.org/10.1037/10409-005

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203–240). American Psychological Association. https://doi.org/10.1037/10409-007

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*, 1–7. https://doi.org/10.3389/fpsyg.2014.00978

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. Routledge.

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2018). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, *22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. https://doi.org/10.1037/a0022955

Obradović, J., Pardini, D. A., Long, J. D., & Loeber, R. (2007). Measuring interpersonal callousness in boys from childhood to adolescence: An examination of longitudinal invariance and temporal stability. *Journal of Clinical Child & Adolescent Psychology*, *36*(3), 276–292. https://doi.org/10.1080/15374410701441633

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development*, *76*(4), 763–782. https://doi.org/10.1111/1467-8624.00498-i1

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/

Robitzsch, A. (2020a). L_p Loss functions in invariance alignment and Haberman linking with few or many groups. *Stats*, *3*(3), 246–283. https://doi.org/10.3390/stats3030019

Robitzsch, A. (2020b). *Sirt: Supplementary item response theory models* [R package version 3.9-4]. https://CRAN.R-project.org/package=sirt

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Ryff, C., Almeida, D., Ayanian, J., Binkley, N., Carr, D. S., Coe, C., Davidson, R., Grzywacz, J., Karlamangla, A., Krueger, R., Lachman, M., Love, G., Mailick, M., Mroczek, D., Radler, B., Seeman, T., Sloan, R., Thomas, D., Weinstein, M., & Williams, D. (2019). Midlife in the United States (MIDUS 3), 2013-2014. https://doi.org/10.3886/ICPSR36346.v7

Ryff, C., Almeida, D. M., Ayanian, J., Carr, D. S., Cleary, P. D., Coe, C., Davidson, R., Krueger, R. F., Lachman, M. E., Marks, N. F., Mroczek, D. K., Seeman, T., Seltzer, M. M., Singer, B. H., Sloan, R. P., Tun, P. A., Weinstein, M., & Williams, D. (2017). Midlife in the United States (MIDUS 2), 2004-2006. https://doi.org/10.3886/ICPSR04652.v7

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shi, D., Song, H., & Lewis, M. D. (2017). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, *26*(7), 1217–1233. https://doi.org/10.1177/1073191117711020

Tyrell, F. A., Yates, T. M., Widaman, K. F., Reynolds, C. A., & Fabricius, W. V. (2019). Data harmonization: Establishing measurement invariance across different assessments of the same construct across adolescence. *Journal of Clinical Child & Adolescent Psychology*, *48*(4), 555–567. https://doi.org/10.1080/15374416.2019.1622124

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association.

Wu, C.-H., Chen, L. H., & Tsai, Y.-M. (2009). Longitudinal invariance analysis of the satisfaction with life scale. *Personality and Individual Differences*, *46*(4), 396–401. https://doi.org/10.1016/j.paid.2008.11.002

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

Table 1

Factor loadings and measurement intercepts for the data generating model across noninvariance conditions.

	r_{ni} = 0	r_{ni} = .25/p_{ni} = .40				r_{ni} = .55/p_{ni} = 1			
Parameter	All Ts	T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4
λ_1	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
λ_2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
λ_3	0.70	0.70	0.70	0.70	0.70	**0.70**	**0.90**	**1.00**	**0.80**
λ_4	0.65	0.65	0.65	0.65	0.65	0.65	**0.60**	0.65	**0.70**
λ_5	0.70	**0.70**	**0.80**	**0.90**	**1.00**	**0.70**	**0.80**	**0.90**	**1.00**
ν_1	0.000	0.000	0.000	0.000	0.000	**0.000**	**0.750**	**0.500**	**0.250**
ν_2	0.500	0.500	0.500	0.500	0.500	0.500	**0.750**	0.500	**1.000**
ν_3	-0.250	-0.250	-0.250	-0.250	-0.250	-0.250	-0.250	-0.250	-0.250
ν_4	0.250	**0.250**	**0.500**	**0.750**	**1.000**	**0.250**	**0.500**	**0.750**	**1.000**
ν_5	-0.500	-0.500	**-0.375**	**-0.625**	-0.500	-0.500	**-0.375**	**-0.625**	-0.500

Note. r_{ni} = proportion of noninvariant parameters. p_{ni} = proportion of noninvariant items. λ = factor loadings. ν = measurement intercepts. Noninvariant parameters are bolded.

Table 2

Root mean squared error (RMSE) and error rates of 95% confidence intervals (CIs) for mean level (κ_1) and level variance (ϕ_1).

			Mean Level (κ_1)						Level Variance (ϕ_1)					
			RMSE			CI Error Rate			RMSE			CI Error Rate		
Model	N	r_{ni}/p_{ni}	PI	FI	AwC	PI	FI	AwC	PI	FI	AwC	PI	FI	AwC
C	100	0	0.10	0.10	0.11	4.8	4.8	4.1	0.16	0.16	0.16	6.0	6.0	3.9
		.25/.40	0.10	0.12	0.11	4.7	7.2	3.9	0.15	0.17	0.16	6.1	**13.8**	4.9
		.55/1	0.24	0.19	0.12	**45.6**	**29.4**	5.4	0.17	0.18	0.18	**9.3**	**15.4**	6.4
	250	0	0.07	0.07	0.07	4.9	4.9	3.7	0.10	0.10	0.10	5.5	5.5	3.5
		.25/.40	0.06	0.08	0.07	3.6	**11.0**	3.1	0.09	0.11	0.10	4.6	**14.5**	3.3
		.55/1	0.22	0.17	0.08	**84.6**	**62.2**	**9.4**	0.11	0.12	0.12	**8.7**	**15.7**	6.8
	1000	0	0.03	0.03	0.03	3.8	3.8	3.0	0.05	0.05	0.05	5.2	5.2	3.5
		.25/.40	0.03	0.06	0.03	4.0	**33.0**	3.3	0.05	0.07	0.05	3.8	**27.4**	2.6
		.55/1	0.22	0.16	0.06	**100.0**	**99.5**	**31.2**	0.07	0.08	0.07	**17.8**	**30.1**	**12.0**
M	100	0	0.10	0.10	0.11	5.3	5.3	3.2	0.16	0.16	0.17	**10.5**	**10.5**	5.8
		.25/.40	0.10	0.11	0.11	5.2	6.1	2.8	0.16	0.19	0.18	**9.9**	**21.7**	7.1
		.55/1	0.23	0.18	0.11	**38.0**	**27.6**	3.1	0.20	0.21	0.20	**15.5**	**25.2**	**8.5**
	250	0	0.07	0.07	0.07	5.2	5.2	2.6	0.11	0.11	0.11	**10.8**	**10.8**	6.4
		.25/.40	0.07	0.07	0.07	4.3	6.9	*2.3*	0.10	0.15	0.12	**9.6**	**29.6**	6.7
		.55/1	0.22	0.16	0.07	**79.1**	**57.1**	4.4	0.15	0.17	0.14	**20.9**	**36.0**	**10.2**
	1000	0	0.03	0.03	0.04	6.5	6.5	3.8	0.07	0.07	0.07	**17.5**	**17.5**	**11.2**
		.25/.40	0.04	0.04	0.04	6.0	**15.0**	3.6	0.07	0.12	0.07	**17.5**	**69.4**	**13.7**
		.55/1	0.21	0.15	0.05	**100.0**	**98.7**	**12.8**	0.12	0.14	0.09	**55.1**	**80.7**	**25.1**

Note. r_{ni} = proportion of noninvariant parameters. p_{ni} = proportion of noninvariant items. PI = partial strong invariance model. FI = full strong invariance model. AwC = alignment-within-confirmatory factor analysis. C = correctly specified model. M = misspecified model. RMSEs are averaged across conditions of average growth rate. Bolded values indicate error rates > 7.5%; For conditions with $r_{ni} = .55/p_{ni} = 1$, the PI model was misspecified as there were no noninvariant items.

Table 3

Root mean squared error (RMSE) and error rates of 95% confidence intervals (CIs) for mean slope (κ_2) and slope variance (ϕ_2).

				Mean Slope (κ_2)						Slope Variance (ϕ_2)					
				RMSE			CI Error Rate			RMSE			CI Error Rate		
Model	κ_2	N	r_{ni}/p_{ni}	PI	FI	AwC	PI	FI	AwC	PI	FI	AwC	PI	FI	AwC
C	0.00	100	0	0.05	0.05	0.05	5.4	5.4	3.6	0.04	0.04	0.04	5.9	5.9	5.5
			.25/.40	0.05	0.07	0.06	5.7	**12.2**	4.6	0.04	0.05	0.05	6.4	**7.7**	5.7
			.55/1	0.09	0.10	0.08	**25.3**	**29.3**	**9.8**	0.04	0.05	0.07	**12.8**	**10.2**	**7.7**
		250	0	0.03	0.03	0.03	4.7	4.7	*2.5*	0.02	0.02	0.03	5.9	5.9	5.4
			.25/.40	0.03	0.05	0.03	5.4	**19.6**	3.6	0.02	0.04	0.03	5.5	**13.0**	6.9
			.55/1	0.08	0.08	0.05	**48.1**	**57.0**	**9.5**	0.03	0.04	0.05	**14.0**	**19.1**	**15.3**
		1000	0	0.02	0.02	0.02	4.8	4.8	*2.4*	0.01	0.01	0.01	5.7	5.7	5.3
			.25/.40	0.02	0.04	0.02	5.3	**52.5**	3.1	0.01	0.03	0.01	5.2	**41.8**	6.4
			.55/1	0.07	0.08	0.03	**96.0**	**98.2**	**15.0**	0.02	0.03	0.03	**28.3**	**57.3**	**31.6**
	0.25	100	0	0.05	0.05	0.05	5.3	5.3	4.1	0.04	0.04	0.04	5.8	5.8	5.5
			.25/.40	0.05	0.09	0.07	6.0	**24.5**	7.3	0.04	0.05	0.05	6.4	**8.9**	5.7
			.55/1	0.09	0.13	0.10	**24.8**	**58.1**	**23.1**	0.04	0.06	0.07	**12.8**	**12.0**	**7.8**
		250	0	0.03	0.03	0.03	4.3	4.3	3.1	0.02	0.02	0.03	6.0	6.0	5.4
			.25/.40	0.03	0.07	0.04	5.3	**48.4**	6.2	0.02	0.04	0.03	5.5	**15.0**	6.9
			.55/1	0.08	0.13	0.07	**46.3**	**91.7**	**27.4**	0.03	0.04	0.05	**14.0**	**23.4**	**15.5**
		1000	0	0.02	0.02	0.02	4.5	4.5	2.8	0.01	0.01	0.01	5.7	5.7	5.2
			.25/.40	0.02	0.07	0.02	4.8	**96.1**	5.6	0.01	0.03	0.01	5.3	**49.0**	6.4
			.55/1	0.07	0.12	0.04	**93.8**	**100.0**	**43.4**	0.02	0.04	0.03	**28.3**	**68.3**	**32.0**
M	0.00	100	0	0.05	0.05	0.05	5.2	5.2	*1.8*	0.04	0.04	0.04	**11.0**	**11.0**	**8.0**
			.25/.40	0.05	0.07	0.06	5.2	**14.0**	3.4	0.04	0.04	0.05	**11.6**	7.3	5.9
			.55/1	0.11	0.10	0.07	**31.4**	**31.2**	7.1	0.05	0.05	0.06	**20.2**	**8.2**	5.5
		250	0	0.03	0.03	0.03	4.9	4.9	*1.4*	0.03	0.03	0.03	**12.4**	**12.4**	**8.9**
			.25/.40	0.03	0.05	0.03	5.3	**23.8**	*2.2*	0.03	0.03	0.03	**13.1**	**7.8**	7.2
			.55/1	0.10	0.08	0.05	**59.0**	**59.4**	**7.6**	0.04	0.03	0.04	**32.0**	**10.1**	7.0
		1000	0	0.02	0.02	0.02	4.6	4.6	*0.9*	0.02	0.02	0.02	**24.8**	**24.8**	**18.0**
			.25/.40	0.02	0.04	0.02	4.9	**63.9**	*1.4*	0.02	0.02	0.02	**25.5**	**10.6**	**10.5**
			.55/1	0.09	0.08	0.03	**98.8**	**98.4**	**12.8**	0.04	0.02	0.02	**72.8**	**21.5**	7.3
	0.25	100	0	0.05	0.05	0.06	5.0	5.0	2.6	0.04	0.04	0.04	**11.2**	**11.2**	**8.0**
			.25/.40	0.05	0.09	0.07	5.0	**28.3**	6.0	0.04	0.04	0.05	**11.5**	7.4	5.8
			.55/1	0.12	0.14	0.10	**33.8**	**60.5**	**18.7**	0.05	0.05	0.06	**20.2**	**9.0**	5.5
		250	0	0.03	0.03	0.03	4.6	4.6	*1.6*	0.03	0.03	0.03	**12.4**	**12.4**	**8.8**
			.25/.40	0.03	0.08	0.04	5.0	**55.8**	5.6	0.03	0.03	0.03	**13.2**	**8.5**	7.2
			.55/1	0.10	0.13	0.07	**62.9**	**92.9**	**23.2**	0.04	0.04	0.04	**32.0**	**12.4**	7.1
		1000	0	0.02	0.02	0.02	4.4	4.4	*1.8*	0.02	0.02	0.02	**24.8**	**24.8**	**18.0**
			.25/.40	0.02	0.07	0.02	4.8	**98.1**	4.8	0.02	0.02	0.02	**25.5**	**13.6**	**10.6**
			.55/1	0.09	0.12	0.05	**99.2**	**100.0**	**42.3**	0.04	0.02	0.02	**72.8**	**32.0**	7.4

Note. r_{ni} = proportion of noninvariant parameters. p_{ni} = proportion of noninvariant items. PI = partial strong invariance model. FI = full strong invariance model. AwC = alignment-within-confirmatory factor analysis. C = correctly specified model. M = misspecified model. Bolded values indicate error rates > 7.5%; italic values indicate error rates < 2.5%. For conditions with $r_{ni} = .55/p_{ni} = 1$, the PI model was misspecified as there were no noninvariant items.

Table 4

Factor loadings, measurement intercepts, and latent means and variances of the longitudinal configural model of the applied example before and after alignment optimization.

	Loading/Variance		Intercept/Mean	
Variable	Pre-aligned	Aligned	Pre-aligned	Aligned
Measurement Parameters				
moody1	0.44	0.44	2.40	2.40
moody2	0.41	0.45	2.18	2.32
moody3	0.43	0.46	2.09	2.21
worry1	0.79	0.79	2.62	2.62
worry2	0.73	0.80	2.38	2.63
worry3	0.72	0.77	2.41	2.62
nervous1	0.77	0.77	2.24	2.24
nervous2	0.68	0.74	1.98	2.21
nervous3	0.71	0.75	2.05	2.25
calm1	0.32	0.32	2.11	2.11
calm2	0.33	0.36	2.17	2.28
calm3	0.34	0.36	2.14	2.24
Structural Parameters				
η_1	1.00	1.00	0.00	0.00
η_2	1.00	0.92	0.00	-0.31
η_3	1.00	0.94	0.00	-0.27

Note. The items moody, worry, and nervous were reversely coded.

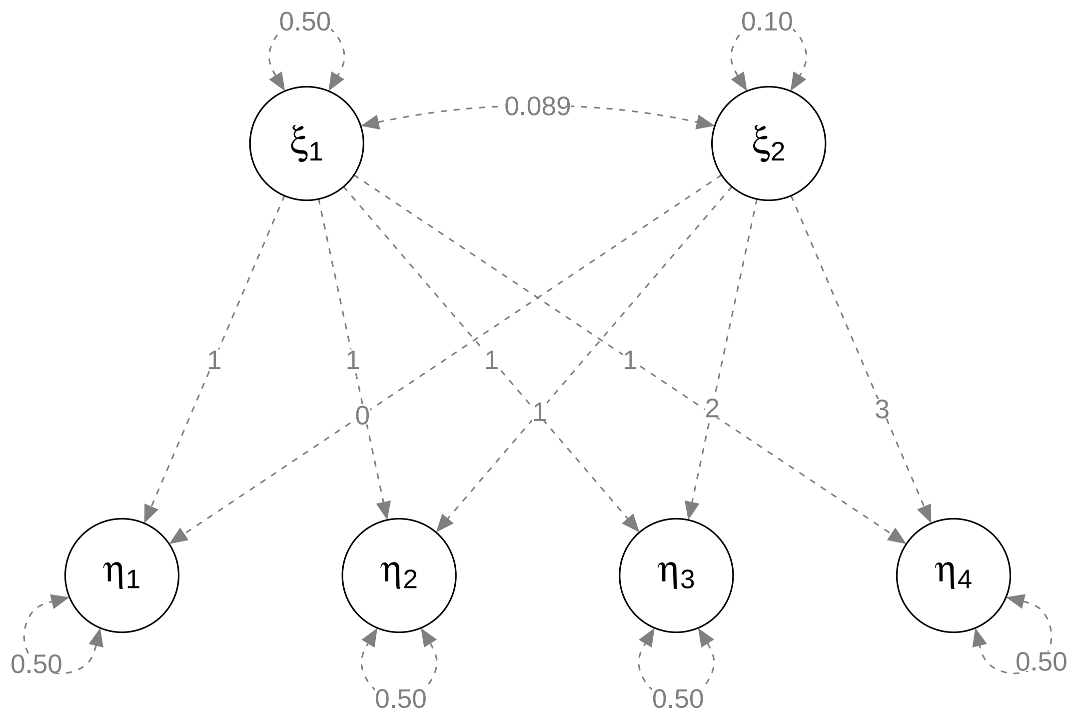

Figure 1. Data generating model for the simulation study. Each of the η variables was measured by five indicators, which were omitted from the diagram.

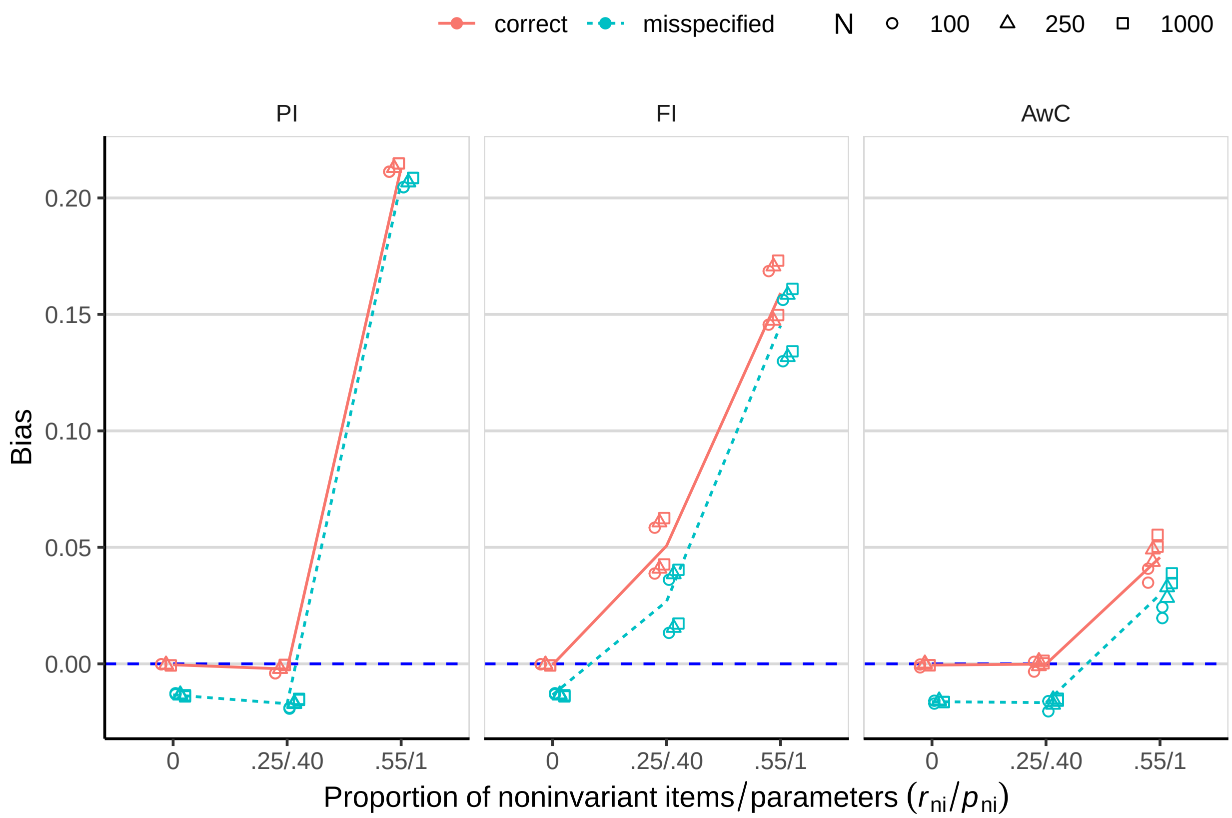

Figure 2. Bias for estimating mean level (κ_1). PI = partial strong invariance model. FI = full strong invariance model. AwC = alignment-within-confirmatory factor analysis. For conditions with $r_{\mathrm{ni}} = .55/p_{\mathrm{ni}} = 1$, the PI model was misspecified as there were no noninvariant items.

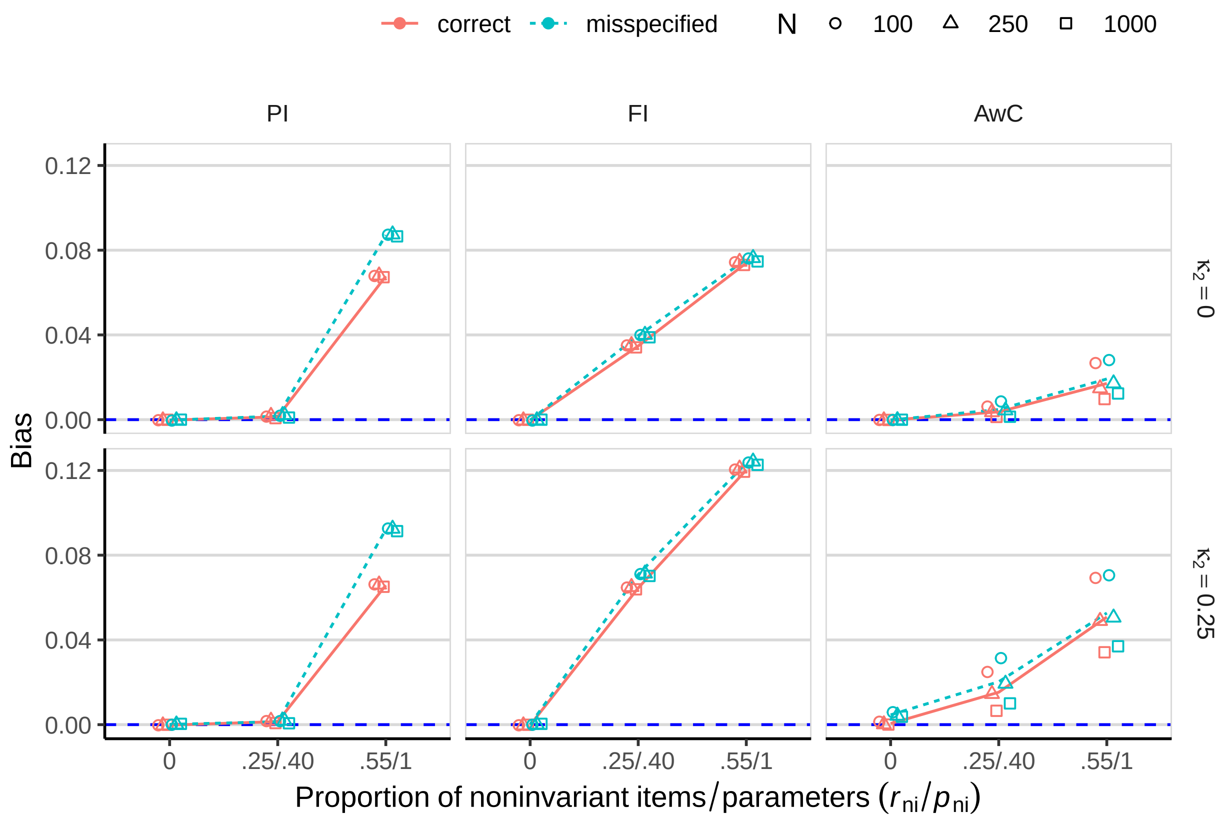

Figure 3. Bias for estimating mean slope (κ_2). PI = partial strong invariance model. FI = full strong invariance model. AwC = alignment-within-confirmatory factor analysis. For conditions with $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, the PI model was misspecified as there were no noninvariant items.

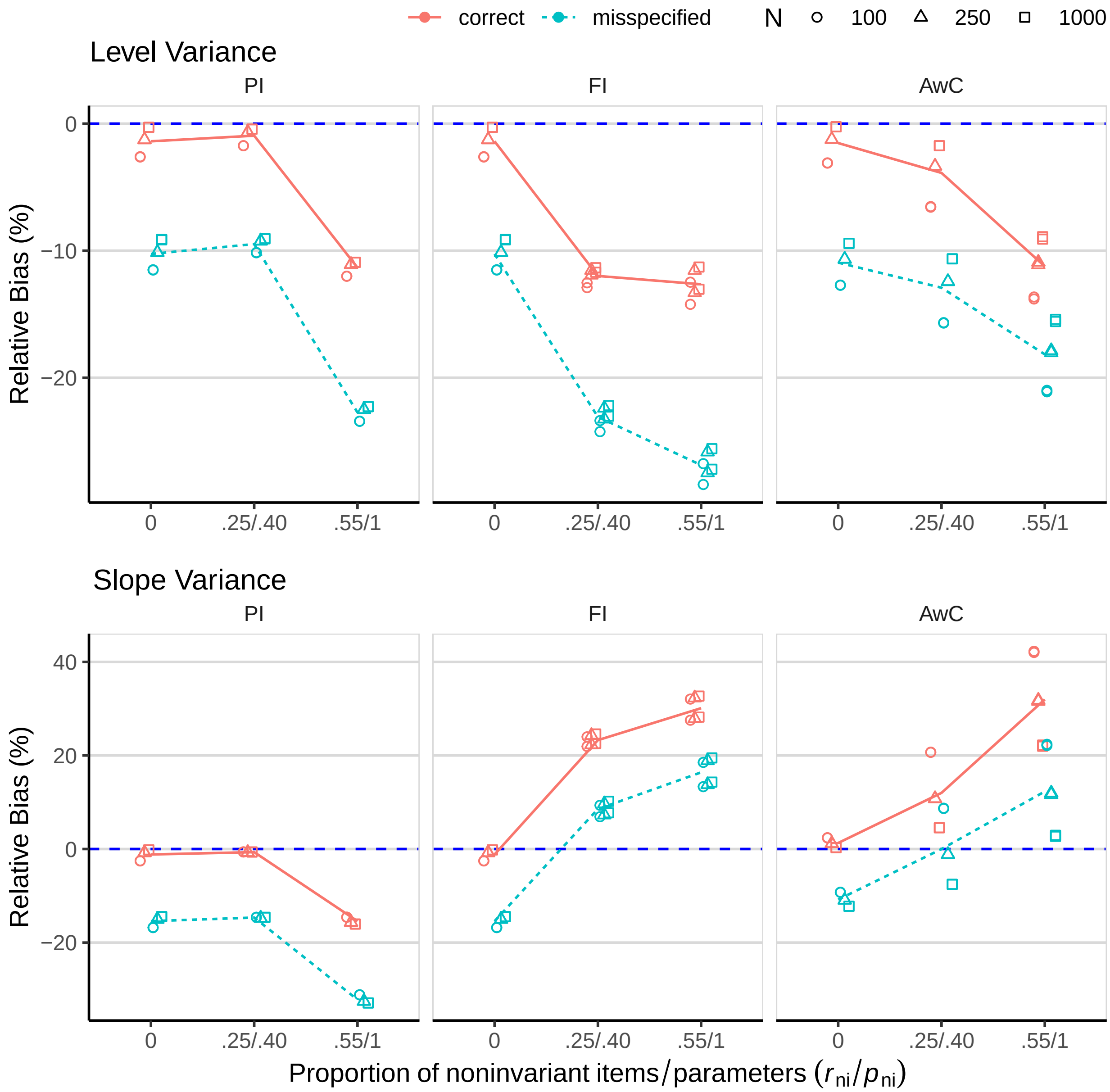

Figure 4. Percentage relative bias for estimating level and slope variance (ϕ_1 and ϕ_2). PI = partial strong invariance model. FI = full strong invariance model. AwC = alignment-within-confirmatory factor analysis. For conditions with $r_{\text{ni}} = .55/p_{\text{ni}} = 1$, the PI model was misspecified as there were no noninvariant items.

Appendix

A Heuristic Example of Alignment Optimization (AO)

As an example of applying the AO loss function defined in equations (2) and (3), consider a scenario where three items are used to measure a latent variable across two waves ($t_1 = 1$ and $t_2 = 2$). Assume that for the first wave, one already knows $\alpha_1 = 0$, $\psi_1 = 1$, $\boldsymbol{\lambda}_1 = [0.9, 0.8, 0.7]$, and $\boldsymbol{\nu}_1 = [0, 0, 0]$. Because of factor indeterminacy, for the second wave, there are infinitely many possible sets of parameter estimates that correspond to the same model-implied means and covariances for the observed variables. For example, consider the following two sets of parameters for the second wave:

- Model 0 (M_0): $\alpha_{2,0} = 0$, $\psi_{2,0} = 1$, $\boldsymbol{\lambda}_{2,0} = [0.81, 0.72, .45]$, $\boldsymbol{\nu}_{2,0} = [0.45, 0.4, 0.4]$
- Model 1 (M_1): $\alpha_{2,1} = 0.5$, $\psi_{2,1} = 0.81$, $\boldsymbol{\lambda}_{2,1} = [0.9, 0.8, 0.5]$, $\boldsymbol{\nu}_{2,1} = [0, 0, 0.15]$

Under both M_0 and M_1, the latent variable accounts for variances of 0.6561, 0.5184, and 0.2025 for the three items (using $\lambda^2\psi$), and the mean of the three items are 0.45, 0.4, and 0.4 (using $\nu + \lambda\alpha$), so they are equivalent models, and there are infinitely many more combinations of α_2, ψ_2, $\boldsymbol{\lambda}_2$, and $\boldsymbol{\nu}_2$ that are equivalent. However, M_0 and M_1 give different implications with respect to factorial invariance, as M_0 implies all items are noninvariant, whereas M_1 implies only item 3 is noninvariant. Because AO aims to identify a set of parameters, among all the equivalent models, that has very few large noninvariant parameters and many approximately invariant parameters, it should prefer M_1 over M_0.

Let's go through equations (2) and (3) to get the component loss (F) values for the parameter differences of M_0 and M_1, with $\epsilon = .001$. For the loading of the first indicator in M_0,

$$f(\lambda_{11,0} - \lambda_{12,0}) = f(0.9 - 0.81) = f(0.09) = \sqrt{\sqrt{(0.09)^2 + .001}} = 0.31,$$

and under M_1,

$$f(\lambda_{11,1} - \lambda_{12,1}) = f(0.9 - 0.9) = f(0) = \sqrt{\sqrt{0^2 + .001}} = 0.18.$$

One can verify that the loss values for the loadings and intercepts under M_0 are 0.31, 0.29, 0.50, 0.67, 0.63, 0.63, and those under M_1 are 0.18, 0.18, 0.45, 0.18, 0.18, 0.39. Summing the loss values as in equation (2), one gets $F_0 = 3.04$ and $F_1 = 1.55$, so M_1 is indeed preferred in AO over M_0.

This heuristic example only considers two sets of parameter values, but the AO algorithm considers all possible sets and identifies the one, denoted as M_a, that gives the smallest F.