Behav Genet (2017) 47:516–536
DOI 10.1007/s10519-017-9864-0

ORIGINAL RESEARCH

Sum Scores in Twin Growth Curve Models: Practicality Versus Bias

Justin M. Luningham[1] · Daniel B. McArtor[1] · Meike Bartels[2] · Dorret I. Boomsma[2] · Gitta H. Lubke[1]

Received: 18 November 2016 / Accepted: 15 July 2017 / Published online: 5 August 2017

Abstract To study behavioral or psychiatric phenotypes, multiple indices of the behavior or disorder are often collected that are thought to best reflect the phenotype. Combining these items into a single score (e.g. a sum score) is a simple and practical approach for modeling such data, but this simplicity can come at a cost in longitudinal studies, where the relevance of individual items often changes as a function of age. Such changes violate the assumptions of longitudinal measurement invariance (MI), and this violation has the potential to obfuscate the interpretation of the results of latent growth models fit to sum scores. The objectives of this study are (1) to investigate the extent to which violations of longitudinal MI lead to bias in parameter estimates of the average growth curve trajectory, and (2) whether absence of MI affects estimates of the heritability of these growth curve parameters. To this end, we analytically derive the bias in the estimated means and variances of the latent growth factors fit to sum scores when the assumption of longitudinal MI is violated. This bias is further quantified via Monte Carlo simulation, and is illustrated in an empirical analysis of aggression in children aged 3–12 years. These analyses show that measurement non-invariance across age can indeed bias growth curve mean and variance estimates, and our quantification of this bias permits researchers to weigh the costs of using a simple sum score in longitudinal studies. Simulation results indicate that the genetic variance decomposition of growth factors is, however, *not* biased due to measurement non-invariance across age, provided the phenotype is measurement invariant across birth-order and zygosity in twins.

Keywords Sum scores · Growth curve models · Twin models · Aggression · Measurement invariance

Introduction

The analysis of longitudinal data from relatives provides two important research opportunities: the opportunity to structure repeated measurements of a phenotype into growth trajectories, and the opportunity to investigate the genetic and environmental contributions the parameters that govern these trajectories (McArdle 1986; Neale and Maes 2004; Neale and McArdle 2000; Prescott and Kendler 1996). In longitudinal settings, however, the meaning of a measurement instrument may change over time, and different questions may be asked to address age-appropriate expressions of the same behavior (Achenbach and Rescorla 2000, 2001; Edwards and Wirth 2009; Hudziak et al. 2003; Rutter and Sroufe 2000). When constructing a simple phenotype to analyze across multiple time points or ages, a common practice is to use a sum score or mean score that aggregates individual questionnaire items (e.g., Forsman et al. 2010; Hudziak et al. 2003; Wang et al. 2013; van Beijsterveldt et al. 2003). Although these scores are easy to compute and use, they are based on the assumption that all items are equally relevant indicators of the phenotype of interest, and they ignore the possibility that items can change in relevance over age. The aim of this paper is to investigate the consequences of analyzing simple aggregate scores as longitudinal phenotypes. More specifically,

Edited by Deborah Finkel.

✉ Justin M. Luningham
jluningh@nd.edu

[1] Department of Psychology, University of Notre Dame, 220 C Haggar Hall, Notre Dame, IN 46556, USA

[2] Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

we provide a quantification of the bias in structured growth curve estimates resulting from the use of sum scores in linear growth curve models (McArdle 1988; Singer and Willet 2003), and we evaluate the impact this bias has on genetic variance decomposition of the growth trajectories.

Ideally, the trait of interest is measured perfectly across different instruments and ages. Identical measurement across ages is not easily accomplished in practice because longitudinal studies may include age-appropriate versions of a questionnaire at different measurement occasions. This is especially common when measuring psychological outcomes or childhood psychopathology over many ages, such as in the study of aggression across childhood and adolescence (Edwards and Wirth 2012; Hudziak et al. 2003; Kan et al. 2013). Items in a measurement tool that change in meaning or content over time imply violations of measurement invariance (MI). Longitudinal MI means that each item's relationship to the scale's underlying construct is constant across repeated measurements and ensures that the underlying construct has the same meaning across measurement occasions (Meredith 1993; Meredith and Horn 2001; Widaman et al. 2010). If the observed items are measurement invariant over time, then changes across time can be confidently interpreted as longitudinal development of the latent trait. From a statistical perspective, violations of MI make it difficult to disentangle true growth in the latent trait from changing characteristics in the measurement instrument. A natural application (in which this challenge is especially relevant) is assessing the behavior of children over multiple ages as they develop and change.

A common practice in the assessment of childhood psychopathology is to ask parents or teachers to rate a child's behavior across a number of items and to form a sum, mean, or other aggregate score based on these items (Achenbach and Rescorla 2001; Goodman 1997; Grimm et al. 2013). These aggregate scores are easy to compute, and they can be a reliable measure if the number of items is sufficiently large (Carmines and Zeller 1979). For example, more than 20 items are needed to ensure a Cronbach's .....at least 0.8, considered to indicate "good" reliability of a scale, if the items have 70% measurement error (Cronbach 1951; George and Mallery 2005). In behavioral applications, the number of items in a sum score is usually much smaller. As a consequence, the measurement error of the composite items does not cancel out in the formation of the aggregate score, leading to low reliability. In addition, scores on the items are commonly summed with all items receiving the same weight, implying that all items are equally important indicators of the construct. If the true measurement model relating the items to the underlying trait is not the same at each time point, the estimated average growth curve of the phenotype may be biased using this approach (Leite 2007; Wirth 2009).

It is therefore important to test for longitudinal MI before fitting growth models to a phenotype that is measured with multiple items. Violations of MI cannot be detected using a simple aggregate score, but MI can be investigated if a latent variable measurement model is fitted to the item-level data. In particular, longitudinal MI can be assessed with a second-order latent growth model (SLGM), which was developed to study longitudinal change in a latent phenotype that is measured by multiple items at each time point (Hancock et al. 2001; McArdle 1988). The SLGM permits estimation of the measurement properties of the items at each occasion and consequently allows the investigation of potential violations of MI, such as changes in factor loadings over time.

In addition, the measurement portion of the SLGM can also be used to investigate *a priori* hypotheses in which violations of MI might be expected due to theoretical or substantive knowledge about the phenotype. For example, researchers may expect a certain item to be more salient at later ages and therefore fit a model allowing the particular item to increase in its importance (i.e., factor loading) over time. It has been argued that if one has hypotheses as to how and why MI would not hold, a model that relaxes MI constraints accordingly will provide a better fit to the data, and the construct can be considered consistent over time (Byrne et al. 1989). The decision to allow for non-invariance should be guided by theory, but it can also be evaluated by comparing model fit because the model imposing MI is nested within a less restricted model allowing time-specific loadings. In a study of childhood aggression, for instance, an item asking if a child threatens others is likely to be more salient for older children due to an increase in cognitive and/or verbal skills as children age. If item loadings change in an *a priori* formulated way, and model fit is improved, it is possible to argue that the differences in loadings might be due to changes in the behavioral expression of the same underlying construct. In other words, it can be argued that the construct is measured age-appropriately.

Previous research has demonstrated that modeling a latent phenotype with sum scores can introduce bias in twin model estimates when measurement non-invariance is present across grouping variables such as gender, environmental exposure group, or twin zygosity (Lubke et al. 2004; Neale et al. 2005; Slof-Op't Landt et al. 2009). Particularly, using sum scores with measurement non-invariance across gender can lead to the incorrect conclusion of sex limitation in cross-sectional estimates of heritability (Lubke et al. 2004). If and how violations of longitudinal MI impact genetic variance components have yet to be explored. With genetically informative longitudinal data, there are two main options. A popular approach is to fit a genetic simplex model to the data (Boomsma and Molenaar 1987). The simplex involves decomposing the observed variance

at each measurement occasion into genetic and environmental components. These components are autoregressive in the simplex model, meaning that the source of variation at a given time point is causally related to the previous time point, reflecting the degree to which genetic and environmental influences are transmitted from one occasion to the next. Unique variance at each occasion is also decomposed into meaningful genetic and environmental components, corresponding to genetic or environmental innovations in phenotypic variance. This framework informs about the degree to which the specific variance components are stable, via a high degree of transmission, or changing, via a high degree of innovation.

An alternative approach, which is the emphasis of this paper, is to specify a particular growth structure in the phenotype, such as a latent growth model (LGM), which permits decomposing the variance of the latent variables that account for the growth structure into genetic and environmental components (Neale and McArdle 2000). Specifically, LGMs summarize the covariance among repeated measures into latent variables that correspond to an intercept factor and one or more change factors, most commonly involving a slope factor that parameterizes linear change in the phenotype across time. The variance decomposition is carried out at the level of these growth factors. The simplex model provides an indication of the stability of heritability from time point to time point, whereas the LGM decomposes the variation of linear or curvilinear phenotypic trajectories into genetic and environmental contributions. The intercept factor, which summarizes the degree of stability over the repeated measurements, has corresponding genetic and environmental components, and the slope factor, summarizing a specified change trajectory, has its own variance components (e.g., Finkel et al. 2015; Lubke et al. 2016; see McArdle and Hamagami 2003, for related models). However, any bias from non-invariance is expected to be realized in the estimates of these structured growth factors. Here, we extend previous research to investigate whether or not the variance decomposition of growth factors is affected by longitudinal violations of MI.

The first aim of this paper is to provide analytic and simulation-based quantifications of bias due to using sum scores in growth models when measurement properties change over time. The second aim of the paper is to investigate, through simulation, how bias in the variance parameters of the growth model may affect genetic and environmental variance decompositions in longitudinal twin analyses. The results can be used to assess if the benefits of using a convenient and simple phenotype definition (i.e., a sum score) in a given analysis outweigh the costs. The paper is organized as follows. The LGM for a single observed variable and its extension to accommodate multiple items are briefly presented. The model notation is then used to derive the bias in the estimated growth factors if MI over time is violated and sum scores are used as scores on the phenotype. We present the bias derivations, give three hypothetical demonstrations of deviations from MI over time that could occur in real data (e.g., individual items change in relevance across age), and confirm our analytic results with Monte Carlo simulations. The derivations of the bias are then illustrated in an exploratory subsample of data concerning the development of aggression during childhood in Dutch twins. Finally, based on the parameters from the empirical illustration, twin data are simulated to investigate potential bias in the heritability estimates of the growth factors.

Latent growth model

The linear LGM is a widely applied model for longitudinal data. It can be used to model univariate observations (e.g., answers to a single question) measured at 3 or more time points (McArdle 1988; Singer and Willet 2003). The linear LGM features two latent variables that represent the intercept and slope of the individuals' linear change trajectories over time, respectively. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ denote subject i's scores ($i = 1, \ldots, n$) on a single variable of interest measured at occasions $t = 1, \ldots, T$, where t can, for instance, indicate the different ages at which children's behaviors have been rated. The linear growth model for the trajectory of individual i measured at T time points can be represented as,

$$\mathbf{y}_i = \mathbf{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i \tag{1}$$

where $\mathbf{\Gamma}$ is a $T \times 2$ matrix of loadings that are fixed to correspond to linear growth, $\boldsymbol{\xi}_i = \begin{bmatrix} \alpha_i & \beta_i \end{bmatrix}$ is a 2×1 vector of factor scores for individual i on intercept (α) and slope (β) growth factors, and $\boldsymbol{\zeta}_i$ is a $T \times 1$ vector of residuals for individual i's observations at the T time points (Hancock et al. 2001). The loadings onto the intercept factor are fixed to one, and the loadings onto the slope factor are fixed to values corresponding to the time intervals between measurement occasions (e.g., the number of years between time points). The average growth curve is described by the means of the intercept and slope factors.

When the outcome at each time point is measured by multiple items, researchers can either aggregate items at each occasion into a single score and then use the LGM presented above, or they can use the SLGM (Hancock et al. 2001; McArdle 1988; Sayer and Cumsille 2001). In the SLGM, a single factor is modeled from the multiple items at each time point, and this factor is simultaneously fitted to a linear growth model, replacing the univariate outcome presented in Eq. (1).

The factor model for the p items[1] ($j=1,\ldots, p$), measured on individual i at time point t is given by,

$$y_{ijt} = \nu_{jt} + \lambda_{jt}\eta_{it} + \varepsilon_{ijt} \quad (2)$$

where ν_{jt} represents the intercept for item j at time t, λ_{jt} is the item loading relating item j to the single latent construct η_{it} at time t, and ε_{ijt} is the item residual for individual i on item $j$at time t. Note that this model formulation corresponds to a single phenotype or construct measured by multiple items. However, the factor model can be extended to incorporate a multi-dimensional phenotype or even multiple phenotypes by including more η's. Using vector notation, $\mathbf{y}_i = \left(y_{i11}, \ldots, y_{ip1}, \ldots, y_{i1T}, \ldots, y_{ipT}\right)'$ is now a $(T * p) \times 1$ vector whose elements are the p items measured at each time t. Equation (2) can therefore be rewritten as,

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad (3)$$

where $\boldsymbol{\eta}_i$ is a $T \times 1$ vector of latent factor scores for subject i at each measurement occasion, $\boldsymbol{\nu}$ is a $(T * p) \times 1$ vector of intercepts for the items comprising $\mathbf{y}_i$, and $\boldsymbol{\varepsilon}_i$ is a $(T * p) \times 1$ vector of residuals for the i^{th} individual on the items comprising $\mathbf{y}_i$. The $(T * p) \times T$ matrix $\boldsymbol{\Lambda}$ contains the p item loadings at each time t. Equation 3 represents the common factor model for p items measured at T time points.

In the SLGM, the factors $\boldsymbol{\eta}_i$ are then subjected to the linear growth model:

$$\boldsymbol{\eta}_i = \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i \quad (4)$$

where the notation is equal to Eq. (1). Combining Eqs. (3) and (4) and fixing the item intercepts $\boldsymbol{\nu}$ to zero (without loss of generality)[2] gives,

$$\mathbf{y}_i = \boldsymbol{\Lambda}(\boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i) + \boldsymbol{\varepsilon}_i \quad (5)$$

$$E\left[\mathbf{y}_i\right] = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\mu} \quad (6)$$

$$\operatorname{cov}\left[\mathbf{y}_i\right] = \boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}\boldsymbol{\Omega}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} = \boldsymbol{\Lambda}\left(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}\right)\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \quad (7)$$

where $\boldsymbol{\mu} = \left[\mu_\alpha \ \mu_\beta\right]'$ is a 2×1 vector of the intercept and slope factor means, $\boldsymbol{\Omega} = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}$ is the (co)variance matrix of the measurement constructs $\boldsymbol{\eta}_i$, $\boldsymbol{\Phi}$ is the (co)variance matrix of the intercept and slope factors $\boldsymbol{\xi}_i$, $\boldsymbol{\Psi}$ is the residual (co)variance matrix of the measurement constructs, and $\boldsymbol{\Theta}$ is the residual (co)variance matrix of the questionnaire items. Commonly, it is assumed that the items have independent and identically distributed errors, i.e., $\boldsymbol{\Theta} = \sigma_\varepsilon^2\mathbf{I}_T$, and σ_ε^2 is the residual variance of the observations. The intercept and slope factor means ($\boldsymbol{\mu}$) determine the average growth trajectory in a sample. A path diagram of a general SLGM is presented in Fig. 1

Note in Eq. (2) the t subscript for λ_{jt}. corresponding to time-specific item loadings. The time-specific loadings allow the items to relate to the underlying measurement factor differently across time, which would indicate a violation of MI. The SLGM allows tests of MI over time by specifying models in which $\lambda_{j1} = \lambda_{j2} = \ldots = \lambda_{jT}$, i.e., estimating only one λ_j which is equal across t. Constraining the item loadings and item intercepts to be time-invariant for the repeated items constitutes "strong MI" (Meredith 1993; Widaman and Reise 1997). Strong MI is generally considered to be the minimum level that should be established to ensure that the factors have the same interpretation and same scale across different ages (Ferrer et al. 2008; Widaman et al. 2010). If parameters of the measurement model are age-specific, the interpretation of the factors is not the same over a developmental span, which in turn makes the interpretation of the growth curve unclear.

Sum or aggregate score models

Sum scores based on multiple items carry the implicit assumption of MI, because the unweighted sums of the items implies that all of the loadings in $\boldsymbol{\Lambda}$ presented in Eq. (3) are all fixed to 1 at each age. Importantly, summing a small number of items also can leave considerable measurement error in the aggregate score. In what follows, we evaluate the bias in estimates of growth means and variances in LGMs when sum scores are analyzed. Evaluating the bias is necessary to understand the costs of using sum scores in twin models (Neale et al. 2005; van den Berg et al. 2007).

To assess the bias in variance and mean parameters of the growth model when sum scores are analyzed, it is assumed that the SLGM presented in Eq. (5) is the true underlying data-generating process. Using the notation of the SLGM, the y_{ijt}'s are summed over all p items at each time point t. Specifically, for the means given in Eq. (6), the equation can be written to show the result of summing over all items at each time point:

$$\sum_{j=1}^{p} E[y_{ijt}] = \sum_{j=1}^{p} \lambda_{jt}\mu_\alpha + \sum_{j=1}^{p} \lambda_{jt}\gamma_t\mu_\beta = (\mu_\alpha + \gamma_t\mu_\beta)\sum_{j=1}^{p} \lambda_{jt} \quad (8)$$

In matrix notation, this can be rewritten as a diagonal matrix of the summed factor loadings at each time point,

[1] The observed scores in this derivation are assumed to be (multivariate) normally distributed.Categorical outcomes can be dealt with by a superseding threshold model (Agresti 2002), whichis omitted here to avoid unnecessary complexity.

[2] The item intercepts are fixed at zero to set them equal over time and also allow for the estimation of the intercept factor mean. Another option for SLGMs is to estimate item means and constrain them to equality over time while setting the intercept mean to zero.

Table 1 Definitions of SLGM notation

Matrix	Definition	Matrix	Definition
$\boldsymbol{\Gamma}$	Growth loadings	η	Measurement factor scores
ξ	Growth factor scores	ε	Item-level residual
μ	Growth factor means	$\boldsymbol{\Omega}$	Construct factor covariance matrix
ζ	Growth model residuals	$\boldsymbol{\Psi}$	Construct-level disturbances
$\boldsymbol{\Sigma}$	Observed variable covariance matrix	$\mathbf{SS}$	Sum scores of items at each time point
$\boldsymbol{\Phi}$	Growth factor covariance matrix	$\boldsymbol{\Lambda}_s$	Sum of item factor loadings at each time point
$\boldsymbol{\Theta}$	Residual covariance matrix	μ_s	Growth factor means fitted to sum scores
$\boldsymbol{\Lambda}$	Item factor loadings	$\boldsymbol{\Phi}_s$	Growth factor covariances fitted to sum scores

denoted $\boldsymbol{\Lambda}_s$, multiplied by the growth loadings and the true growth means,

$$E\left[\mathbf{SS}_i\right] = \boldsymbol{\Lambda}_s\boldsymbol{\Gamma}\mu \tag{9}$$

Summing the items therefore results in a vector of T total sum scores, called $\mathbf{SS}_i$. An explicit presentation of the sum score loading matrix is presented in "Appendix 1". To achieve the first aim of the paper, bias derivations based on these estimates are presented below.

Bias derivations

Bias is first evaluated in the means of the growth factors. From Eq. (6), matrix operations are used to solve for $\boldsymbol{\mu}$. Detailed derivations are available in "Appendix 2", and Table 1 defines the matrices used in these models. This gives the expression for the growth factor means,

$$\hat{\mu} = \left(\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'E\left[\mathbf{y}_i\right] \tag{10}$$

Equation (10) can be used to obtain estimates $\hat{\mu}$ for any set of values for $\boldsymbol{\Lambda}$, $\boldsymbol{\Gamma}$, or $E[\mathbf{y}_i]$. Of particular interest in this investigation is a given constrained measurement loading matrix, which will be generally referred to as $\tilde{\boldsymbol{\Lambda}}$. Equation (6) provides the expectation of $\mathbf{y}_i$, $E\left[\mathbf{y}_i\right] = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\mu$, which is invariant to the constraints of $\tilde{\boldsymbol{\Lambda}}$, so estimates are calculated under this measurement model by modifying (10) to:

$$\hat{\boldsymbol{\mu}} = \left(\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\mu} \tag{11}$$

The bias can be evaluated in the mean growth factors for incorrect measurement models by calculating $\mu - \hat{\mu}$ and the relative bias by

$$\text{bias}_r = \frac{\hat{\mu} - \mu}{\mu} \tag{12}$$

Similarly, an expression that calculates the variance of the parameter estimates from constrained measurement models can be derived. Equation (9) can be used to solve for $\boldsymbol{\Phi}$, the variance–covariance of the growth parameters. This result gives

$$\hat{\boldsymbol{\Phi}} = \left(\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'(\boldsymbol{\Sigma}_y - \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' - \boldsymbol{\Theta})\boldsymbol{\Lambda}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\right)^{-1} \tag{13}$$

Equation (12) can be used to obtain estimates for $\hat{\boldsymbol{\Phi}}$ by replacing $\boldsymbol{\Lambda}$ with the constrained loading matrix $\tilde{\boldsymbol{\Lambda}}$ and using Eq. (9) for $\boldsymbol{\Sigma}_y$,

$$\boldsymbol{\Phi} = \left(\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\left(\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\right)\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\right)^{-1} \tag{14}$$

Notice that the $\boldsymbol{\Lambda}$'s in the middle parentheses are not the constrained $\tilde{\boldsymbol{\Lambda}}$'s. These represent the population-level covariance structure, just as the known expected value of $\mathbf{y}_i$ was used for the means above. Then, $\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}$ gives the bias in the variance–covariance of the growth parameters and the relative bias is given by

$$\text{bias}_r = \frac{\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}}{\boldsymbol{\Phi}} \tag{15}$$

If these biased values are proportional to the true parameters, then the bias is a trivial scaling issue. If, however, is not proportional to $\boldsymbol{\Phi}$, then systematic bias is present, which can lead to misleading results in practice. Furthermore, an estimated $\hat{\boldsymbol{\Phi}}$ can be found for any constrained loading matrix $\tilde{\boldsymbol{\Lambda}}$, and bias can be evaluated thusly. Again, further details of the derivations are outlined in "Appendix 2".

Now consider fitting a univariate LGM to unweighted sums of items at each time point, that is, to standard sum scores computed at each measurement. The sum score model fits a univariate LGM to the$\mathbf{SS}_i$. This model treats the constrained loading matrix $\tilde{\boldsymbol{\Lambda}}$ essentially as an identity matrix because each item is equally weighted. This can be written,

$$E[\mathbf{SS}_i] = \boldsymbol{\Gamma}\mu_s \tag{16}$$

where $\mathbf{SS}_i$ is the aforementioned vector of sum scores for individual i and μ_s is a vector of intercept and slope means when the univariate model is fitted to the sum scores. The

estimated intercept and slope means for the sum score then have the form,

$$\hat{\mu}_{\mathbf{s}} = \left(\mathbf{\Gamma}'\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}' E[\mathbf{SS}_i] \tag{17}$$

with $E[\mathbf{SS}_i]$ given by Eq. (9). The estimated intercept and slope covariance matrix is similarly derived for sum scores, where $\tilde{\mathbf{\Lambda}}$ is treated as identity from Eq. (13) and $\mathbf{\Lambda}_{\mathbf{s}}$ from Eq. (9) is the population-level $\mathbf{\Lambda}$ for the summed items, giving,

$$\hat{\mathbf{\Phi}}_{\mathbf{s}} = (\mathbf{\Gamma}'\mathbf{\Gamma})^{-1}\mathbf{\Gamma}'(\mathbf{\Lambda}_{\mathbf{s}}\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}'\mathbf{\Lambda}_{\mathbf{s}}')\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1} \tag{18}$$

Practically speaking, the estimates of the growth means and variances using a sum score are a function of the sums of the true underlying item-level factor loadings. What remains to be seen is whether or not the growth parameter estimates are simply rescaled or are systematically biased under violations of MI. In the next section we therefore use the derived expressions to calculate the bias in growth factor means and variances for different violations of MI, and we validate the analytic results in a simulation study.

Calculation of bias for three different violations of MI

Conditions

Three hypothetical scenarios were used to illustrate the results of the above derivations. In all three examples, there were 4 time points with 6 items at each time point. The population values remained consistent across conditions, and are presented below:

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \mathbf{\Gamma} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \quad \mathbf{\Theta} = \sigma^2\mathbf{I}_{24} \quad \sigma^2 = 1$$

$$\mathbf{\Phi} = \begin{bmatrix} .5 & .2 \\ .2 & .5 \end{bmatrix} \quad \mathbf{\Psi} = \begin{bmatrix} .5 & 0 & 0 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & 0 & .5 \end{bmatrix}$$

where $\mathbf{I}_{24}$ is a 24 × 24 identity matrix,[3] and $\mathbf{\Theta} = \sigma^2\mathbf{I}_{24}$ indicates that the items have independent and identical error variances. For ease of presentation, assume that error variances are invariant across item and across time. The only manipulation across the three conditions involved the factor loading matrix $\mathbf{\Lambda}$. The bias estimates were calculated based on the bias derivation presented above, and Monte Carlo simulations were then conducted to support these results empirically.

The three item conditions are as follows: (1) item loadings are not all equivalent within a particular time point, but MI holds over age; (2) MI does not hold over age as certain items decline in their relation to the factor after the second time-point; and (3) MI does not hold, but 2 items decrease systematically in their loadings while 2 increase systematically, such that the changes essentially balance out. The explicit form of the $\mathbf{\Lambda}$'s for the demonstrations is presented in "Appendix 3". These three loading matrices could correspond to three different conditions that might occur in practice in an investigation of aggression in children. In the first condition, the items are not all equally important in how well they measure aggression, but their importance is consistent over time. In the second condition, items 5 and 6 become less important as children age, corresponding to a particular behavior that is relevant for aggression in young children but not in older children. The third condition reflects a situation in which some items are highly related to aggression at a younger age only to decrease in importance over development, while others are not important for the youngest children but become more important as children age.

Bias calculations

Using the information above, the sum score estimates were calculated under Condition 1, in which MI holds but the items do not have consistent weights. From Eq. (9), the expected values of the sum scores are known to be $E[\mathbf{SS}_i] = \mathbf{\Lambda}_{\mathbf{s}}\mathbf{\Gamma}\mu$, which were obtained with the sum of the item loadings at each time point, $\mathbf{\Lambda}_{\mathbf{s}}$,

$$\mathbf{\Lambda}_{\mathbf{s}} = \begin{bmatrix} 4.4 & 0 & 0 & 0 \\ 0 & 4.4 & 0 & 0 \\ 0 & 0 & 4.4 & 0 \\ 0 & 0 & 0 & 4.4 \end{bmatrix} \tag{19}$$

Growth factor estimates for the sum scores were calculated by applying these expected sum scores to Eq. (15):

$$\hat{\mu}_{\mathbf{s}} = (\mathbf{\Gamma}'\mathbf{\Gamma})^{-1}\mathbf{\Gamma}'\mathbf{\Lambda}_{\mathbf{s}}\mathbf{\Gamma}\mu = \begin{bmatrix} 4.4 \\ 4.4 \end{bmatrix} \tag{20}$$

$$\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \text{bias}_r(\hat{\mu}_{\mathbf{s}}) = \begin{bmatrix} 3.4 \\ 3.4 \end{bmatrix} \tag{21}$$

From Eq. (16) the growth factor covariance matrix was calculated:

$$\hat{\mathbf{\Phi}}_{\mathbf{s}} = (\mathbf{\Gamma}'\mathbf{\Gamma})^{-1}\mathbf{\Gamma}'(\mathbf{\Lambda}_s\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}'\mathbf{\Lambda}_s')\mathbf{\Gamma}(\mathbf{\Gamma}'\mathbf{\Gamma})^{-1} = \begin{bmatrix} 9.680 & 3.872 \\ 3.872 & 9.680 \end{bmatrix} \tag{22}$$

[3] As discussed above, the dimensions of $\mathbf{\Theta}$ are $(T * p) \times (T * p)$; in this case, $(4 * 6) \times (4 * 6)$, hence the need for a 24 × 24 identity matrix here.

$$\boldsymbol{\Phi} = \begin{bmatrix} .5 & .2 \\ .2 & .5 \end{bmatrix}; \text{bias}_r(\hat{\boldsymbol{\Phi}}_s) = \begin{bmatrix} 18.360 & 18.360 \\ 18.360 & 18.360 \end{bmatrix} \quad (23)$$

The estimates are proportional to true parameters, that is, the bias terms are all the same, and the relative bias is constant across the sets of parameters. Therefore, when MI holds, the sum score estimates are simply weighted or rescaled values of the true parameters. In other words, the bias is due to scaling, and is therefore inconsequential.

The same procedure was repeated for Condition 2, where the only change was the manipulation of the item-level loading matrix $\boldsymbol{\Lambda}$. Here, some items decreased in their true loadings over time. Again, as demonstrated in Eq. (9), the item loadings were summed at each time point to obtain the sum scores computed from the SLGM generation,

$$\boldsymbol{\Lambda}_s = \begin{bmatrix} 4.5 & 0 & 0 & 0 \\ 0 & 4.3 & 0 & 0 \\ 0 & 0 & 3.9 & 0 \\ 0 & 0 & 0 & 3.9 \end{bmatrix} \quad (24)$$

The matrix of summed loadings is applied to Eqs. (15) and (16):

$$\hat{\mu}_s = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}_s\boldsymbol{\Gamma}\mu = \begin{bmatrix} 4.64 \\ 3.64 \end{bmatrix} \quad (25)$$

$$\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \text{bias}_r(\hat{\mu}_s) = \begin{bmatrix} 3.64 \\ 2.64 \end{bmatrix} \quad (26)$$

$$\hat{\boldsymbol{\Phi}}_s = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'(\boldsymbol{\Lambda}_s\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}_s')\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1} = \begin{bmatrix} 10.335 & 3.268 \\ 3.268 & 7.134 \end{bmatrix} \quad (27)$$

$$\boldsymbol{\Phi} = \begin{bmatrix} .5 & .2 \\ .2 & .5 \end{bmatrix}; \text{bias}_r(\hat{\boldsymbol{\Phi}}_s) = \begin{bmatrix} 19.670 & 15.338 \\ 15.338 & 13.269 \end{bmatrix} \quad (28)$$

In this condition, it is apparent that the estimates of the growth factor means are systematically biased. Instead of the intercept and slope means having a 1-to-1 relationship as when they are generated, estimates from a univariate sum score lead to an intercept mean that is 1.27 that of the slope means. In other words, the average growth trajectory is underestimated, whereas the average baseline level is overestimated. Similarly, the variance for the intercept is proportionally larger than the variance of the slope, though they have the same value in the population.

The calculations for Condition 3 again followed the same procedure, but in this condition the loadings increased for two items from time 1 to time 4 and decreased for two other items from time 1 to time 4. These items increased and decreased in a reciprocal pattern. From the true loading matrix, sum scores were created, resulting in the following loading matrix,

$$\boldsymbol{\Lambda}_s = \begin{bmatrix} 3.5 & 0 & 0 & 0 \\ 0 & 3.5 & 0 & 0 \\ 0 & 0 & 3.5 & 0 \\ 0 & 0 & 0 & 3.5 \end{bmatrix} \quad (29)$$

Although the item loadings for some items changed across measurement occasions, the sum of the item loadings at each time remained constant. The growth estimates were obtained from Eqs. (15) and (16),

$$\hat{\mu}_s = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}_s\boldsymbol{\Gamma}\mu = \begin{bmatrix} 3.5 \\ 3.5 \end{bmatrix} \quad (30)$$

$$\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \text{bias}_r(\hat{\mu}_s) = \begin{bmatrix} 2.5 \\ 2.5 \end{bmatrix} \quad (31)$$

$$\hat{\boldsymbol{\Phi}}_s = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'(\boldsymbol{\Lambda}_s\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}_s')\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1} = \begin{bmatrix} 6.125 & 2.450 \\ 2.450 & 6.125 \end{bmatrix} \quad (32)$$

$$\boldsymbol{\Phi} = \begin{bmatrix} .5 & .2 \\ .2 & .5 \end{bmatrix}; \text{bias}_r(\hat{\boldsymbol{\Phi}}_s) = \begin{bmatrix} 11.250 & 11.250 \\ 11.250 & 11.250 \end{bmatrix} \quad (33)$$

These results demonstrate that there was no bias in the growth factor estimates based on sum scores, even though the items themselves were not perfectly invariant. Clearly, this is due to the fact that in this particular scenario the sum of the item loadings remains constant over time.

Verification with simulated data

Monte Carlo simulations were conducted to validate the results of the analytic derivations. Using the population values given above, response data were generated using the SLGM. Data were simulated with a mean structure given by inserting population values into Eq. (9), and the covariance structure was given by Eq. (14). One thousand simulations were conducted for each condition with a sample size of $N = 1000$ each. Data were simulated using *R* (R Development Core Team, 2014). Sum scores of the item-level data were calculated and LGMs were fit to the sum scores using the *lavaan* package in *R* (Rosseel 2012; R Core Team 2017).

The results of the 1000 Monte Carlo simulations were directly compared to the analytic estimates for the means and covariance matrix of the intercept and slope factors. The average estimates and empirical standard errors are presented side-by-side with the analytic results in Tables 2 (means) and 3 (covariance matrices) below. The average values from the Monte Carlo simulations were consistent with the analytic calculations for all conditions for both the means and covariance matrix.

Table 2 Calculated and simulated growth factor means

Cond.		True $\boldsymbol{\mu}$	$\boldsymbol{\mu}_s$	$\hat{\boldsymbol{\mu}}_s$	SE ($\hat{\boldsymbol{\mu}}_s$)	bias$_r$ ($\boldsymbol{\mu}_s$)	$\boldsymbol{\mu}/\boldsymbol{\mu}_s$
1	μ_α	1.0	4.40	4.3939	0.1426	3.40	0.2272
	μ_β	1.0	4.40	4.3979	0.1133	3.40	0.2272
2	μ_α	1.0	4.64	4.6343	0.1456	3.64	0.2158
	μ_β	1.0	3.64	3.6347	0.1002	2.64	0.2751
3	μ_α	1.0	3.50	3.4951	0.1200	2.50	0.2857
	μ_β.	1.0	3.50	3.4983	0.0926	2.50	0.2857

Analytic and empirical estimates of growth factor means. μ_s refers to analytic derivations, and $\hat{\mu}_s$ refers to the averaged empirical estimates, μ_α and μ_β are the intercept and slope factors, respectively. The relative bias of estimates and proportions of true parameters to the estimates are also presented

Table 3 Calculated and simulated growth factor variances and covariances

Cond.		$\boldsymbol{\Phi}$	$\boldsymbol{\Phi}_s$	$\hat{\boldsymbol{\Phi}}_s$	SE ($\hat{\boldsymbol{\Phi}}_s$)	Bias$_r$ ($\boldsymbol{\Phi}_s$)	$\boldsymbol{\Phi}/\boldsymbol{\Phi}_s$
1	σ^2_α	0.5	9.680	9.5684	1.1489	18.360	0.0517
	σ^2_β	0.5	9.680	9.6630	0.5949	18.360	0.0517
	$\sigma_{\alpha\beta}$	0.2	3.872	3.8772	0.5158	18.360	0.0517
2	σ^2_α	0.5	10.3347	9.9368	1.1636	19.6694	0.0484
	σ^2_β	0.5	7.1343	7.0610	0.4841	13.2686	0.0701
	$\sigma_{\alpha\beta}$	0.2	3.2675	3.3927	0.5562	15.3375	0.0612
3	σ^2_α	0.5	6.125	6.0413	0.8339	11.250	0.0816
	σ^2_β	0.5	6.125	6.1133	0.4025	11.250	0.0816
	$\sigma_{\alpha\beta}$	0.2	2.450	2.4578	0.4225	11.250	0.0816

Analytic and empirical estimates of growth parameter covariances. $\boldsymbol{\Phi}_s$ refers to analytic derivations, and $\hat{\boldsymbol{\Phi}}_s$ refers to the averaged empirical estimates. The relative bias of estimates and proportions of true parameters to the estimates are also presented

Summary of bias calculation

The calculations from the three hypothetical conditions analytically quantified the bias in growth factors that results from a violation of MI in the item loadings. The first condition, in which item weights were not equal across all items, resulted in growth factor estimates proportional to the true estimates because the pattern of item loadings was consistent across measurement occasion. In this case, using a sum score would lead to the correct inference on the intercept and slope means. Additionally, the estimated intercept-slope covariance matrix was proportional to the true covariance matrix.

In Condition 2, MI was violated in two of the six items, which were characterized by declining factor loadings over time. Fitting a LGM to sum scores from Condition 2 results in systematic bias of the intercept and slope factor means and covariances. In particular, the population means stipulated that the intercept and the growth means were equal to each other; the estimates from the Condition 2 resulted in a slope mean that is approximately 78% of the intercept estimate. Additionally, the amount of variance in the intercept and slope factors, which was equal in the data generating model, was drastically different in the sum score estimates. Inference is often made on the slope means relative to the baseline level in a growth model. Condition 2 reveals a case in which this inference would be particularly misguided. Condition 3 represents a case in which MI did not hold because item loadings were consistently increasing for some items and decreasing for others, yet the sum of the items at each time remained the same. This reflects a scenario in which individual items may shift to accommodate age-salient changes in behavior, but the underlying behavior itself remains constant, and the overall aggregate of the instrument consistently measures the construct of interest.

Empirical illustration

In support of the bias derivation and simulations, an application illustrating how to understand sum score bias in growth parameters due to longitudinal non-invariance is presented. The application assesses the costs and benefits of using a univariate phenotype in the "ACTION:

Table 4 CBCL items and respective factor loadings by age and gender

Item	Age 3	Age 7	Age 10	Age 12
Females				
Destroys things belonging to his/her family or other children	0.57	0.95	0.92	0.93
Gets in many fights	0.74	0.65	0.64	0.65
Physically attacks others	0.73	0.64	0.66	0.72
Hits others	0.82	–	–	–
Hurts animals or children without meaning to	0.60	–	–	–
Destroys his/her own things	–	0.90	0.92	0.90
Threatens people	–	0.37	0.44	0.58
Males				
Destroys things belonging to his/her family or other children	0.59	0.94	0.91	0.91
Gets in many fights	0.74	0.63	0.70	0.72
Physically attacks others	0.77	0.70	0.77	0.81
Hits others	0.83	–	–	–
Hurts animals or children without meaning to	0.62	–	–	–
Destroys his/her own things	–	0.89	0.86	0.88
Threatens people	–	0.64	0.70	0.76

Physical aggression items identified, and their corresponding factor loadings for both males and females. Notice that the items "destroys own things" and "threatens people" were not administered at age 3, while "hits others" and "hurts others without meaning to" were only administered at age 3

Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies" consortium. ACTION was formed to investigate genetic and environmental contributions to the development of aggression throughout childhood and adolescence (Boomsma 2015). Aggression in childhood is heritable (e.g. Eley et al. 1999; Hudziak et al. 2003; van der Valk et al. 1998). Literature suggests that physical aggression maintains a consistent level as children develop, though overall aggression may decline across childhood on average (Burt 2009; Tremblay 2003). Genetic factors have been shown to account for individual differences in childhood aggression at specific ages and longitudinally across development (Porsch et al. 2016; van Beijsterveldt et al. 2003). Given the goal of ACTION to unravel the environmental and genetic contributions to the development of aggression, the construction of an aggression phenotype is critically important to provide the basis for accurate inferences concerning aggression trajectories.

The Young Netherlands Twin Register (YNTR) is a population-based sample of Dutch twins observed from ages 2 to 16 (van Beijsterveldt et al. 2013). From 1986 to 2011, families with newborn twins were recruited to participate in the YNTR, and most families (89%) were registered within 12 months of birth. This illustration uses a small subsample of the NTR data that was previously set aside for exploratory data analyses of the aggression factor structure (Lubke et al. in press). This subsample consisted of 591 male individuals and 1009 female individuals. Due to established gender differences in aggressive behaviors, the growth parameter bias calculation was carried out for males and females separately (Bartels et al. 2003; Hudziak et al. 2003).

At age 3, children's aggressive behaviors were measured by mother report of the aggressive behaviors subscale of the Child Behavior Checklist (CBCL) 1.5-5 (Achenbach and Rescorla 2000), a version of the CBCL adapted for preschool-aged children. The CBCL 1.5–5 contains 19 items rated on a 3-point likert scale (0 = "not true", 1 = "somewhat/sometimes true", 2 = "very/often true of my child in the past six months"). At ages 7, 10, and 12, mother-reported data were collected with the aggressive behaviors subscale of the CBCL 6–18 (Achenbach and Rescorla 2001). The CBLC 6–18 contains 18 items rated on the same likert scale as the CBCL 1.5–5. Exploratory Factor Analysis of these data identified 5 items for each CBCL version that loaded onto a physical aggression factor (detailed in Lubke et al. in press). The illustration presented in this paper will focus on the physical aggression phenotype. Confirmatory Factor Analysis (CFA) was conducted to establish a factor structure at each age. Table 4 contains the items included and their factor loadings at each age.

Table 4 shows that the physical aggression items identified from the CBLC 1.5–5 are not exactly the same as the physical aggression items for the CBCL 6–18. In fact, there are three consistent items and two unique items across survey form. Two physical aggression items are removed and two new items are added in the different versions of the CBCL. Additionally, the consistent item "destroys others' things" increases in its item loading from around 0.58 to greater than 0.9 after age 3. In practice, researchers have a few options on how to proceed: using all available items

at each time point in a sum score, using all available items in an SLGM (thus allowing for the non-overlapping items to have different loadings at age 3), or using only the three consistent items in a sum score or SLGM.

The illustration presented below uses a sum score of all available items because both versions of the CBCL subscale aim to measure the same underlying construct of physical aggression across all ages and because composite scores of slightly varying item sets are often used in practice (Forsman et al. 2010; Wang et al. 2013; van Beijsterveldt et al. 2003). To assess the consequences of this approach, the bias attributable to using a sum score of all available items is calculated for hypothetical growth parameters in an LGM. The CFA results (Table 4) are used as the factor structure for each measurement occasion. Previous literature has indicated that physically aggressive behaviors in children are best characterized by multiple trajectories: starting low and increasing, starting low and remaining low, starting high and decreasing, or remaining high (Burt 2009; Cui et al. 2016; NICHD 2004). For simplicity in this illustration, and because our simulations had a positive slope, this example posits the low-increasing trajectory of aggression, with growth parameters specified as $E[\xi] = [\mu_\alpha \ \mu_\beta]' = [1 \ 0.2]'$. This represents a slight increase over age, with the rate of growth modest relative to the initial level (intercept to slope ratio of 5 to 1). The slope loadings in $\boldsymbol{\Gamma}$differ from the simulation above to correspond to the appropriate age intervals in this example; namely, the vector of slope loadings is $[0 \ 4 \quad 7 \ 9]'$. The other population values from the simulation are used to fill in the remaining parameters, in order to simply evaluate the bias in growth parameter estimates due to forcing measurement non-invariant items into a sum score.

Empirical results

The estimates for $\hat{\mu}_s$ and $\hat{\boldsymbol{\Phi}}_s$ are calculated based on sum scores formed from the items with the factor structure described in Table 4. The results for aggression based on the data from females are presented first:

$$\hat{\mu}_s = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}_s\boldsymbol{\Gamma}\mu = \begin{bmatrix} 3.336 \\ 0.781 \end{bmatrix}$$
$$\text{bias}_r(\hat{\mu}_s) = \begin{bmatrix} 2.336 \\ 2.905 \end{bmatrix} \tag{34}$$

$$\hat{\boldsymbol{\Phi}}_s = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'(\boldsymbol{\Lambda}_s\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}_s')\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1} = \begin{bmatrix} 5.358 & 1.804 \\ 1.804 & 7.052 \end{bmatrix}$$
$$\text{bias}_r(\hat{\boldsymbol{\Phi}}_s) = \begin{bmatrix} 9.717 & 8.018 \\ 8.018 & 13.104 \end{bmatrix} \tag{35}$$

It is clear that the intercept and slope estimates under of the sum score model are biased. Importantly, the relative bias does not indicate the bias is a trivial scaling issue. This is apparent when we consider that the ratio of the intercept to slope means should be 5 to 1, but it is closer to 4.25 to 1, indicating that the slope estimate resulting from the use of sum-scores would imply a faster rate of growth relative to the intercept than is stipulated in the population. The estimated variance in the slope is also upwardly biased relative to the intercept variance. The intercept-slope correlation is 0.29 in these calculations, but it is 0.4 in the population.

The results for the males are similar:

$$\hat{\mu}_s = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}_s\boldsymbol{\Gamma}\boldsymbol{\mu} = \begin{bmatrix} 3.468 \\ 0.870 \end{bmatrix}$$
$$\text{bias}_r(\hat{\mu}_s) = \begin{bmatrix} 2.468 \\ 3.350 \end{bmatrix} \tag{36}$$

$$\hat{\boldsymbol{\Phi}}_s = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'(\boldsymbol{\Lambda}_s\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}_s')\boldsymbol{\Gamma}(\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1} = \begin{bmatrix} 5.795 & 2.104 \\ 2.104 & 8.344 \end{bmatrix}$$
$$\text{bias}_r(\hat{\boldsymbol{\Phi}}_s) = \begin{bmatrix} 10.589 & 9.520 \\ 9.520 & 15.688 \end{bmatrix} \tag{37}$$

The pattern of results is the same as the females, with non-ignorable bias introduced into the intercept and slope means and variances due to sum score measurement model. The estimated slope mean shows upward bias relative to the intercept, as the ratio of the intercept to the slope is 4 to 1 rather than 5 to 1, and the estimated slope variance is disproportionately large compared to the estimated intercept variance. These results imply that a univariate LGM fit to sum scores of the identified physical aggression items would result in biased growth estimates.

Implication of bias for variance decomposition

An effective longitudinal twin modeling strategy is to fit an LGM to data collected on a longitudinal phenotype and decompose the variance in the intercept and slope factors (Finkel et al. 2015; Lubke et al. 2016; Neale and McArdle 2000). This is often carried out with a genetic model, which decomposes the variance of a phenotype into portions due to additive genetic effects (A), common environment (C), and non-shared environment (E) components. The twin modeling framework estimates these variance components by modeling the expected relationships between pairs of twins. For example, monozygotic (MZ) twins are genetically identical, but dizygotic (DZ) twins share on average 50% of their genetic information. The same model is fit to pairs of MZ twins and

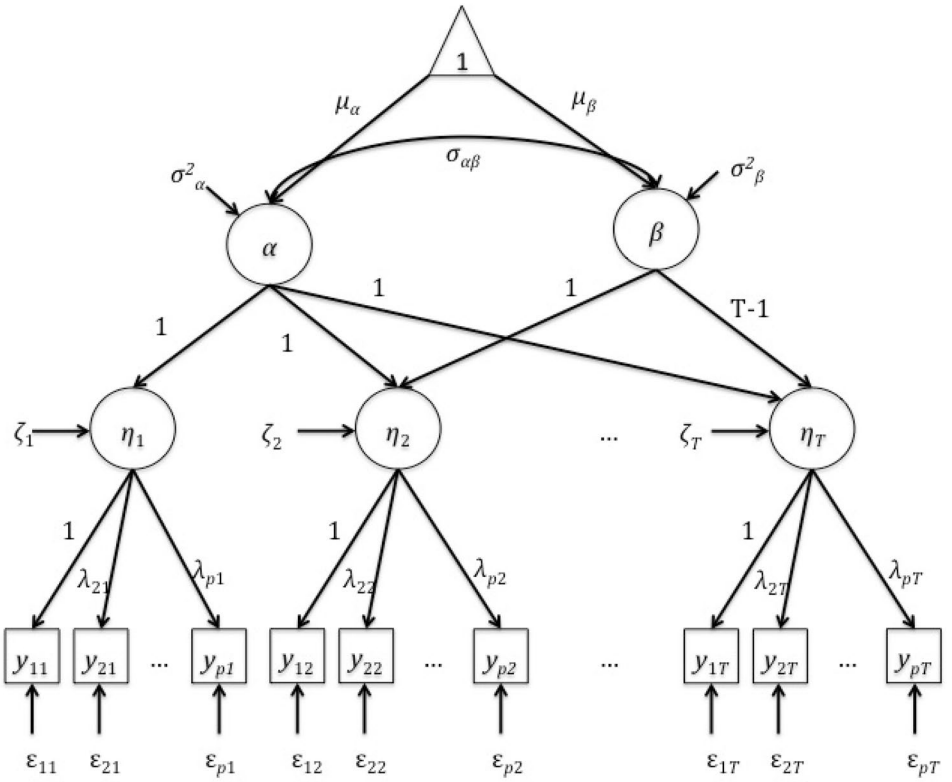

Fig. 1 Example of a second-order latent growth model with $j = 1, \ldots, p$ items and $t = 1, \ldots, T$ time points. α = intercept and β = slope represent latent growth constructs that influence the latent phenotypes $\boldsymbol{\eta}$ which in turn define observed indicators $\mathbf{y}$

DZ twins, but the A component is perfectly correlated among MZ twins and correlated 0.5 for DZ twins. The C component is perfectly correlated across twins of both zygosities to capture shared variance not found in the A component, and the unique component E is completely uncorrelated across twins of both zygosities. The E component includes all non-shared environmental effects as well as measurement error. In the aggression example above, the variances of the intercept and slope fitted to sum scores (denoted $\hat{\boldsymbol{\Phi}}_s$ in Eq. 35) would be decomposed into the A, C, and E components. It was established above that the use of sum scores when MI is violated over time leads to a biased variance–covariance matrix for the intercept and slope factors. We conducted a simulation to investigate if bias due to longitudinal non-invariance leads to biased variance decomposition.

Longitudinal twin data were simulated by generating growth factors under an ACE path coefficients model composing linear intercept and slope factors. Figure 2 depicts a linear LGM of sum scores for twin 1 of a twin pair, and Fig. 3 depicts the ACE decomposition of intercept and slope factors. The variance of the intercept and slope factors were decomposed into A, C, and E components, where the path coefficients were set to $a = \sqrt{.5}, c = \sqrt{.25}$, and $e = \sqrt{.25}$. These path coefficients represent the proportion of the intercept and slope factors due to each component because the A, C, and E factors have unit variance. The correlations among the A, C, and E factors were all set to 0.5 to maintain a marginal correlation of 0.5 between the intercept factor and slope factor for a given individual. The zygosity coefficient, which we call z here,[4] corresponds to the correlation of

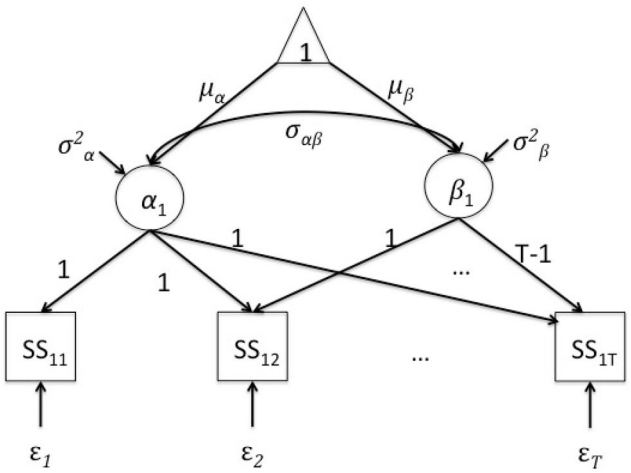

Fig. 2 Linear LGM of sum scores for twin 1 in a twin modeling framework. "SS_{11}" indicates the sum score at the first time-point for twin 1. The intercept and slope factors are denoted α and β

[4] This "zygosity coefficient" is conventionally labeled α, but we use z so as not to confuse it with the label for the intercept factor.

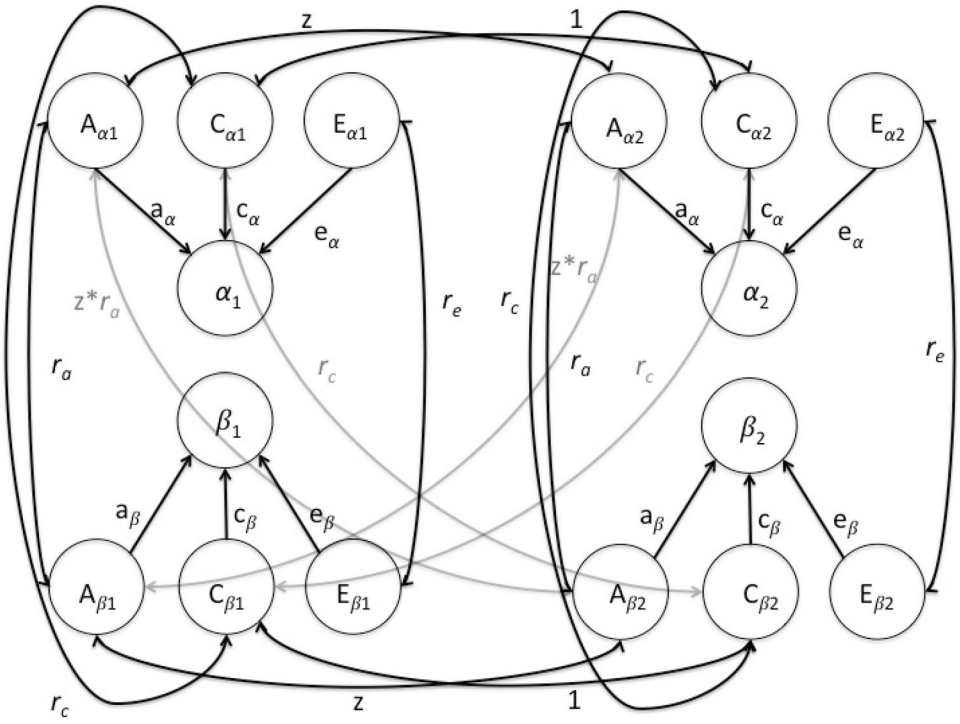

Fig. 3 Path diagram of ACE decomposition of intercept and slope growth factors with intercept-slope correlation. r indicates correlation of intercept and slope in respective ACE factors; α and β subscripts indicate intercept and slope, respectively; 1 and 2 indicates twin 1 and twin 2 in a particular twin pair; $z = 1$ for MZ twins, 0.5 for DZ twins

the A component between MZ twins or DZ twins. One thousand simulations were conducted per condition with $N = 1000$ twin pairs generated in each simulation (500 MZ and 500 DZ twins). All model parameters were set equal across twin pairs and MZ and DZ twins, with the exception of the z coefficient.

Data were generated in *R* (R Development Core Team 2017), and the model fitting was carried in Mplus Version 7.3 using the external Monte Carlo feature (Muthén and Muthén 1998–2015). Abbreviated Mplus code is available in "Appendix 4". Data were generated under the three different conditions described above, varying only by the item-level measurement model. Sum scores of the items were computed at each time point, the LGM was fitted to the sum scores, and the growth factor variances were decomposed into ACE components. The ACE simulations were evaluated on their proportional variance components estimates by dividing the squared path coefficients that represent genetic and environmental variance components by the total variance. In other words, the heritability estimate of interest was $\frac{a^2}{a^2 + c^2 + e^2}$ for both the intercept and slope variance. This provided the estimated proportion of the variance in the intercept or slope due to A, C, and E. Because the growth factors in the sum score model are on a different scale than the factors in the item level model, we consider proportional variance components to be able to compare the results.

Variance decomposition results

The proportional path coefficient estimates corresponding to the intercept and slope are presented in Table 5, along with their empirical standard errors, mean squared error (MSE), coverage rates, and percentage significant across replications (power). The proportional decomposition of the intercept and slope factors was accurate for all three conditions, whether or not MI held over time. The common environment path coefficient for the intercept factors tended to have a larger standard error compared to the other estimates of interest, but the point estimates for all proportional path coefficients were true to the population values. The estimated intercept and slope means were identical to the derivations above, as expected, indicating that there is bias in the average growth trajectories in Condition 2. The total variance of the intercept and slope is also biased in Condition 2, but the proportional decomposition in the ACE path coefficients is correct. The conditions each had convergence rates of 99.3, 99.7, and 98.9%, respectively.

Discussion

This paper had two aims: to calculate expected bias in the intercept and slope factors when MI of the items does not hold over time but sum scores are used for LGMs, and to investigate the consequences of this bias for variance

Table 5 Results from simulated ACE decomposition of growth factor variances

Cond.		Pop.	Avg. estimate	Empirical SE	MSE	95% cover.	% signif.
1	a^2_α	0.50	0.4948	0.1509	0.0023	0.943	0.904
	c^2_α	0.25	0.2555	0.1245	0.0155	0.935	0.521
	e^2_α	0.25	0.2497	0.0535	0.0029	0.937	0.997
	a^2_β	0.50	0.4969	0.0957	0.0092	0.941	0.999
	c^2_β	0.25	0.2517	0.0833	0.0069	0.944	0.843
	e^2_β	0.25	0.2514	0.0281	0.0008	0.944	1.000
2	a^2_α	0.50	0.5072	0.1424	0.0203	0.951	0.935
	c^2_α	0.25	0.2542	0.1189	0.0141	0.953	0.522
	e^2_α	0.25	0.2387	0.0495	0.0026	0.943	0.997
	a^2_β	0.50	0.4991	0.1001	0.0100	0.942	0.997
	c^2_β	0.25	0.2528	0.0890	0.0079	0.936	0.807
	e^2_β	0.25	0.2481	0.0296	0.0009	0.947	1.000
3	a^2_α	0.50	0.5038	0.1576	0.0248	0.939	0.866
	c^2_α	0.25	0.2452	0.1304	0.0170	0.937	0.402
	e^2_α	0.25	0.2510	0.0587	0.0034	0.958	0.993
	a^2_β	0.50	0.4972	0.0973	0.0095	0.952	1.000
	c^2_β	0.25	0.2508	0.0867	0.0075	0.944	0.814
	e^2_β	0.25	0.2519	0.00293	0.0009	0.948	1.000

Results of ACE decomposition for simulated longitudinal twin data. Reported estimates are proportions of intercept and slope variances due to A, C, and E components. cond, condition; pop., population value; *MSE* mean squared error; α and β subscripts indicate intercept and slope, respectively; cover., coverage; signif., significant

decomposition. To address the first aim, expressions for estimating the growth factor means and (co)variance matrix with a sum score (or other constrained measurement model) were derived. These derivations were used to calculate bias under conditions with and without MI. Results showed that using a sum score when MI does not hold over time can result in an incorrect interpretation of growth. In Condition 2, for example, a subset of item loadings decreased over time. The resulting LGM fitted to the sum scores underestimated the true growth trajectory. This clearly illustrates how changes in the measurement model are confounded with the true change in the construct of interest. Condition 3 represented the special case of measurement changes balancing out over time, resulting in correct growth estimates. The first simulation showed that our derived estimates were correct. The empirical illustration depicted an example of calculating expected bias in the growth factors due to longitudinal measurement non-invariance. In this case, the estimated growth trajectories would be incorrect, in that growth would be overestimated due simply to measurement changes. Fitting a second order growth model would permit the estimation of age-specific loadings. In case loadings in fact change over time as hypothesized *a priori*, this would then allow the interpretation of an age-appropriately measured aggression phenotype.

The second aim of the paper was carried out by conducting a variance decomposition of simulated twin data. This simulation study revealed that though the interpretation of growth may be altered by longitudinal non-MI in sum scores, the variance decompositions of the intercept and slope variances derived from the sum score model were correct. In other words, the growth factors were biased by longitudinal non-invariance, but the proportion of variation in the growth factors due to genetic and environmental influences remained the same, and was accurately estimated. These results can be understood as follows. The variance decomposition into genetic, shared and unique components depends on MI across twins in a pair and across MZ and DZ groups. The measurement model is assumed to be the same across twin groups in the simulation, so the growth factors are systematically biased by longitudinal non-MI *in the same way* for each twin and the MZ and DZ groups.

Estimating the genetic and environmental components is only possible by assuming that the model is the same across groups, with the only exception being the zygosity coefficient specifying different correlations of genetic effects for MZ and DZ groups. As a result, the proportion of variance due to the different ACE components is not biased, even if the variance estimate itself is biased. Violations of MI longitudinally do not change how the intercept and slope are related across twins, provided that the measurement of the construct is the same across twins and groups.

The longitudinal MI problem presented in this paper often arises in longitudinal studies spanning early childhood through adolescence and young adulthood, where questionnaires that measure behaviors in young children are no longer age-appropriate as participants grow older. In some cases, questionnaires are replaced to adapt to the age of participants, and the method of reporting can even switch from parent-report to self-reported surveys. Changes in the measurement instrument itself, such as including new items at different ages, are apparent to the researcher and can be incorporated in models fitted to the data. The simulations presented here demonstrate that ignoring such changes in a sum score introduce bias of longitudinal growth patterns. An alternative approach is fitting the previously described SLGM, in which age-specific measurement models are fitted to the age-specific item sets.

Additionally, questionnaire items can be included at multiple ages and allowed to demonstrate non-invariance over time. An intuitive example of this is an item regarding behaviors that have different meanings at different ages. In our study of childhood aggression, for example, the item "threatens others" increases in loadings from age 7 to age 12. A 7-year-old child may not have the social tools or awareness to understand what a threat is and how to manipulate others by threats that a 12-year old has, meaning that a child can have the same level of true aggression at age 7 and 12, yet the child will score higher at age 12. The SLGM also allows researchers to incorporate this hypothesis in the model by allowing the factor loading on the item to be freely estimated over time. Researchers can formally evaluate this specification by carrying out Chi square tests of model comparisons (Bollen 1989; Byrne et al. 1989). However, one should exercise caution in permitting non-invariance over time simply to improve model fit. Questions remain regarding if the same construct is truly measured over time when MI fails to hold (Edwards and Wirth 2009, 2012; Meredith and Horn 2001; Widaman et al. 2010). Without some degree of consistent measurement over time, it is not possible to disentangle measurement differences from true change in the construct, and it is recommended that a majority of overlapping items is fixed to invariance in SLGMs (Widaman et al. 2010).

It should be noted that the work presented here pertains to the modeling of a single construct, and it is assumed that all items are in fact indicators of the same common phenotype. In other words, the items should be at least congeneric, pertaining to the same underlying true trait. Though the measurement properties and content of the instruments can manifest age-specific changes, the same underlying construct is measured at all times. In practice, exploratory psychometric analyses could be carried out to assess dimensionality and identify items not loading onto a common phenotype.

Longitudinal models coupled with twin data can provide information about genetic and environmental contributions to a baseline measure and a growth trajectory of the phenotype. Measurement non-invariance over time leads to complications in modeling growth in the phenotype, but the genetic decomposition of the growth factors remains valid. This is because the measurement properties *are* invariant with respect to twin zygosity and across twins within pairs. Any bias introduced into the growth factors is the same for all twins participating in the study. Therefore, the variability in estimated phenotypic baseline and trajectory is still captured for all twins. The genetic and environmental contribution to the trajectory of the phenotype can then be accurately estimated in the presence of measurement non-invariance over time, provided that MI is realistic across twins and other grouping variables. This result holds for the LGM framework, but it may not hold for other longitudinal twin models, such as the genetic simplex model. The genetic simplex decomposes the phenotype at each measurement occasion and models transmission (autocorrelation) and innovation (time-specific phenotypic variance not explained by prior measurements) of the ACE variance components. The simplex model provides an indication of the transmission of genetic and environmental effects over time, whereas the LGM estimates genetic and environmental contributions to variation in a modeled phenotypic trajectory. With measurement non-invariance over time, which may indicate different degrees of measurement error at different time points, the E component could be inflated when invariance is imposed, thus shrinking proportional contribution of the A and C components. The LGM framework does not encounter this problem because it captures the structural change of the phenotype over time in the latent growth factors. Future research is needed to evaluate the impact of measurement non-invariance on the variance decomposition in a genetic simplex.

In contrast to the possibility of age-specific measurement across development, it is quite reasonable to assume the measurement model is the same across twin pairs and across MZ/DZ twin groups. The designation of twins in a pair as twin 1 and twin 2 is typically arbitrary, and there are limited cases where one might believe that twin

pairs are measured differently depending on if they are MZ or DZ twins (c.f., Neale et al. 2005). Furthermore, MI across zygosity can be explicitly tested by comparing the fits of item-level factor models with and without MI constraints.

It is important to note, however, that unbiased ACE variance estimates can only be obtained in longitudinal sum score models if MI holds across zygosity and across twins within pairs. If, for instance, data from boys and girls are used in a joint analysis, then measurement non-invariance across sex would lead to a spurious detection of sex limitation (Lubke et al. 2004). This problem can manifest in a longitudinal study in multiple ways: either through a scale that operates differently across males and females at all time points, or a scale that follows different patterns of longitudinal non-MI for males and females. Other researchers have extended the incorrect detection of genotype by environment interactions because of sum scores in other cases of measurement instrument issues, such as heterogeneous measurement errors or scaling problems (Molenaar and Dolan 2014; Schwabe and van den Berg 2014). Problems from these types of measurement issues would only be exacerbated in the longitudinal case. Careful consideration of these other measurement concerns must also be taken for longitudinal modeling.

The benefits of item-level measurement models for twin data are extolled in the literature (Molenaar and Dolan 2014; Neale et al. 2005; Schwabe and van den Berg 2014; van den Berg et al. 2007). This paper does not intend to downplay the importance or usefulness of testing for longitudinal MI when possible. On the contrary, this paper directly shows that the average growth trajectories will be misinterpreted when MI does not hold over time but sum scores are used. There are cases, however, in which sum scores or other summary data are the only pieces of information available, or computational constraints necessitate simpler measurement models. In the case of decomposing the variance of the intercept and the variance of the slope estimated from sum scores, conclusions about genetic and environmental proportional effects are still reasonable. Caution is necessary due to their oft-ignored assumptions, but sums scores are useful in these contexts.

Acknowledgements We acknowledge grant FP7-602768 "ACTION: Aggression in Children: Unraveling gene-environment interplay to inform Treatment and InterventiON strategies" from the European Commission/European Union Seventh Framework Program. GL was in addition supported by DA-018673. The Netherlands Twin Register is supported by multiple grants from the Netherlands Organization for Scientific Research (NWO) and MagW/ZonMW (Grants 904-61-090, 985-10-002, 904-61-193, 480-04-004, 400-05-717, 463-06-001, 451-04-034, Middelgroot-911-09-032).

Compliance with Ethical Standards

Conflict of interest Justin M. Luningham, Daniel B. McArtor, Meike Bartels, Dorret I. Boomsma, and Gitta H. Lubke declares they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all participants included in the study and/or from their parents.

Appendix 1

Under the assumption that the SLGM is the true data-generating model, summing the individual items at each time point results in a score that is a function of underlying factor loadings as well as the proposed growth in the factor. As an example, consider sum scores of 4 time points. A more explicit formulation of Eq. (9) is then,

$$\begin{bmatrix} \sum_{j=1}^{p} E[y_{ij1}] \\ \sum_{j=1}^{p} E[y_{ij2}] \\ \sum_{j=1}^{p} E[y_{ij3}] \\ \sum_{j=1}^{p} E[y_{ij4}] \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{p} \lambda_{j1} & 0 & 0 & 0 \\ 0 & \sum_{j=1}^{p} \lambda_{j2} & 0 & 0 \\ 0 & 0 & \sum_{j=1}^{p} \lambda_{j3} & 0 \\ 0 & 0 & 0 & \sum_{j=1}^{p} \lambda_{j4} \end{bmatrix} \boldsymbol{\Gamma}\mu \tag{38}$$

In practice, fitting a univariate LGM to the sum score results in,

$$\mathbf{SS}_i = \boldsymbol{\Gamma}\mu_{\mathbf{s}} + \varepsilon_{\mathbf{s}i} \qquad \text{(14, repeated)}$$

where $\mu_{\mathbf{s}}$ is inflated relative to μ as a function of the summed factor loadings and $\varepsilon_{\mathbf{s}i}$ is the sum of p item residuals at each measurement occasion.

Appendix 2

Bias derivations

Bias is first evaluated in the means of the growth factors. Although the focus is on linear growth, these derivations could be extended to curvilinear growth. Recall that Eq. (6) gives $E[\mathbf{y}_i] = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\mu$ because the item intercepts ν are set to 0 without loss of generality and $E[\xi_i] = \mu = [\mu_\alpha \ \mu_\beta]'$. After pre-multiplying both sides of Eq. (6) with transposed loading matrices, the growth factor means can be isolated by pre-multiplying with $(\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma})^{-1}$,

$$
\begin{aligned}
&E\left[\mathbf{y}_i\right] = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\mu} \\
&\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'E\left[\mathbf{y}_i\right] = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\mu} \\
&\left(\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'E\left[\mathbf{y}_i\right] = \boldsymbol{\mu}
\end{aligned} \tag{39}
$$

As described in the text, a fitted model may include some constrained factor loading matrix $\tilde{\boldsymbol{\Lambda}}$, but $E\left[\mathbf{y}_i\right]$is invariant to the constraints of $\tilde{\boldsymbol{\Lambda}}$, giving Eq. (11). This result generalizes to any form of $\tilde{\boldsymbol{\Lambda}}$, and the bias can be computed for the comparison of any misspecified measurement model to a known population value of $\boldsymbol{\Lambda}$.

Similarly, an expression that calculates the variance of the parameter estimates from constrained measurement models can be derived. Equation (7) establishes that the variance in the SLGM model is $\boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$. To solve for $\boldsymbol{\Phi}$, pre- and post-multiply both sides by the factor and growth loading matrices as follows,

$$
\begin{aligned}
&\boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}\left(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}\right)\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \\
&\boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \\
&(\boldsymbol{\Sigma}_y - \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' - \boldsymbol{\Theta}) = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}' \\
&\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'(\boldsymbol{\Sigma}_y - \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' - \boldsymbol{\Theta})\boldsymbol{\Lambda}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma} \\
&\left(\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'(\boldsymbol{\Sigma}_y - \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' - \boldsymbol{\Theta})\boldsymbol{\Lambda}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\right)^{-1} = \boldsymbol{\Phi}
\end{aligned} \tag{40}
$$

By replacing $\boldsymbol{\Lambda}$ with the constrained loading matrix $\tilde{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Phi}}$ can be obtained,

$$
\left(\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'(\boldsymbol{\Sigma}_y - \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' - \boldsymbol{\Theta})\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\right)^{-1} = \hat{\boldsymbol{\Phi}} \tag{41}
$$

$$
\begin{aligned}
&\left(\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\left(\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} - \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' - \boldsymbol{\Theta}\right) \\
&\quad\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\right)^{-1} = \hat{\boldsymbol{\Phi}}
\end{aligned} \tag{42}
$$

$$
\hat{\boldsymbol{\Phi}} = \left(\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\left(\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\right)\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\tilde{\boldsymbol{\Lambda}}'\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}\right)^{-1} \tag{43}
$$

Again, it is important to note that the $\boldsymbol{\Lambda}$'s in the middle parentheses are not the constrained $\tilde{\boldsymbol{\Lambda}}$'s. These represent the population-level covariance structure, which needs to be assumed or fixed to evaluate the bias of a misspecified measurement model.

Now consider the form of $\tilde{\boldsymbol{\Lambda}}$ when fitting a univariate LGM to unweighted sums of items at each time point. The univariate model simply applies the linear growth function to one value at each time point, and the underlying individual item loadings λ_{jt} are no longer considered once the items are summed to form $\mathbf{y}_{\mathbf{s}_i}$. Therefore, the univariate model treats the constrained loading matrix $\tilde{\boldsymbol{\Lambda}}$ essentially as an identity matrix. This results in fitting the univariate growth model of Eq. (1) to the aggregated sums of multiple items at each time point,

$$
E[\mathbf{SS}_i] = \boldsymbol{\Gamma}\hat{\boldsymbol{\mu}}_{\mathbf{s}} \tag{14, repeated}
$$

where $\mathbf{SS}_i$ is the aforementioned vector of sum scores for individual i and $\hat{\boldsymbol{\mu}}_{\mathbf{s}}$ is a vector of intercept and slope means estimated when the univariate model is fit to the sum scores. The derivation for the estimated intercept and slope means for the sum score then has the form,

$$
\hat{\boldsymbol{\mu}}_{\mathbf{s}} = (\boldsymbol{\Gamma}'\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'E[\mathbf{SS}_i] \tag{15, repeated}
$$

with the true form of $E[\mathbf{SS}_i]$ given by Eq. (9) under the assumption that the true data-generating process is actually the SLGM for the item-level data. The estimated intercept and slope covariance matrix is similarly derived for sum scores, where $\tilde{\boldsymbol{\Lambda}}$ is treated as identity from Eq. (13) and $\boldsymbol{\Lambda}_{\mathbf{s}}$ from Eq. (9) is the population-level $\boldsymbol{\Lambda}$ for the summed items, giving,

$$
\hat{\boldsymbol{\Phi}}_{\mathbf{s}} = \left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\left(\boldsymbol{\Lambda}_{\mathbf{s}}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}_{\mathbf{s}}'\right)\boldsymbol{\Gamma}\left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1} \tag{18, repeated}
$$

Appendix 3

The measurement loading matrices that were used in the analytic demonstrations are given below:

$$
\boldsymbol{\Lambda}_1 = \begin{bmatrix}
1 & 0 & 0 & 0 \\ .7 & 0 & 0 & 0 \\ .5 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ .7 & 0 & 0 & 0 \\ .5 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\ 0 & .7 & 0 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & .7 & 0 & 0 \\ 0 & .5 & 0 & 0 \\
0 & 0 & 1 & 0 \\ 0 & 0 & .7 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & .7 & 0 \\ 0 & 0 & .5 & 0 \\
0 & 0 & 0 & 1 \\ 0 & 0 & 0 & .7 \\ 0 & 0 & 0 & .5 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & .7 \\ 0 & 0 & 0 & .5
\end{bmatrix}
\quad
\boldsymbol{\Lambda}_2 = \begin{bmatrix}
1 & 0 & 0 & 0 \\ .7 & 0 & 0 & 0 \\ .7 & 0 & 0 & 0 \\ .7 & 0 & 0 & 0 \\ .7 & 0 & 0 & 0 \\ .7 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\ 0 & .7 & 0 & 0 \\ 0 & .7 & 0 & 0 \\ 0 & .7 & 0 & 0 \\ 0 & .6 & 0 & 0 \\ 0 & .6 & 0 & 0 \\
0 & 0 & 1 & 0 \\ 0 & 0 & .7 & 0 \\ 0 & 0 & .7 & 0 \\ 0 & 0 & .7 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & .3 & 0 \\
0 & 0 & 0 & 1 \\ 0 & 0 & 0 & .7 \\ 0 & 0 & 0 & .7 \\ 0 & 0 & 0 & .7 \\ 0 & 0 & 0 & .5 \\ 0 & 0 & 0 & .3
\end{bmatrix}
\quad
\boldsymbol{\Lambda}_3 = \begin{bmatrix}
1 & 0 & 0 & 0 \\ .7 & 0 & 0 & 0 \\ .7 & 0 & 0 & 0 \\ .3 & 0 & 0 & 0 \\ .3 & 0 & 0 & 0 \\ .5 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & .5 & 0 & 0 \\ 0 & .5 & 0 & 0 \\
0 & 0 & 1 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & .5 & 0 \\
0 & 0 & 0 & 1 \\ 0 & 0 & 0 & .3 \\ 0 & 0 & 0 & .3 \\ 0 & 0 & 0 & .7 \\ 0 & 0 & 0 & .7 \\ 0 & 0 & 0 & .5
\end{bmatrix}
$$

These correspond to the description of the three scenarios described in the text.

Appendix 4

```
TITLE:          ACE SS sim MONTE CARLO cond 1
        Model building step 5
        Decompose growth params of a sum score model

DATA: FILE = ACE_growth_sim_all.txt;
TYPE = MONTECARLO;

VARIABLE:
NAMES ARE
 zyg     ai_t1    ci_t1    ei_t1    as_t1    cs_t1    es_t1    ai_t2
 ci_t2    ei_t2    as_t2    cs_t2    es_t2     i_t1     s_t1     i_t2
 s_t2     m1_t1    m2_t1    m3_t1    m4_t1    m1_t2    m2_t2    m3_t2
 m4_t2    y1_m1_t1 y2_m1_t1 y3_m1_t1 y4_m1_t1 y5_m1_t1 y6_m1_t1 y1_m2_t1
 y2_m2_t1 y3_m2_t1 y4_m2_t1 y5_m2_t1 y6_m2_t1 y1_m3_t1 y2_m3_t1 y3_m3_t1
 y4_m3_t1 y5_m3_t1 y6_m3_t1 y1_m4_t1 y2_m4_t1 y3_m4_t1 y4_m4_t1 y5_m4_t1
 y6_m4_t1 y1_m1_t2 y2_m1_t2 y3_m1_t2 y4_m1_t2 y5_m1_t2 y6_m1_t2 y1_m2_t2
 y2_m2_t2 y3_m2_t2 y4_m2_t2 y5_m2_t2 y6_m2_t2 y1_m3_t2 y2_m3_t2 y3_m3_t2
 y4_m3_t2 y5_m3_t2 y6_m3_t2 y1_m4_t2 y2_m4_t2 y3_m4_t2 y4_m4_t2 y5_m4_t2
 y6_m4_t2
 ss1_t1 ss2_t1 ss3_t1 ss4_t1 ss1_t2 ss2_t2 ss3_t2 ss4_t2
 !names: t1 indicates twin 1, t2 indicates twin 2
 !                m1 indicates measurement factor at time 1, m1_t1: eta1 for twin 1
 !      y1_m1_t1 = item 1 at time 1 for twin 1, eg y3_m2_t2 = itm 3 at time 2 on twin 2
;
USEVARIABLES =
 zyg
 ss1_t1 ss2_t1 ss3_t1 ss4_t1 ss1_t2 ss2_t2 ss3_t2 ss4_t2
 ;

GROUP = zyg(1=mz 2=dz);

 MODEL:
! sum score measurement model

 !Factor Growth model
        i_t1 s_t1 | ss1_t1@0 ss2_t1@1 ss3_t1@2 ss4_t1@3;
        i_t2 s_t2 | ss1_t2@0 ss2_t2@1 ss3_t2@2 ss4_t2@3;

        !SS variance equal across twins
        ss1_t1-ss4_t1 (sig1-sig4);
        ss1_t2-ss4_t2 (sig1-sig4);

!!!covariance structure

Ai1-Es2@1;
Ai1 WITH Ai2@1;
Ci1 WITH Ci2@1     ;
Ei1 WITH Ei2@0 ;

As1 WITH As2@1;
Cs1 WITH Cs2@1     ;
Es1 WITH Es2@0 ;
```

```
Ai1 WITH As1*.5 (ra);
Ai2 WITH As2*.5 (ra);
Ai1 WITH As2*.5 (ra);
As1 WITH Ai2*.5 (ra);

Ci1 WITH Cs1*.5 (rc);
Ci2 WITH Cs2*.5 (rc);
Ci1 WITH Cs2*.5 (rc);
Ci2 WITH Cs1*.5 (rc);

Ei1 WITH Es1*.5 (re);
Ei2 WITH Es2*.5 (re);

Ai1 WITH Ci1-Es2@0;
Ai2 WITH Ci1-ES2@0;
As1 WITH Ci1-Es2@0;
As2 WITH Ci1-Es2@0;
Ei1 WITH Ci1-Cs2@0;
Ei2 WITH Ci1-Cs2@0;
Es1 WITH Ci1-Cs2@0;
Es2 WITH Ci1-Cs2@0;
Ei1 WITH Ei2-Es2@0;
Es1 WITH Ei2-Es2@0;
i_t1-s_t2@0;

!!!means
        [Ai1-Es2@0];
        [i_t1*1 i_t2*1] (mui);
        [s_t1*1 s_t2*1] (mus);
        [ss1_t1-ss4_t2@0];

        !ACE decomposition if I and S
! all A's together, C's together etc. to facilitate setting cov's at zero
Ai1 BY i_t1*.7071 (ai);
Ai2 BY i_t2*.7071 (ai);
As1 BY s_t1*.7071 (as);
As2 BY s_t2*.7071 (as);

Ci1 BY i_t1*.5 (ci);
Ci2 BY i_t2*.5 (ci);
Cs1 BY s_t1*.5 (cs);
Cs2 BY s_t2*.5 (cs);

Ei1 BY i_t1*.5 (ei);
Ei2 BY i_t2*.5 (ei);
Es1 BY s_t1*.5 (es);
Es2 BY s_t2*.5 (es);

!need to specify that means are same across group
!Grouping
                MODEL dz:
 !Factor Growth model
        i_t1 s_t1 | ss1_t1@0 ss2_t1@1 ss3_t1@2 ss4_t1@3;
        i_t2 s_t2 | ss1_t2@0 ss2_t2@1 ss3_t2@2 ss4_t2@3;

        !SS variance equal across twins
```

```
        ss1_t1-ss4_t1 (sig1-sig4);
        ss1_t2-ss4_t2 (sig1-sig4);

Ai1 BY i_t1*.7071 (ai);
Ai2 BY i_t2*.7071 (ai);
As1 BY s_t1*.7071 (as);
As2 BY s_t2*.7071 (as);

Ci1 BY i_t1*.5 (ci);
Ci2 BY i_t2*.5 (ci);
Cs1 BY s_t1*.5 (cs);
Cs2 BY s_t2*.5 (cs);

Ei1 BY i_t1*.5 (ei);
Ei2 BY i_t2*.5 (ei);
Es1 BY s_t1*.5 (es);
Es2 BY s_t2*.5 (es);

!!!covariance structure

Ai1-Es2@1;
Ai1 WITH Ai2@.5;
Ci1 WITH Ci2@1     ;
Ei1 WITH Ei2@0 ;

As1 WITH As2@.5;
Cs1 WITH Cs2@1     ;
Es1 WITH Es2@0 ;

Ai1 WITH As1*.5 (ra);
Ai2 WITH As2*.5 (ra);
Ai1 WITH As2*.25 (dzra);
As1 WITH Ai2*.25 (dzra);

Ci1 WITH Cs1*.5 (rc);
Ci2 WITH Cs2*.5 (rc);
Ci1 WITH Cs2*.5 (rc);
Ci2 WITH Cs1*.5 (rc);

Ei1 WITH Es1*.5 (re);
Ei2 WITH Es2*.5 (re);

Ai1 WITH Ci1-Es2@0;
Ai2 WITH Ci1-ES2@0;
As1 WITH Ci1-Es2@0;
As2 WITH Ci1-Es2@0;

 Ei1 WITH Ci1-Cs2@0;
 Ei2 WITH Ci1-Cs2@0;
 Es1 WITH Ci1-Cs2@0;
 Es2 WITH Ci1-Cs2@0;
 Ei1 WITH Ei2-Es2@0;
 Es1 WITH Ei2-Es2@0;
 i_t1-s_t2@0;

!!!means
[Ai1-Es2@0];
[i_t1*1 i_t2*1] (mui);
[s_t1*1 s_t2*1] (mus);
[ss1_t1-ss4_t2@0];

        MODEL mz:
 !Factor Growth model
        i_t1 s_t1 | ss1_t1@0 ss2_t1@1 ss3_t1@2 ss4_t1@3;
        i_t2 s_t2 | ss1_t2@0 ss2_t2@1 ss3_t2@2 ss4_t2@3;

        !SS variance equal across twins
        ss1_t1-ss4_t1 (sig1-sig4);
        ss1_t2-ss4_t2 (sig1-sig4);

!ACE decomposition of I and S
! all A's together, C's together etc. to facilitate setting cov's at zero
Ai1 BY i_t1*.7071 (ai);
Ai2 BY i_t2*.7071 (ai);
As1 BY s_t1*.7071 (as);
As2 BY s_t2*.7071 (as);

Ci1 BY i_t1*.5 (ci);
Ci2 BY i_t2*.5 (ci);
Cs1 BY s_t1*.5 (cs);
Cs2 BY s_t2*.5 (cs);

Ei1 BY i_t1*.5 (ei);
Ei2 BY i_t2*.5 (ei);
Es1 BY s_t1*.5 (es);
Es2 BY s_t2*.5 (es);

!!!covariance structure

Ai1-Es2@1;
Ai1 WITH Ai2@1;
Ci1 WITH Ci2@1     ;
Ei1 WITH Ei2@0 ;

As1 WITH As2@1;
Cs1 WITH Cs2@1     ;
Es1 WITH Es2@0 ;

Ai1 WITH As1*.5 (ra);
Ai2 WITH As2*.5 (ra);
Ai1 WITH As2*.5 (ra);
As1 WITH Ai2*.5 (ra);

Ci1 WITH Cs1*.5 (rc);
Ci2 WITH Cs2*.5 (rc);
Ci1 WITH Cs2*.5 (rc);
Ci2 WITH Cs1*.5 (rc);

 Ei1 WITH Es1*.5 (re);
 Ei2 WITH Es2*.5 (re);
```

```
Ai1 WITH Ci1-Es2@0;
Ai2 WITH Ci1-ES2@0;
As1 WITH Ci1-Es2@0;
As2 WITH Ci1-Es2@0;
Ei1 WITH Ci1-Cs2@0;
Ei2 WITH Ci1-Cs2@0;
Es1 WITH Ci1-Cs2@0;
Es2 WITH Ci1-Cs2@0;
Ei1 WITH Ei2-Es2@0;
Es1 WITH Ei2-Es2@0;
i_t1-s_t2@0;

!!!means
[Ai1-Es2@0];
[i_t1*1 i_t2*1] (mui);
[s_t1*1 s_t2*1] (mus);
[ss1_t1-ss4_t2@0];

MODEL CONSTRAINT:
NEW(hi2 hs2 ci2 cs2 ei2 es2);
hi2 = ai**2/(ai**2 + ci**2 + ei**2);
ci2 = ci**2/(ai**2 + ci**2 + ei**2);
ei2 = ei**2/(ai**2 + ci**2 + ei**2);
hs2 = as**2/(as**2 + cs**2 + es**2);
cs2 = cs**2/(as**2 + cs**2 + es**2);
es2 = es**2/(as**2 + cs**2 + es**2);
dzra = ra*.5;
 OUTPUT: TECH1 ;
```

References

Achenbach TM, Rescorla LA (2000) Manual for the ASEBA preschool forms & profiles. University of Vermont, Research Center for Children, Youth, & Families. Burlington

Achenbach TM, Rescorla LA (2001) Manual for the ASEBA school-age forms & profiles. University of Vermont, Research Center for Children, Youth, & Families. Burlington

Agresti A (2002) Categorical data analysis, 2nd edn, Wiley ser in probability and statistics, Hoboken

Bartels M, Hudziak JJ, van den Oord EJCG, van Beijsterveldt, CEM, Rietveld MJH, Boomsma DI (2003) Co-occurrence of aggressive behavior and rule-breaking behavior at age 12: Multi-rater analyses. Behav Genet 33(5):607–621

Bollen KA (1989) Structural equations with latent variables. Wiley, New York

Boomsma DI. (2015) Aggression in Children: Unravelling the interplay of genes and environment through (epi)genetics and metabolomics. J Ped Neo Ind Med. 4(2):e040251)

Boomsma DI, Molenaar PC (1987) The genetic analysis of repeated measures: I. simplex models. Behav Genet 17(2):111–123

Burt SA (2009) Are there meaningful etiological differences within antisocial behavior? results of a meta-analysis. Clin Psychol Rev 29(2):163–178

Byrne BM, Shavelson RJ, Muthén B (1989) Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychol Bull 105(3):456–466

Carmines EG, Zeller RA (1979) Reliability and validity assessment. Sage Publications, Newbury Park

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334

Cui L, Colasante T, Malti T, Ribeaud D, Eisner MP (2016) Dual trajectories of reactive and proactive aggression from mid-childhood to early adolescence: Relations to sensation seeking, risk taking, and moral reasoning. J Abnorm Child Psychol 44(4):663–675

Edwards MC, Wirth RJ (2009) Measurement and the study of change. Res Hum Dev 6(2–3):74–96

Edwards MC, Wirth RJ (2012) Valid measurement without factorial invariance: A longitudinal example. In: Harring JR, Hancock GR (eds) Advances in longitudinal methods in the social and behavioral sciences. IAP Information Age Publishing, Charlotte, pp 289–311

Eley TC, Lichenstein P, Stevenson J (1999) Sex differences in the etiology of aggressive and nonaggressive antisocial behavior: results from two twin studies. Child Dev 70(1):155–168

Ferrer E, Balluerka N, Widaman KF (2008) Factorial invariance and the specification of second-order latent growth models. Methodology 4(1):22–36.

Finkel D, Davis DW, Turkheimer E, Dickens WT (2015) Applying biometric growth curve models to developmental synchronies in cognitive development: The Louisville twin study. Behav Genet 45(6):600–609

Forsman M, Lichtenstein P, Andershed H, Larsson H (2010) A longitudinal twin study of the direction of effects between psychopathic personality and antisocial behaviour. J Child Psychology. Psychiatry 51(1):39–47

George D, Mallery P (2005) SPSS for windows step by step: a simple guide and reference 12.0 update. 5th edn. Pearson Education New Zealand, Auckland

Goodman R (1997) The strengths and difficulties questionnaire: a research note. Child Psychol Psychiatry Allied Discip 38(5):581–586

Grimm KJ, Kuhl AP, Zhang Z (2013) Measurement models, estimation, and the study of change. Struct Equat Model 20(3):504–517

Hancock GR, Kuo W, Lawrence FR (2001) An illustration of second-order latent growth models. Struct Equat Model 8(3):470–489

Hudziak JJ, van Beijsterveldt CEM, Bartels M, Rietveld MJH, Rettew DC, Derks EM, Boomsma DI (2003) Individual differences in aggression: genetic analyses by age, gender, and informant in 3-, 7-, and 10-year-old dutch twins. Behav Genet 33(5):575–589

Kan K, Dolan CV, Nivard MG, Middeldorp CM, van Beijsterveldt CEM, Willemsen G, Boomsma DI (2013) Genetic and environmental stability in attention problems across the lifespan: evidence from the netherlands twin register. J Am Acad Child Adoles Psychiatry 52(1):12–25

Leite WL (2007) A comparison of latent growth models for constructs measured by multiple items. Struct Equat Model 14(4):581–610

Lubke GH, Neale MC, Dolan CV (2004) Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction. Twin Res 7:292–298

Lubke GH, Miller PJ, Verhulst B, Bartels M, van Beijsterveldt CEM, Willemsen G, Boomsma DI, Middeldorp CM (2016) A powerful phenotype for gene-finding studies derived from trajectory analyses of symptoms of anxiety and depression between age seven and 18. Am J Med Genet B: Neuropsychiatric Genet 171(7):948–957

Lubke GH, McArtor DB, Boomsma DI, Bartels M (in press) Genetic and environmental contributions to the development of childhood aggression. Dev Psych

McArdle JJ (1986) Latent variable growth within behavior genetic models. Behav Genet 16(1):163–200

McArdle JJ (1988) Dynamic but structural equation modeling of repeated measures data. In: Nesselroade JR, Cattell RB (eds) 2nd edn. Handbook of multivariate experimental psychology. Plenum Press, New York, pp 561–614

McArdle JJ, Hamagami F (2003) Structural equation models for evaluating dynamic concepts within longitudinal twin analyses. Behav Genet 33(2):137–159

Meredith W (1993) Measurement invariance, factor analysis and factorial invariance. Psychometrika 58(4):525–543

Meredith W, Horn J (2001) The role of factorial invariance in modeling growth and change. In: Collins LM, Sayer AG (eds) New methods for the analysis of change. American Psychological Association, Washington, pp 203–240

Molenaar D, Dolan CV (2014) Testing systematic genotype by environment interactions using item level data. Behav Genet 44(3):212–231

Muthén LK, Muthén BO (1998–2015) Mplus user's guide. Seventh Edition. Muthén & Muthén, Los Angeles

Neale MC, Maes, HHM (2004) Methodology for genetic studies of twins and families. Kluwer Academic/Plenum Publishers, New York. Available at http://ibgwww.colorado.edu/workshop2006/cdrom/HTML/book2004a.pdf

Neale MC, McArdle JJ (2000) Structured latent growth curves for twin data. Twin Res 3(3): 165–177.

Neale MC, Lubke G, Aggen SH, Dolan CV (2005) Problems with using sum scores for estimating variance components: Contamination and measurement noninvariance. Twin Res Hum Genet 8(6):553–568

NICHD Early Child Care Research Network (2004) Trajectories of physical aggression from toddlerhood to middle childhood. Monogr Soc Res Child Dev 69(4):102–119

Porsch RM, Middeldorp CM, Cherny SS, Krapohl E, van Beijsterveldt CEM, Loukola A, Korhonen T, Pulkkinen L, Corley R, Rhee S et al (2016) Longitudinal heritability of childhood aggression. Am J Med Genet B 171(5):697–707

Prescott CA, Kendler KS (1996) Longitudinal stability and change in alcohol consumption among female twins: contributions of genetics. Dev Psychopathol 8(4):849–866

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL http://www.R-project.org/

Rosseel Y (2012). lavaan: An R Package for Structural Equation Modeling. J Stat Software, 48(2):1–36. URL: http://www.jstatsoft.org/v48/i02/

Rutter M, Sroufe LA (2000) Developmental psychopathology: concepts and challenges. Dev Psychopathol 12(3):265–296

Sayer AG, Cumsille PE (2001) Second-order latent growth models. In: Collins LM, Sayer AG (eds) New methods for the analysis of change. American Psychological Association, Washington, pp 179–200

Schwabe I, van den Berg SM (2014) Assessing genotype by environment interaction in case of heterogeneous measurement error. Behav Genet 44(4):394–406. doi:10.1007/s10519-014-9649-7

Singer JD, Willett JB (2003) Applied longitudinal data analysis: Modeling change and event occurrence. Oxford University Press, New York

Slof-Op't Landt MC, van Furth EF, Rebollo-Mesa I, Bartels M, van Beijsterveldt CEM, Slagboom PE, Boomsma DI, Meulenbelt I, Dolan CV (2009) Sex differences in sum scores may be hard to interpret: the importance of measurement invariance. Assessment 16(4):415–423

Tremblay RE (2003) Why socialization fails: The case of chronic physical aggression. In: Lahey BB, Moffitt TE, Caspi A (eds) Causes of conduct disorder and juvenile delinquency. Guilford Press, New York, pp 182–224

van Beijsterveldt CEM, Bartels M, Hudziak JJ, Boomsma DI (2003) Causes of stability of aggression from early childhood to adolescence: a longitudinal genetic analysis in dutch twins. Behav Genet 33(5):591–605

van Beijsterveldt, CEM, Groen-Blokhuis M, Hottenga JJ, Franić S, Hudziak JJ, Lamb D, Huppertz C, de Zeeuw E, Nivard M, Schutte N et al (2013) The young netherlands twin register (YNTR): longitudinal twin and family studies in over 70,000 children. Twin Res Hum Genet 16(1):252–267

van den Berg SM, Glas CAW, Boomsma DI (2007) Variance decomposition using an IRT measurement model. Behav Genet 37(4):604–616

van der Valk JC, Verhulst FC, Neale MC, Boomsma DI (1998) Longitudinal genetic analysis of problem behaviors in biologically related and unrelated adoptees. Behav Genet 28(5):365–380

Wang P, Niv S, Tuvblad C, Raine A, Baker LA (2013) The genetic and environmental overlap between aggressive and non-aggressive antisocial behavior in children and adolescents using the self-report delinquency interview (SR-DI). J Crim Justice 41(5):277–284

Widaman KF, Reise SP (1997) Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. editors. In: Bryant KJ, Windle M, West SG (eds) The science of prevention: Methodological advances from alcohol and substance abuse research. American Psychological Association, Washington, pp 281–324

Widaman KF, Ferrer E, Conger RD (2010) Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. Child Development. Perspectives 4(1):10–18

Wirth RJ. 2009. The effects of measurement non-invariance on parameter estimation in latent growth models. (Order No. AAI3331053)