ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-193

Estimating Student Growth on Psychological and Social-emotional Constructs: A Comparison of Multiple Scoring Approaches

Megan Kuhfeld
NWEA

James Soland
University of Virginia

A huge portion of what we know about how humans develop, learn, behave, and interact is based on survey data. Researchers use longitudinal growth modeling to understand the development of students on psychological and social-emotional learning constructs across elementary and middle school. In these designs, students are typically administered a consistent set of self-report survey items across multiple school years, and growth is measured either based on sum scores or scale scores produced based on item response theory (IRT) methods. While there is great deal of guidance on scaling and linking IRT-based large-scale educational assessment to facilitate the estimation of examinee growth, little of this expertise is brought to bear in the scaling of psychological and social-emotional constructs. Through a series of simulation and empirical studies, we produce scores in a single-cohort repeated measure design using sum scores as well as multiple IRT approaches and compare the recovery of growth estimates from longitudinal growth models using each set of scores. Results indicate that using scores from multidimensional IRT approaches that account for latent variable covariances over time in growth models leads to better recovery of growth parameters relative to models using sum scores and other IRT approaches.

VERSION: January 2020

Suggested citation: Kuhfeld, Megan, and James Soland. (2020). Estimating Student Growth on Psychological and Social-emotional Constructs: A Comparison of Multiple Scoring Approaches. (EdWorkingPaper: 20-193). Retrieved from Annenberg Institute at Brown University: https://www.edworkingpapers.com/ai20-193

Estimating Student Growth on Psychological and Social-emotional Constructs:

A Comparison of Multiple Scoring Approaches

Megan Kuhfeld
Research Scientist, NWEA

James Soland
Assistant Professor, University of Virginia

January 3, 2020

CONTACT INFORMATION:
Megan Kuhfeld
121 N.W. Everett Street
Portland, OR 97209
Ph. 503-548-5295
megan.kuhfeld@nwea.org

Abstract

A huge portion of what we know about how humans develop, learn, behave, and interact is based on survey data. Researchers use longitudinal growth modeling to understand the development of students on psychological and social-emotional learning constructs across elementary and middle school. In these designs, students are typically administered a consistent set of self-report survey items across multiple school years, and growth is measured either based on sum scores or scale scores produced based on item response theory (IRT) methods. While there is great deal of guidance on scaling and linking IRT-based large-scale educational assessment to facilitate the estimation of examinee growth, little of this expertise is brought to bear in the scaling of psychological and social-emotional constructs. Through a series of simulation and empirical studies, we produce scores in a single-cohort repeated measure design using sum scores as well as multiple IRT approaches and compare the recovery of growth estimates from longitudinal growth models using each set of scores. Results indicate that using scores from multidimensional IRT approaches that account for latent variable covariances over time in growth models leads to better recovery of growth parameters relative to models using sum scores and other IRT approaches.

Keywords: psychological development, social-emotional learning, developmental trajectories, growth modeling, multidimensional item response theory (MIRT).

Estimating Student Growth on Psychological and Social-emotional Constructs: A Comparison of Multiple Scoring Approaches

Oftentimes, psychologists and researchers from related disciplines are interested not only in a person's position on a latent psychological construct at a point in time, but also that individual's growth on the construct over time. Particularly among developmental psychologists, the trajectories of children on latent psychological and social-emotional[1] constructs like self-efficacy, motivation, and self-concept that are not directly observable represent the primary topic of inquiry. Such trajectories provide insight into what normative developmental patterns look like, including whether children may not be developing at a typical pace. Further, growth on latent constructs is often used to evaluate programs and interventions to see if those actions change trajectories for the better. In short, a huge swath of the psychological, educational, and broader social science literature is devoted to understanding how people grow on constructs over time.

Since these psychological constructs of interest cannot be directly observed, they are often measured by a set of survey items administered at multiple time points. After the item responses are collected, there are two central approaches currently used to examine children's developmental trajectories. The first follows a two-step procedure, where (a) scores are produced to quantify a child's trait level at each time point, and (b) these scores are used in statistical modeling procedures such as repeated measures analysis of variance (ANOVA), multilevel models, or latent growth curve models (LGCMs). Said differently, growth is estimated based on a single composite score (e.g., total sum score or estimated factor score) that is used to represent the construct at a given timepoint, rather than the item responses used to estimate those scores.

The second approach is to use structural equation modeling (SEM) with a measurement submodel (Bollen & Curran, 2006). This approach is used to directly model the latent construct at each timepoint using observed item responses and estimate growth based on those time-specific latent variables in a single modeling framework. Oftentimes, this approach is referred to as "curve of factors" or as second-order growth modeling (Hancock, Kuo, & Lawrence, 2001; McArdle, 1988). Research has demonstrated many advantages of the second-order growth model over the two-stage approach in the particular case when observed (sum or mean) scores are used for the latter (Geiser, Keller, Lockhart, 2013; Bishop, Geiser, & Cole, 2015).

However, as reviewed and investigated by Bauer and Curran (2015), simultaneously modeling measurement and estimating growth is frequently hampered by computational and practical concerns, including the large samples needed to obtain stable parameter estimates. Perhaps as a result, in a review conducted by Isiordia and Ferrer (2018), of 100 articles on growth in the Education Resources Information Center (ERIC) database from 2005 to 2015, none used a second-order growth model. In our own brief review of the literature on psychological and social-emotional development, we found only a handful of studies using an approach comparable to a second-order growth model (e.g., Caprara et al., 2008; Caprara, Vecchione, Alessandri, Gerbino, & Barbaranelli, 2011; Soland, Jensen, Keys, Bi, & Wolk, 2019; Soland, Kuhfeld, Wolk, & Bi, 2019).

Given that the two-step procedure is still most widely used to study children's developmental trajectories, more careful consideration is needed to understand the impacts of the scoring approach used in the first stage on our understanding of how children develop. As Bauer and Curran (2016, pg. 2) point out, "there is a lack of alignment between how outcomes are measured and the models subsequently used to analyze individual differences in stability and

change". There are many ways in which scores can be produced from a single data source, and these choices are increasingly considered to be an underemphasized cause of the well-documented replication issues in psychological studies (Fried & Flake, 2018; McNeish & Wolf, 2019). Many researchers in the psychological literature prefer to measure relevant constructs using observed scores (either sum scores or means of items) from survey instruments (Bauer & Curran, 2016). However, there are known limitations of sum scores for inferences about psychological constructs. At a single point in time, sum scores make a range of strong assumptions, including that all students have received and answered the same item set, all items are equally difficult, and items do not display any measurement invariance (e.g., that all items are equally developmentally appropriate across age groups and across gender/race/language levels). However, most research that uses sum/mean scores to study psychological constructs does not examine the veracity of these assumptions prior to using the scores (Crutzen & Peters, 2017).

Furthermore, though largely unstudied, issues with sum scores are likely compounded when used in growth models. Beyond the large (oftentimes untenable) assumptions made at a single point in time, sum scores used in longitudinal models require several additional assumptions, including that items function similarly across timepoints (e.g., items maintain the same level of difficulty and construct relevance across multiple observations). Research has shown that such failures of longitudinal measurement invariance can change fundamental inferences about developmental trajectories (e.g., Widaman, Ferrer, & Conger, 2010; Willoughby, Wirth, & Blair, 2012). Thus, much of what we know about growth on latent constructs is, at best, likely impacted by strong assumptions implicit in sum scores.

Given the limitations of sum score approaches, there are a bevy of item response theory (IRT) and other latent variable models that have been employed in the context of educational measurement (particularly large-scale assessments). Further, a set of multidimensional IRT-based methods have been developed specifically to scale tests to appropriately account for multiple timepoints (e.g., Koran, 2009; Paek, Li, & Park, 2016). However, these longitudinal measurement models are rarely used in the context of measuring students' psychological and social-emotional development. Furthermore, little evidence exists on which latent variable approach does the best job not only of recovering true scores, but also recovering true growth parameters when the scores are used in a latent growth model. That is, if estimates of true population-level slope parameters are the estimand of interest, which scoring approach best reproduces those parameters? Despite the emergence of multiple model-based alternatives to scoring, little is known about which does a better job of recovering growth parameters (nor do we know how much better these models perform relative to using observed scores in growth models).

Our study involves simulation and empirical studies to investigate how to optimally score longitudinal survey data if the parameters of interest are students' latent growth trajectories. In the simulation study, we generate item response data under a number of different conditions, produce scores using sum scores and multiple IRT approaches, and compare those scores using latent growth models. Through this approach, we can determine how effectively using the observed scores in growth models recovers true trajectories under a number of conditions, and see whether IRT-based models perform better, including relative to each other. Perhaps more importantly, we can determine which IRT approach best recovers the true scores and growth parameters so that researchers interested in psychological development have guidelines on which

scoring strategy to use. The empirical study compares the various scoring approaches using a widely used survey measure of growth mindset.

Background

Approaches for Calibrating and Scoring Multi-Timepoint Surveys

Before we can study students' developmental trajectories, we must (a) administer a survey measure at multiple timepoints, and then (b) scale and score the measure based on the item responses obtained (a process referred to as *calibration*). One of the most common scoring approaches used in psychology and education for survey measures is to simply add up the item responses. While this method works in some situations when a set of strong assumptions hold (Bauer & Curran, 2016), it also has known limitations. As McNeish and Wolf (2019) point out, using a sum score is mathematically equivalent to fitting a measurement model with strong assumptions about the items and their reliability. For example, in a traditional SEM framework, such a model assumes that all items are equally related to the construct of interest (e.g., all the loadings are one) and that the error variances for all the items are equal. Further, a sum score model fails to account for potential differences in the severity (or in IRT parlance the "difficulty") of the items. For example, sum scores from a math test would place equal weight on the items "What is 2+2?" and "What is the square root of 325?" even though these items vary greatly with regards to their difficulty. Sum scores additionally assume that the measure functions identically across groups (e.g., boys do not have a higher probability than girls of endorsing an item, controlling for the underlying latent trait) as well as across time, an assumption that is frequently violated in longitudinal studies that span developmental periods (Millsap, 2012). Given these assumptions, sum scores tend to be much less reliable, which has

consequences for uses of related scales, including for how students are classified on the basis of a scale (McNeish & Wolf, 2019).

There are a number of IRT-based alternatives that do not have the same limitations as sum scores. Under these approaches, a measurement model is fit to the item response data and then scores are produced based on the calibrated item parameters (Embretson & Reise, 2013; Wirth & Edwards, 2007). While IRT is now widely used in educational and psychological measurement (Reise & Waller, 2009; van der Linden & Hambleton, 2013), applying this approach to produce scores in a longitudinal design can be less than straightforward. There are a number of choices that a researcher can make when calibrating/scoring a scale, including (a) the calibration sample used (e.g., a single timepoint from the data collected in a study or multiple timepoints of item response data), (b) the IRT model used to calibrate (e.g., a unidimensional or multidimensional model), and (c) the scoring approach used (e.g., maximum likelihood, expected a posteriori [EAP], and modal a posteriori [MAP] scoring). While questions of IRT model type and scoring approach have been examined in depth for scores at a single point in time (Kolen & Tong, 2010; Maydeu-Olivares, Drasgow, & Mead, 1994), choices around the calibration sample and IRT model used for longitudinal measurement are less well understood.

We now describe a set of possible calibration samples and IRT models that can be used to produce scores for growth modeling, starting with the simplest and moving towards increased complexity. All of these approaches have been used in practice, examined empirically in research, or both (Bauer & Curran, 2015). We also discuss the accompanying IRT model used to score those different calibration samples. Perhaps the most fundamental consideration under investigation is whether to use an IRT model that calibrates based on a single timepoint versus a MIRT model that includes latent variables for scores at all timepoints.

1. Cross-sectional IRT Calibration, Unidimensional IRT Model. The first approach we considered is to calibrate the measure using cross-sectional data from a single time point. That is, even though students have item responses from multiple timepoints, only responses from the first timepoint are used. Such an approach might be taken in the event that parameters are calibrated before subsequent waves of data are collected, and those parameter values continue to be used to score subsequent waves. Depending on the data collection design, scores from just a single age group or multiple age groups at Time 1 could be used. Once the item parameters are obtained using a unidimensional IRT model, the parameters are treated as fixed for the later waves and used to score the item responses in the remaining timepoints. Figure 1(a) presents the path diagram for the unidimensional cross-sectional IRT model.

As described below, in our own study, we calibrated item parameters based on scores from just a single age and timepoint (what we refer to as a "restricted age" calibration). In a hypothetical scenario where there are four years of data, Table 1 shows a set of possible combinations of students that can be used for calibration. The top panel of Table 1 shows the respondents used in the "restricted age" calibration, where item parameters were estimated using only students in 5^{th} grade at Time 1. If the study design allows for the collection of data from multiple ages at the first timepoint, an alternative cross-sectional IRT approach could be used. This cross-sectional "range of ages" design combines the item responses from students across multiple grade levels/age cohorts observed during the first timepoint (middle panel of Table 1). However, since this "range of ages" design is less common than a single cohort design in longitudinal research, we have chosen to examine only the "restricted age" calibration condition.[2]

SCORING LONGITUDINAL SURVEY DATA

For Likert-type items, item calibration and scoring can be accomplished using the graded response model or GRM (Samejima, 1969). Let there be $j = 1, \dots, n$ items and $i = 1, \dots, N$ individuals. Let the response from individual i to item j at timepoint t be y_{tij}, where y_{tij} has K response categories. It can be assumed that y_{tij} takes integer values from $(0, \dots, K-1)$. Let the cumulative category response probabilities be

$$P(y_{tij} \geq 1 | \theta_i) = \frac{1}{1 + \exp[-(c_{j1} + a_j \theta_i)]}$$

$$\vdots \qquad (1)$$

$$P(y_{tij} \geq K-1 | \theta_i) = \frac{1}{1 + \exp[-(c_{j,K-1} + a_j \theta_i)]}$$

The category response probability is the difference between two adjacent cumulative probabilities

$$P(y_{tij} = k | \theta_i) = P(y_{tij} \geq k | \theta_i) - P(y_{tij} \geq k+1 | \theta_i), \qquad (2)$$

where $\mathrm{P}(y_{tij} \geq 0 | \theta_i)$ is equal to 1 and $\mathrm{P}(y_{tij} \geq K | \theta_i)$ is zero. The item parameter a_j is the slope parameter describing the relationship between item j and the latent factor and $b_{\mathrm{j}1}, \dots, b_{j,K-1}$ are a set of $K-1$ (strictly ordered) parameters. The thresholds denote the point on the latent variable separating category k from category $k+1$.

In the unidimensional case, the logit in Equation 1 can be re-expressed in a more convenient slope-threshold form as $c_{jk} + a_j \theta_i = a_j(\theta_i - b_{jk})$, where $b_{jk} = -c_{jk}/a_j$ is the threshold (also referred to as severity or difficulty) parameter for category k. The kth threshold denotes the point on the latent variable separating category k from category $k+1$. However, the slope-threshold form does not generalize well to multidimensional models, so we adopt the slope–intercept parameterization here and for all remaining IRT models presented.

There are multiple possible limitations in using just a single timepoint to calibrate scores for longitudinal research. First, in the single age group design, we are assuming that the items function similarly in the age group studied and in all of the other ages at which students may be assessed. Second, in both approaches, the calibration approach contains only one observation per person and does not provide any information about whether the construct of interest varies within persons across time. When, for example, EAP or MAP scoring is used, we assume a single mean. Thus, in a scenario with true population growth, the score estimates are shrunken (and therefore biased) towards a mean that assumes no change across time.

2. Longitudinal IRT Calibration, Unidimensional IRT Model. The second possible approach is to calibrate the item response data using all available timepoints from a given cohort in a single unidimensional model. That is, item responses across different timepoints from a single individual are treated as independent observations in a single (long) data file, and a unidimensional model is estimated based on the pooled (across-years) item responses. Returning to bottom panel of Table 1, item responses within each timepoint from students in the 5th grade cohort would be included. The path diagram for this model is shown in Figure 1(b). One advantage of this approach is that it makes use of all of the available data for a cohort. The downsides of the approach are that (a) the longitudinal data are treated as coming from a single normally distributed population rather than freely estimating the latent means/variances of the later timepoints separately, (b) since we assume responses across time for an individual are independent, the serial correlation due to observing the same respondents across timepoints is not modeled directly, and (c) measurement invariance of all of the item parameters is assumed but not directly testable.

3. Longitudinal IRT Calibration, Multidimensional IRT Model. The third approach uses a multidimensional item response theory (MIRT) model to estimate latent change across time in an IRT framework. Item response data from each timepoint are combined and calibrated simultaneously across the *T* timepoints. The path diagram for this model is shown in Figure 1(c). As shown in Figure 1(c), items are calibrated for Cohort 1 such that each grade/timepoint has its own latent variable estimate. We use a multidimensional extension of the GRM. Let the cumulative category response probabilities be

$$P\left(y_{tij} \geq 1 \middle| \boldsymbol{\theta_i}\right) = \frac{1}{1 + \exp[-(c_{j1} + \boldsymbol{a}_j'\boldsymbol{\theta_i})]}$$

$$\vdots \qquad (3)$$

$$P\left(y_{tij} \geq K - 1 \middle| \boldsymbol{\theta_i}\right) = \frac{1}{1 + \exp[-(c_{j,K-1} + \boldsymbol{a}_j'\boldsymbol{\theta_i})]}$$

As with the unidimensional model, the category response probability is the difference between two adjacent cumulative probabilities. The difference between the unidimensional and multidimensional GRM is that $\boldsymbol{\theta_i}$ is now a $T \times 1$ vector of latent traits and $\boldsymbol{a}_j'$ a vector of slope parameters.

Given the same items are repeated across time points, we include a set of equality constraints for the item parameters of the repeated items. Let $\varphi_{tj}=\{\boldsymbol{a}_{tj}', c_{tj1}, \ldots, c_{tj,K-1}\}$ be the vector of item parameters for item *j* observed at the first time point ($t = 1$). We assume that $\varphi_{1j} = \varphi_{2j} = \cdots = \varphi_{Tj}$, where φ_{Tj} is the item parameter vector for item *j* measured at time T. The first latent dimension is often (but not always) constrained for identification purposes to follow a standard normal distribution $\theta_{T1} \sim N(0,1)$, and the mean, variance, and covariance of the other latent factors are freely estimated relative to the first timepoint.

Thus, unlike Approaches 1 and 2, the MIRT approach explicitly accounts for the over-time correlations in the model. While it is expected that the scores generated under Approaches 1 and 2 would have attenuated over-time correlations, this problem should be mitigated by the inclusion of this information in the MIRT model. However, this approach does not model the possible serial correlation due to each item on the survey being administered repeatedly across the timepoints (Paek et al., 2014).

4. Longitudinal IRT Calibration, Multidimensional IRT Model with Serial Correlation. The fourth approach involves an adapted version of the two-tier full-information IRT framework (Cai, 2010b, 2010a). This approach uses the same sample and latent psychological factors as in Approach 3, but adds a set of secondary dimensions that capture the serial correlation due to the same item being repeated across multiple timepoints. Each observation of item *j* (e.g., at Time 1, Time 2, etc.) loads on a single secondary (or specific) factor. In the two-tier formulation, the GRM cumulative response probabilities are

$$P\left(y_{tij} \geq 1 \middle| \boldsymbol{\theta_i}\right) = \frac{1}{1 + \exp[-(c_{j1} + \boldsymbol{a}'_{jp}\boldsymbol{\theta_{ip}} + a_{js}\theta_s)]} \tag{4}$$

$$\vdots$$

$$P\left(y_{tij} \geq K - 1 \middle| \boldsymbol{\theta_i}\right) = \frac{1}{1 + \exp[-(c_{j,K-1} + \boldsymbol{a}'_{jp}\boldsymbol{\theta_{ip}} + a_{js}\theta_s)]}$$

where $\boldsymbol{a}'_{jp}$ is the $T\times 1$ vector of item slopes on the primary (p) factors and a_{js} the item slope on specific factor *s*. For model identification, we fix the distribution of the first primary latent factor $\theta_{T1}\sim N(0,1)$ and each of the specific dimensions ($\theta_s\sim N(0,1)$, for s=1,…,*S*), and free the mean and variance of the remaining primary dimensions ($t = 2, \ldots, T$). Additionally, the slope parameters of each specific factor are set equal. The path diagram for this model is shown in Figure 1(d).

Prior comparisons of scoring approaches

A handful of researchers have proposed MIRT models to scale test scores accounting for the longitudinal nature of the data and compared estimated growth parameters across models. Paek, Park, Cai, and Chi (2014) examined three different IRT approaches to estimate growth in a single-group anchor test design with dichotomous items, where the same students took two multiple-choice mathematics assessments that were linked with seven common (anchor) items. They examined three approaches: (a) separate calibrations of each timepoint's data, (b) a two-dimensional MIRT model, and (c) a nine-dimensional MIRT model containing two primary factors and seven specific factors. Examinee change scores were then compared across models. In terms of tracking individual growth, growth patterns, and model–data fit, their results demonstrated the importance of modeling serial correlation over multiple time when producing scores for use in growth models. For example, 14% of examinees at Time 1 and 36.4% at Time 2 had more than a |.2| difference (one fifth of a standard deviation on the θ scale at Time 1) between models that did and did not properly account for serial correlation.

Bauer and Curran (2015) compared different scoring methods with dichotomously-scored longitudinal data, including the cross-sectional IRT model (our first approach) and the longitudinal MIRT model (our third approach). Using a single simulated dataset with 12 dichotomous items, Bauer and Curran (2015) found that the cross-sectional IRT calibration resulted in muted age trends and underestimated the degree of individual differences in the trajectories across time, whereas the longitudinal IRT calibration resulted in better estimates of the random effect variances.

The primary limitations of this literature are that (a) most of the research has focused on scaling dichotomous items rather than Likert-type items frequently used on surveys, (b) the

shortest measure considered was nine items, whereas psychological constructs are frequently assessed with a small number of items (in fact, Flake, Pek, & Hehman [2017] found that the average scale reported in the *Journal of Personality and Social Psychology* was only 4.7 items long), (c) none have provided a comparison of the IRT approaches with sum scores, and (d) most previous studies used either an empirical example or a single simulated data file. These limitations reduce the applicability of these studies for the analysis of psychological constructs, which are frequently assessed with relatively short Likert-type survey measures that contain consistent items across time.

Study Purpose

The purpose of this study is to demonstrate the effect of calibration and scoring techniques on the results from longitudinal latent growth analyses. Specifically, we compare the performance of five scoring approaches (sum scores and the four IRT calibration/scoring approaches outlined above) that could be applied to multi-timepoint survey data in terms of recovery of the true latent growth parameters. In the first simulation study, we generate data for three timepoints assuming students' true latent scores follow a linear growth model and examine parameter recovery under various conditions. In the second simulation study, we generate data for four timepoints assuming students' true latent scores follow a quadratic growth model. In the empirical analyses, we apply all five scoring approaches to item responses from 3,266 students who were administered the CORE survey of growth mindset once a year from 5th to 8th grade.

Simulation Study 1

In the first simulation study, we examined the recovery of the generating latent linear growth parameters under various conditions. We assume a structure consisting of a latent trait measured by the same (either four or eight) Likert-type items at each of three time points, where

the latent trait is assumed to grow linearly across time. We varied the number of items in each time point (*n*=4 or *n*=8), the number of individuals within the cohort (*N*=500, 1,000, and 2,000), the degree of linear change across time (0, .2, and .5 SD per year), and the difficulty of the item parameters (low difficulty items vs. mix of low/high difficulty). Crossing the levels, these four factors result in $(2 \times 3 \times 3 \times 2) = 36$ total conditions. The data for all conditions were generated using the "Simulation" mode in flexMIRT (Cai, 2017). The generating model was based on user-supplied item parameters, latent mean vector, and a latent variance–covariance matrix. All parameters were based on estimates from empirical data in previous studies, including by the Authors (under review).

Population model. We assume that individual *i's* vector of true latent scores $\boldsymbol{\theta_i}$ follow a linear latent growth model. That is to say, we assume that

$$\boldsymbol{\theta_i} = \boldsymbol{\Lambda\eta_i} + \boldsymbol{\epsilon_i} \quad (5)$$

where $\boldsymbol{\Lambda}$ is a fixed factor loading matrix, $\boldsymbol{\eta_i}$ is a vector representing a student's latent intercept and growth term, and $\boldsymbol{\epsilon_i}$ is a vector of time-specific random disturbance terms assumed to be normally distributed with means of zero and variance $\boldsymbol{\Psi}$. Each of the individual's $\boldsymbol{\eta_i}$ can be decomposed into two parts:

$$\boldsymbol{\eta_i} = \boldsymbol{\alpha} + \boldsymbol{\zeta_i}, \quad (6)$$

where $\boldsymbol{\alpha}$ is the population average and $\boldsymbol{\zeta_i}$ represents an individual deviation from that average. The model-implied mean vector and variance-covariance matrix for the latent growth model are

$$\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{\theta_i}) = \boldsymbol{\Lambda\alpha} \quad (7)$$

$$\boldsymbol{\Sigma} = \mathrm{V}(\boldsymbol{\theta_i}) = \boldsymbol{\Lambda\Phi\Lambda'} + \boldsymbol{\Psi}, \quad (8)$$

where $\boldsymbol{\Phi}$ is variance-covariance matrix for the latent factors and $\boldsymbol{\Psi}$ is the residual variance-covariance matrix. The generating parameters for the three timepoint condition are below.

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \boldsymbol{\alpha} = \begin{pmatrix} 0 \\ g \end{pmatrix}, \boldsymbol{\Phi} = \begin{bmatrix} 0.47 & -0.066 \\ -0.066 & 0.08 \end{bmatrix}, \boldsymbol{\Psi} = \begin{bmatrix} .53 & 0 & 0 \\ 0 & .582 & 0 \\ 0 & 0 & .474 \end{bmatrix},$$

where g can take the value of 0, 0.2, or 0.5 depending on the condition. We selected these generating values based on a prior analysis of social-emotional growth across three years (Authors, under review). The generating parameters were used to supply the model-implied mean vector and variance-covariance matrix to flexMIRT for data generation.

Item parameters. After generating the true $\boldsymbol{\theta}$ values, the item responses were simulated using a multidimensional GRM for five response categories. The generating item parameters were slightly modified from an existing measure of children's interpersonal competencies (DeWalt et al., 2013). The two sets of generating item parameter are shown in Table 2. A known problem in the measurement of psychological and social-emotional constructs through self-report surveys is that individuals tend to select primarily from the top two response categories on Likert-type scales, which typically results in item thresholds ($\boldsymbol{b_j}$) that are primarily concentrated on the lower end of the latent scale and little ability to differentiate reliably among students at the top end of the scale (see, for example, Dewalt et al., 2013; Kuhfeld, 2019). Therefore, in our first condition, we chose items only with low thresholds (ranging from -3.01 to .44) to mirror conditions frequently observed with existing psychological measures. However, to ensure our conditions generalized beyond just self-report surveys, our second condition included a range of thresholds across all items (ranging from roughly -2.96 to 2.35). In this study, we assumed full measurement invariance, and so the same item slopes ($\boldsymbol{a_j}$) and threshold ($\boldsymbol{b_j} = b_{1j}, b_{2j}, b_{3j}, b_{4j}$) parameters were used for item *j* across each timepoint.

Calibration and Scoring

IRT calibration/scoring was conducted under each of the four IRT approaches in flexMIRT (Cai & Wirth, 2013). The unidimensional models were estimated using Bock-Aitken EM estimation, while the two MIRT models were estimated using the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010b, 2010a). Estimates of the person-level scores were produced based on the calibrated item parameters using the EAP scoring approach (Bock & Mislevy, 1982). For the cross-sectional calibrations (Approach 1), item parameters calibrated based on the first timepoint only were used to estimate scores for later timepoints. Additionally, we produced sum scores within each timepoint (e.g., summing all item responses for individual *i* in time *t*). Sum scores in each timepoint were standardized by the first timepoint mean and variance estimates. All models converged to a possible local maximum.

Growth models

After scoring all the simulated respondents, a set of parallel linear latent growth models were estimated in lavaan (Rosseel, 2012). One use the true underlying $\boldsymbol{\theta}$ scores, another used sum scores, and four used the estimated $\widehat{\boldsymbol{\theta}}_{EAP}$ scores from the different calibration conditions. In each analysis, the model shown in Equations 5 and 6 was estimated, with the $\widehat{\boldsymbol{\theta}}_{EAP}$ scores substituted for $\boldsymbol{\theta}$ when examining the estimated score results.

Results

In this section, two primary results are evaluated. First, we examine growth model parameter estimates across calibration approaches and simulation conditions. Second, we present and discuss correlations of estimated scores with true scores within each timepoint. We should note that results for IRT Approaches 3 and 4 (the two MIRT models) were extremely similar. Therefore, we only report results from Approach 3.

Table 3 provides a set of growth model parameter estimates across simulation conditions, averaged across 100 replications. We only interpret the growth parameters from the "low difficulty" condition, but the full set of simulation conditions are presented in Appendix Tables A1 and A2. Whereas the MIRT model estimates tend to overstate the slope slightly, the non-MIRT estimates tend to understate the slope significantly, in some cases by almost half the magnitude of the slope (e.g., true slope = .5 based on observed scores). Further, the other IRT models often do not perform demonstrably better than when using observed scores. In general (and as expected), most of the variance estimates are also understated when using non-MIRT models. To make this point clearer, Figure 2 presents estimated growth trajectories for a random sample of 100 simulees by scoring approach. The degree of individual differences in growth trajectories is greatly underestimated when standardized sum scores are used. Instead of having score trajectories that span the y-axis, students' sum scores are highly compressed, poorly reflecting the true variability in starting point and growth observed with the true scores. The two sets of scores from the unidimensional calibrations are an improvement over the sum scores, but only the scores produced from the MIRT model appear to properly capture the wide variation in growth trajectories observed based on the true scores.

Finally, we compared correlations of estimated scores with true scores within each timepoint. Those results are presented in Table 4. In general, correlations were high, especially for the MIRT model. (Correlations were lower between true scores and sum scores across simulation conditions.) However, under some conditions, correlations diminished over time, especially at the third timepoint. This diminution mainly occurred with easy items, and was especially pronounced with a true slope of .5 and only four items. The relatively poor recovery of true scores at timepoint three under these conditions may have occurred due to ceiling effects.

That is, given the item difficulty ranges of the items within the survey, there is little information to distinguish reliably among students at the upper end of the latent continuum. In terms of growth, with a slope of .5, simulees grew quickly, but were already at the high end of the observed response scale in Time 1. By comparison, correlations hardly diminished over time when item difficulty was more varied.

Simulation Study 2

In the second simulation study, we generate true scores across four timepoints following a nonlinear latent growth model and examine the recovery of the linear and quadratic latent means and variances. Here, we generate data for an eight-item measure with *N*= 2,000 simulated respondents.

Data Generation

As with the previous study, the data were generated using the "Simulation" mode in flexMIRT (Cai, 2017). We used the "low difficulty" set of generating item parameters shown in the left set of columns of Table 2. The generating structural parameters used to produce the model-implied mean vector and variance-covariance matrix for $\boldsymbol{\theta}$ in the four-timepoint condition were

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}, \boldsymbol{\alpha} = \begin{pmatrix} 0 \\ .077 \\ -.017 \end{pmatrix}, \boldsymbol{\Phi} = \begin{bmatrix} .337 & .0155 & -.001 \\ .016 & .172 & -.026 \\ -.001 & -.026 & 0.017 \end{bmatrix}, \quad (9)$$

$$\boldsymbol{\Psi} = \begin{bmatrix} 0.666 & 0 & 0 & 0 \\ 0 & 0.570 & 0 & 0 \\ 0 & 0 & .500 & 0 \\ 0 & 0 & 0 & .150 \end{bmatrix}.$$

Calibration and Scoring

The calibration and scoring conditions for this study mirrored Simulation Study 1. One hundred replications were conducted. All of the models converged to a possible local maximum.

Results

Table 5 shows results from the simulation with four timepoints. As in the prior simulation, we present means, variances, and the covariances for the intercept, slope, and quadratic term. In general, the MIRT model marginally outperforms the sum score and cross-sectional models for the fixed effect estimates of the slope and quadratic means. Whereas the MIRT model overstates the linear slope mean slightly, the other three approaches tend to understate the slope.

Turning to variances, the intercept and linear slope variances are much better recovered by the MIRT model. While the MIRT model comes close to the generating parameters, the other models tend to understate the variance substantively. Differences for the covariance estimates tend to be fairly negligible.

Empirical Study

Our empirical study focuses on the impact of various scoring approaches on estimates of students' growth trajectories using a widely-studied construct: growth mindset. There is expanding research on how students' growth mindset develops across elementary and middle school. For example, researchers have used the GPCM to estimate scores on growth mindset, then used those scores to examine growth in the construct over time, as well as teacher and school contributions to that growth (Loeb et al., 2019; West et al., 2016, 2018). Some limited research has shown that growth mindset tends to decrease during middle school (Pintrich & Zusho, 2002), though some of the work by West et al. (2016) and the Authors (2019) actually shows increases on average during this time period. Even given the relative sparseness of the literature on growth for this construct, there is already disagreement on the trends, which may

arise in part due to how the construct is being measured and scored (Duckworth & Yeager, 2015).

Sample. Our study used a sample of students from a California district that is urban, high-poverty, and serves a high proportion of English learners. To avoid conflating across-grade differences in test scores with growth on the underlying construct, we follow a single cohort of students from 5th to 8th grade. The cohort was not intact, with approximately half of the students taking the survey at all three timepoints and half taking the survey during only one or two of the school years. The sample size for our analyses ranged from 2,319 to 3,266 students depending on the number of complete survey responses for a given year. Roughly 20% of students were English learners, 10% were receiving special education services, and most were at approximately the 30th percentile nationally in reading and math achievement (Thum & Hauser, 2015).

Measures. Students in the sample took surveys administered by the district each spring to measure academic growth mindset. Specific items in the survey can be found in Table A3 in the Appendix. Each item uses a five-category Likert scale. Mean growth mindset scale scores tend to increase over time, with a mean of 3.4 in 5th grade and a mean of 3.62 in 8th grade. Longitudinal measurement invariance for the growth mindset survey used in this study was examined previously; configural, weak factorial, and strong factorial invariance were all found to hold (Soland & Kuhfeld, 2019).

Analytic Strategy. We calibrated and scored the growth mindset item responses using IRT approaches 1-3 using flexMIRT (Cai, 2017). Given how similar the scores were in the simulation study when using IRT approaches 3 and 4, we did not estimate the latter. We then fit the same LGCMs as in the simulation studies using lavaan (Rosseel, 2012).

These models were fit with and without a quadratic term. Ultimately, the model that included a quadratic term fit best based on the root mean square error of approximation (RMSEA; Steiger, 2000), the comparative fit index (CFI; Bentler, 1990), and changes in chi-square statistics between models. Thus, we report results for a model with a quadratic term. Like in the simulation studies, we used results from this model for each of the three IRT approaches to compare several parameter estimates. Specifically, across IRT scoring approaches, we examined the means of the latent intercept, linear slope, and quadratic slope. We also compared the variances and covariances of the latent intercept, linear, and quadratic terms.

Results

Table 6 presents parameter estimates from our empirical analyses across the three IRT approaches. As in the simulation results, models using scores from the first two IRT approaches differed little, but both differed substantively from the MIRT model used under the third approach. For example, like in the second simulation, estimates of the intercept and slope means were substantively higher for IRT approach 3 than for the other two. Also like the simulation study, the variances were higher. For instance, the variance of the intercept (growth mindset in 5^{th} grade) was .49 for the MIRT model compared to ~.24 for the other two IRT models.

Similar results tended to hold when comparing sum score results to those from the MIRT model. Though, in some cases, the sum score models actually produced results that were more similar to the MIRT results than the unidimensional IRT models. For example, the estimated slope mean was roughly .09 using sum scores, .06 using unidimensional models, and .12 when using the MIRT model results. The estimated intercept variance based on sum scores also fell between those estimates from the unidimensional and MIRT models.

The most pronounced differences across models were for the covariances of the random effects. In particular, the covariance between the intercept and linear slope was .13 for the MIRT model, but was indistinguishable from zero for the other approaches. While this difference in the estimates is large, it does match the direction of the difference in the second simulation study. Further, the covariance between the intercept and quadratic term is roughly twice as large for the MIRT model compared to the other two. Once again, this difference matched the direction of the difference in the simulation study. While we obviously do not know the true growth parameters in the empirical study, the fact that the results match so closely between the empirical and simulation growth estimates suggests that the latter is not likely driven purely by how we generated the data.

Discussion

Quantifying how children and students develop on a range of latent constructs is vital to understanding how best to support their psychological and educational needs. The associated knowledge base is generally built on research that takes scores from a survey like, say, a self-efficacy scale, and uses those scores in a growth model. Despite the importance of understanding developmental trajectories, most of this research uses observed scores (and sum scores in particular) from surveys in growth models. Research shows that observed scores are often imprecise (if not biased) estimates of the construct of interest, and can lead to biased classifications and rank orderings of respondents based on those scores (Bauer & Curran, 2015; McNeish & Wolf, 2019). Further, while a range of IRT models have been developed specifically to score longitudinal data in a way that accounts for such shortcomings (e.g., Paek, Li, & Park, 2016), these models have rarely been used in the longitudinal survey research literature. In short, we remain unclear on how much reliance on observed scores has likely affected our

understanding of growth trajectories, and which methods of calibration and scoring do the best job of recovering true growth trajectories when scores are used in growth models.

We begin to close this gap in the literature through two simulation studies and one empirical study. In the first simulation study, we simulated true scores for three timepoints with known properties, including the growth trajectory that underlies them. We then scored them using a range of IRT and MIRT models, used those scores in growth models, and compared results to a similar procedure that used sum scores. Results indicate that MIRT models do a superior job of recovering a range of growth parameters. For example, growth estimates using observed scores or unidimensional IRT models tend to understate the true slope in ways that are statistically and practically significant. By contrast, the MIRT models we used closely recover the true slope parameter. Further, non-MIRT models tend to understate the variance of the growth parameters, likely because the calibration approach does not account for covariances in the scores over time. While these results did not tend to differ across simulation conditions, we did find that correlations between estimated and true scores tended to diminish over time in ways that were most pronounced when using shorter surveys that consisted of relatively easy items (though correlations were still high regardless).

Meanwhile, our second simulation study examined similar issues, but using four timepoints instead of three such that nonlinear growth trends could be generated and recovered. Specifically, the generating model included a polynomial term, as did the models used to estimate growth. Findings from this study indicated that, as in the simulation with three timepoints, the MIRT model did a better job of recovering not only linear slope means, but also variances of the growth parameters. Specifically, the non-MIRT models substantively

understated the variances and the mean slope compared to the other three approaches (observed score, cross-sectional, and unidimensional longitudinal).

Finally, our empirical study used results from four timepoints and tended to corroborate our findings from the first simulation study. Though we could not know true growth trajectories for the empirical growth mindset data, we did find that estimates of the slope and most estimated variances were larger when using the MIRT model than observed scores or unidimensional IRT models. Given these differences across models parallel results from the first simulation study, and that the MIRT model in that simulation study did a better job of recovering true growth parameters, findings from the empirical study are congruous with an argument that simulation study results are not purely due to the assumptions we used to generate those data.

Limitations and Future Research

A few limitations of this study bear mention. First, like in any simulation study, we were limited in terms of the range of conditions, assumptions, and models we could use to test our hypotheses. For example, results could differ dependent on the specific data generating growth model. We also did not compare all possible variations on the available IRT and MIRT models, though we tried to capture the most common and likely models in the literature. Thus, results should be replicated using different data generating assumptions and scoring models.

Second, our empirical study was limited to only a single district that serves a high proportion of low-income and English learner students. Further, that district only administered a single growth mindset survey (once per timepoint). On one hand, the survey has been tested for measurement invariance between English learner and native English-speaking students. On the other, broader issues of generalizability remain. The study should be repeated using a different students and measures to ensure results hold.

Beyond the limitations of our empirical and simulation studies, there are other issues future researchers may wish to consider. For example, we did not compare IRT model performance in recovering growth parameters when there are failures of longitudinal measurement invariance, whether minor or severe. If we were to loosen the measurement invariance assumptions we made, the cross-sectional and longitudinal unidimensional IRT results would most likely not appear as parallel as what we show. We also did not examine what might happen if the composition of the items used to measure the construct shifts across timepoints, such as when a few anchor items are maintained and others are changed. Additionally, while we examined multiple IRT calibration approaches in this study, we only used a single method to produce latent scores (e.g., EAP scoring). Future research should examine other potential scoring approaches, including approaches that condition on key background variables that may influence growth trajectories (Curran et al., 2018).

Finally, we did not directly compare our two-stage approach with results from second-order (multiple-indicator) LGCMs (e.g., McArdle, 1988; Hancock, Kuo, & Lawrence, 2001). We chose to focus only the two-stage approach (calibration/scoring followed by first-order LGCMs) because (a) this approach is widely-used by applied researchers, and (b) little is known about the impact of the calibration model in the context where a short (4-8 item) self-report Likert-type survey is administered across multiple timepoints. Prior researchers have found strong advantages of second-order LGCMs over first-order LGCMs that use simple (sum or mean) scores as the composite indicator (Geiser, Keller, Lockhart, 2013). However, future research should compare first-order and second-order LGCMs across a range of possible calibration conditions to understand whether the benefits of second-order LGCMs are maintained when the

composite scores in the first-order model is based on a MIRT model that accurately captures the residual correlations inherent in the multi-timepoint data.

Conclusion

This study addresses the discrepancy between the measurement of outcomes and the modeling of growth in the context of psychological and social-emotional development. We compare multiple approaches to calibrate and score Likert-type measures administered across multiple timepoints when the goal is to accurately capture the average individual's growth trajectory as well as the variability in change across time. We find that, while MIRT models are not a panacea for all measurement issues that can arise when attempting to quantify growth, a MIRT approach is likely preferable when survey responses from all timepoints are available at the time of calibration. We show that these models generally do a better job of recovering true developmental trajectories (in particular linear growth trends), as well as variance components from growth models. By contrast, using sum scores or even unidimensional IRT models can understate the true slope, in some cases by a magnitude of 50% of the true growth. Estimated growth parameter variances for sum score and unidimensional IRT models are also much more muted relative to the true variance of growth parameters, as well as those generated by our MIRT models.

Notes

1. The Collaborative for Academic, Social, and Emotional Learning (CASEL) defines social-emotional learning (SEL) as the fostering of social and emotional competencies through explicit instruction and through student-centered learning approaches that help students engage in the learning process and develop analytical, communication, and collaborative skills. Competencies are skills that can be taught and learned through appropriate pedagogy. Social-emotional learning strategies focus on development of skills like building healthy peer relationships, responsible decision making, self-management, self-awareness, and social awareness to succeed in school. However, SEL does not encompass mental health conditions such as depression and anxiety disorders, nor certain psychological constructs less directly related to classroom performance.

2. We did do an additional simulation using a cross-sectional (range of ages) design (middle panel of Table 1). However, this simulation required a different data generating model and produced results that were nearly identical to those using the top panel of Table 1. Therefore, we do not report those results.

References

Bauer, D. J., & Curran, P. J. (2016). The discrepancy between measurement and modeling in longitudinal data analysis. In J. R. Harring, L. M. Stapleton & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research* (pp. 3–38). Charlotte, NC. Information Age Publishing.

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

Bishop, J., Geiser, C., & Cole, D. A. (2015). Modeling latent growth with multiple indicators: A comparison of three approaches. *Psychological Methods, 20*(1), 43-62.

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

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Cai, L. (2017). flexMIRT R version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G. M., Barbaranelli, C., & Bandura, A. (2008). Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. *Journal of Educational Psychology, 100*(3), 525.

Caprara, G. V., Vecchione, M., Alessandri, G., Gerbino, M., & Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology*, *81*(1), 78–96.

Crutzen, R., & Peters, G.-J. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*, *11*(3), 242–247.

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., & Hussong, A. M. (2018). Recovering predictor–criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(6), 860-875.

DeWalt, D. A., Thissen, D., Stucky, B. D., Langer, M. M., Morgan DeWitt, E., Irwin, D. E., … Taylor, O. (2013). PROMIS pediatric peer relationships scale: Development of a peer relationships item bank as part of social health measurement. *Health Psychology*, *32*(10), 1093.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. London, UK: Psychology Press.

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, *31*(3).

Geiser, C., Keller, B. T., & Lockhart, G. (2013). First-versus second-order latent growth curve models: Some insights from latent state-trait theory. *Structural equation modeling: a multidisciplinary journal*, *20*(3), 479-503.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, *29*(3), 8–14.

Koran, J. (2009). *An Integrated Item Response Model for Evaluating Individual Students' Growth in Educational Achievement* (PhD Thesis).

Loeb, S., Christian, M. S., Hough, H., Meyer, R. H., Rice, A. B., & West, M. R. (In Press). School differences in social–emotional learning gains: Findings from the first large-scale panel survey of students. *Journal of Educational and Behavioral Statistics*.

Hancock, G., Kuo, W., & Lawrence, F. (2001). *An illustration of second-order latent growth models. Structural Equation Modeling, 8* (3), 470–489.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among paranletric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18*(3), 245–256.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. Nesselroade & R. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561–614). New York, NY: Plenum Press.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Abingdon, UK: Routledge.

Paek, I., Li, Z., & Park, H.-J. (2016). Specifying ability growth models using a multidimensional item response model for repeated measures categorical ordinal item response data. *Multivariate Behavioral Research, 51*(4), 569–580.

Paek, I., Park, H.-J., Cai, L., & Chi, E. (2014). A comparison of three IRT approaches to examinee ability change modeling in a single-group anchor test design. *Educational and Psychological Measurement, 74*(4), 659–676.

Pintrich, P. R., & Zusho, A. (2002). The development of academic self-regulation: The role of cognitive and motivational factors. In *Development of achievement motivation* (pp. 249–284). Elsevier.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *17*(4).

Samejima, F. (2011). The general graded response model. In *Handbook of polytomous item response theory models* (pp. 87–118). New York, NY: Routledge.

Soland, J., Jensen, N., Keys, T. D., Bi, S. Z., & Wolk, E. (2019). Are test and academic disengagement related? Implications for measurement and practice. *Educational Assessment*, 1–16.

Soland, J., Kuhfeld, M., Wolk, E., & Bi, S. (2019). Examining the state-trait composition of social-emotional learning constructs: Implications for practice, policy, and evaluation. *Journal of Research on Educational Effectiveness, 12*(3), 550–577.

van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. New York, NY: Springer Science & Business Media.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F., & Gabrieli, J. D. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, *38*(1), 148–170.

West, M. R., Pier, L., Fricke, H., Loeb, S., Meyer, R. H., & Rice, A. B. (2018). Trends in student social-emotional learning: Evidence from the CORE districts. Stanford, CA: *Policy Analysis for California Education, PACE*.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18.

Willoughby, M. T., Wirth, R. J., & Blair, C. B. (2012). Executive function in early childhood: Longitudinal measurement invariance and developmental change. *Psychological Assessment*, *24*(2), 418.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58.

Table 1

Comparison of Cross-sectional (Restricted-Age), Cross-Sectional (Range of Ages), and Longitudinal Calibration Approaches

Grade	Time 1	Time 2	Time 3	Time 4
Restricted-Age Cross-Sectional Design				
5	X			
6				
7				
8				
Cross-Sectional (Range of Ages) Design				
5	X			
6	X			
7	X			
8	X			
Longitudinal (Cohort) Design				
5	X			
6		X		
7			X	
8				X

Table 2

Generating Parameters for Simulation Studies 1 and 2

	"Easy" Item Difficulty Condition					"Mixed" Item Difficulty Condition				
Item	a	b1	b2	b3	b4	a	b1	b2	b3	b4
v1	2.69	-1.99	-1.32	-0.88	-0.19	2.69	-1.99	-0.88	0.50	1.50
v2	2.00	-1.93	-1.57	-0.80	-0.05	2.00	-1.93	-0.69	1.23	2.04
v3	1.94	-2.18	-1.76	-0.59	0.10	1.94	-2.18	-1.76	-0.59	0.10
v4	1.90	-2.96	-2.25	-1.00	-0.06	1.90	-2.96	-2.25	-1.00	-0.06
v5	1.83	-2.51	-2.15	-0.70	0.37	1.83	-2.51	-2.15	-0.70	0.37
v6	1.86	-2.69	-2.19	-1.08	-0.15	1.76	-2.69	-2.19	-1.08	-0.15
v7	1.94	-2.65	-2.11	-0.78	0.44	1.54	-1.50	-0.65	0.78	1.95
v8	1.74	-3.01	-2.53	-1.36	-0.48	1.74	-1.25	-0.05	1.22	2.35

Note. For the *n*=4 condition, the item parameters for items v1-v4 were used to generate data. We show the item threshold (*b*) parameters instead of intercept (*c*) parameters to allow for clearer picture of the range of the latent trait covered by the items. However, the generating intercepts can be calculated as $c = -ab$. We assume strong measurement invariance, and so the same generating item parameters are used at each timepoint.

Table 3

Growth Model Parameter Estimates for Specific Conditions within Simulation Study 1

Condition	Parameter	Population	True Scores	Sum Scores	Cross-sectional	Long. UniD.	Long. MIRT
			True Slope = 0				
N=2000, J=4 items	Latent Means						
	Intercept Mean	0.000	0.002	0.024	0.014	0.013	0.006
	Slope Mean	0.000	-0.001	-0.001	-0.001	-0.001	0.004
	(Co)variance Estimates						
	Intercept Variance	0.470	0.478	0.289	0.277	0.277	0.459
	Slope Variance	0.080	0.084	0.052	0.049	0.049	0.081
	Covariance	-0.066	-0.071	-0.043	-0.041	-0.041	-0.064
N=2000, J=8 items	Latent Means						
	Intercept Mean	0.000	0.002	0.001	0.001	0.001	0.004
	Slope Mean	0.000	-0.001	-0.001	-0.001	-0.001	0.001
	(Co)variance Estimates						
	Intercept Variance	0.470	0.478	0.238	0.345	0.345	0.475
	Slope Variance	0.080	0.084	0.042	0.060	0.060	0.082
	Covariance	-0.066	-0.071	-0.035	-0.051	-0.051	-0.068
			True Slope = 0.2				
N=2000, J=4 items	Latent Means						
	Intercept Mean	0.000	0.002	0.003	0.002	0.002	0.005
	Slope Mean	0.200	0.199	0.136	0.144	0.145	0.212
	(Co)variance Estimates						
	Intercept Variance	0.470	0.478	0.283	0.282	0.285	0.458
	Slope Variance	0.080	0.084	0.045	0.047	0.047	0.080
	Covariance	-0.066	-0.071	-0.059	-0.052	-0.052	-0.065
N=2000, J=8 items	Latent Means						
	Intercept Mean	0.000	0.199	0.124	0.163	0.163	0.206
	Slope Mean	0.200	0.002	0.003	0.002	0.002	0.003
	(Co)variance Estimates						
	Intercept Variance	0.470	0.478	0.233	0.348	0.347	0.475
	Slope Variance	0.080	0.084	0.037	0.058	0.057	0.081
	Covariance	-0.066	-0.071	-0.048	-0.061	-0.060	-0.068
			True Slope = 0.5				
N=2000, J=4 items	Latent Means						
	Intercept Mean	0.000	0.002	0.024	0.014	0.013	0.006
	Slope Mean	0.500	0.499	0.277	0.324	0.326	0.526
	(Co)variance Estimates						
	Intercept Variance	0.470	0.478	0.268	0.287	0.284	0.460
	Slope Variance	0.080	0.084	0.042	0.048	0.048	0.077
	Covariance	-0.066	-0.071	-0.081	-0.075	-0.073	-0.069
N=2000, J=8 items	Latent Means						
	Intercept Mean	0.000	0.002	0.020	0.010	0.008	0.005
	Slope Mean	0.500	0.499	0.257	0.374	0.368	0.515
	(Co)variance Estimates						
	Intercept Variance	0.470	0.478	0.223	0.342	0.323	0.476
	Slope Variance	0.080	0.084	0.034	0.056	0.054	0.078
	Covariance	-0.066	-0.071	-0.065	-0.080	-0.074	-0.070

Table 4

Correlations between True and Estimated Scores by Timepoint and Model

	Time			Sum Scores			Unidimensional - T1 Only			Unidimensional - Long			MIRT		
N	points	Items	Gain	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3	Time 1	Time 2	Time 3
Easy Item Difficulty Condition															
500	3	4	0	0.84	0.84	0.84	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.88	0.88
500	3	4	0.2	0.84	0.82	0.80	0.87	0.85	0.83	0.87	0.86	0.83	0.87	0.87	0.85
500	3	4	0.5	0.84	0.79	0.71	0.87	0.83	0.76	0.87	0.83	0.76	0.87	0.84	0.78
500	3	8	0	0.89	0.89	0.89	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
500	3	8	0.2	0.89	0.88	0.85	0.92	0.91	0.90	0.92	0.91	0.90	0.92	0.92	0.90
500	3	8	0.5	0.89	0.85	0.78	0.92	0.89	0.84	0.92	0.89	0.84	0.92	0.90	0.85
1000	3	4	0	0.84	0.84	0.84	0.87	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.88
1000	3	4	0.2	0.84	0.82	0.80	0.87	0.86	0.84	0.87	0.86	0.84	0.88	0.87	0.85
1000	3	4	0.5	0.84	0.79	0.71	0.87	0.83	0.76	0.87	0.83	0.76	0.88	0.84	0.78
1000	3	8	0	0.89	0.89	0.89	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
1000	3	8	0.2	0.89	0.88	0.86	0.92	0.91	0.90	0.92	0.91	0.90	0.92	0.92	0.90
1000	3	8	0.5	0.89	0.85	0.78	0.92	0.89	0.84	0.92	0.89	0.84	0.92	0.90	0.85
2000	3	4	0	0.84	0.84	0.85	0.87	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.88
2000	3	4	0.2	0.84	0.83	0.80	0.87	0.86	0.84	0.87	0.86	0.84	0.88	0.87	0.85
2000	3	4	0.5	0.84	0.79	0.72	0.87	0.83	0.76	0.87	0.83	0.76	0.88	0.84	0.78
2000	3	8	0	0.89	0.89	0.89	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
2000	3	8	0.2	0.89	0.87	0.86	0.92	0.91	0.90	0.92	0.91	0.90	0.92	0.92	0.90
2000	3	8	0.5	0.89	0.85	0.79	0.92	0.89	0.84	0.92	0.89	0.84	0.92	0.90	0.86
Mixed Item Difficulty Condition															
500	3	4	0	0.88	0.88	0.88	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
500	3	4	0.2	0.88	0.88	0.87	0.90	0.89	0.89	0.90	0.89	0.89	0.90	0.90	0.90
500	3	4	0.5	0.88	0.87	0.85	0.90	0.89	0.87	0.90	0.89	0.87	0.90	0.89	0.88
500	3	8	0	0.92	0.92	0.92	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
500	3	8	0.2	0.92	0.92	0.92	0.94	0.93	0.93	0.94	0.93	0.93	0.94	0.94	0.93
500	3	8	0.5	0.92	0.91	0.90	0.94	0.93	0.92	0.94	0.93	0.92	0.94	0.93	0.92
1000	3	4	0	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
1000	3	4	0.2	0.89	0.88	0.88	0.90	0.90	0.89	0.90	0.90	0.89	0.90	0.90	0.90
1000	3	4	0.5	0.89	0.87	0.85	0.90	0.89	0.87	0.90	0.89	0.87	0.90	0.90	0.88
1000	3	8	0	0.92	0.92	0.92	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
1000	3	8	0.2	0.92	0.92	0.92	0.94	0.94	0.93	0.94	0.94	0.93	0.94	0.94	0.93
1000	3	8	0.5	0.92	0.91	0.90	0.94	0.93	0.92	0.94	0.93	0.92	0.94	0.93	0.92
2000	3	4	0	0.88	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.91	0.90
2000	3	4	0.2	0.88	0.88	0.88	0.90	0.90	0.89	0.90	0.90	0.89	0.90	0.90	0.90
2000	3	4	0.5	0.88	0.87	0.85	0.90	0.89	0.87	0.90	0.89	0.87	0.90	0.90	0.88
2000	3	8	0	0.92	0.92	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
2000	3	8	0.2	0.92	0.92	0.92	0.94	0.93	0.93	0.94	0.93	0.93	0.94	0.94	0.93
2000	3	8	0.5	0.92	0.91	0.90	0.94	0.93	0.92	0.94	0.93	0.92	0.94	0.93	0.92

Table 5

Growth Model Parameter Estimates for Simulation Study 2

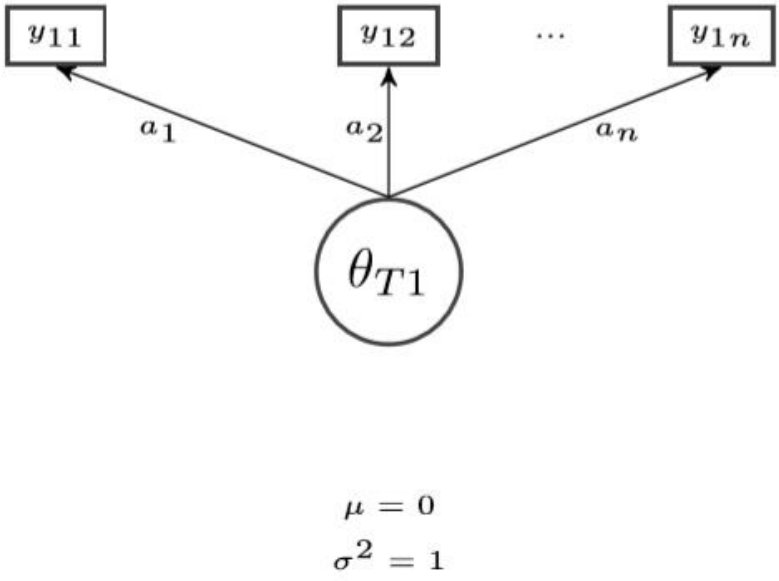
Condition	Parameter	Population	True Scores	Sum Scores	Cross-sectional	Long. UniD.	Long. MIRT
N=2000, J=8 items	Latent Means						
	Intercept Mean	0.000	0.001	0.001	0.001	0.001	0.004
	Slope Mean	0.077	0.077	0.059	0.066	0.072	0.085
	Quad. Mean	-0.017	-0.018	-0.028	-0.022	-0.025	-0.020
	Variance Estimates						
	Intercept Variance	0.337	0.346	0.168	0.216	0.256	0.348
	Slope Variance	0.172	0.177	0.08	0.107	0.127	0.168
	Quad. Variance	0.018	0.017	0.008	0.010	0.011	0.015
	Covariance Estimates						
	Int. Lin. Covariance	0.016	0.008	0.004	0.001	0.001	0.010
	Int. Quad. Covariance	-0.010	-0.008	-0.005	-0.006	-0.007	-0.008
	Lin. Quad. Covariance	-0.026	-0.027	-0.012	-0.017	-0.02	-0.022

Table 6

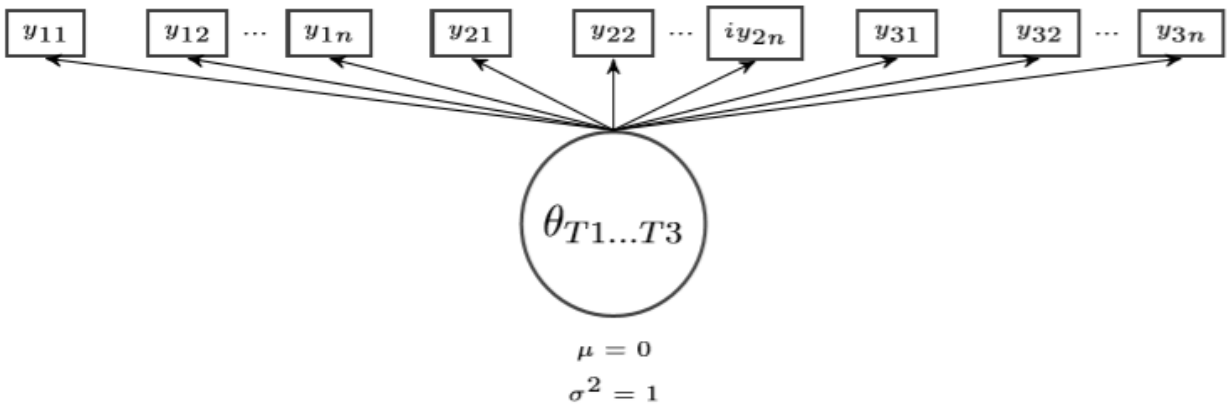
Parameter Estimates from Empirical Analyses Using Growth Mindset

Parameter	Sum Scores	Cross-sectional	Long. UniD.	Long. MIRT
Intercept Mean	-0.016	-0.039	-0.090	0.009
	(0.025)	(0.020)	(0.020)	(0.021)
Slope Mean	0.093	0.060	0.061	0.122
	(0.033)	(0.026)	(0.026)	(0.022)
Quad. Mean	0.003	0.004	0.003	-0.001
	(0.010)	(0.008)	(0.008)	(0.007)
Intercept Variance	0.381	0.236	0.229	0.491
	(0.040)	(0.026)	(0.025)	(0.026)
Slope Variance	0.235	0.106	0.106	0.231
	(0.081)	(0.052)	(0.051)	(0.035)
Quad. Variance	0.017	0.008	0.008	0.017
	(0.008)	(0.005)	(0.005)	(0.003)
Int. Lin. Covariance	-0.020	0.026	0.026	0.131
	(0.043)	(0.027)	(0.027)	(0.021)
Int. Quad. Covariance	-0.003	-0.011	-0.011	-0.041
	(0.012)	(0.008)	(0.008)	(0.006)
Lin. Quad. Covariance	-0.060	-0.026	-0.027	-0.055
	(0.025)	(0.016)	(0.016)	(0.010)

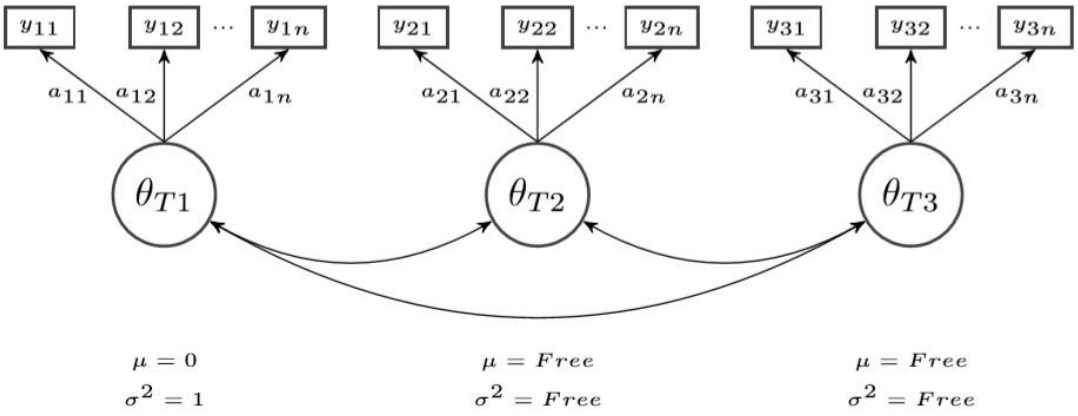
(1a) Approach 1: Cross-sectional IRT model for item response at time 1



(1b) Approach 2: IRT model for item responses at all timepoints from one longitudinal cohort (data in long format)



(1c) Approach 3: Longitudinal MIRT model for three timepoints of data



(1d) Approach 4: Longitudinal MIRT model for three timepoints of data with specific factors

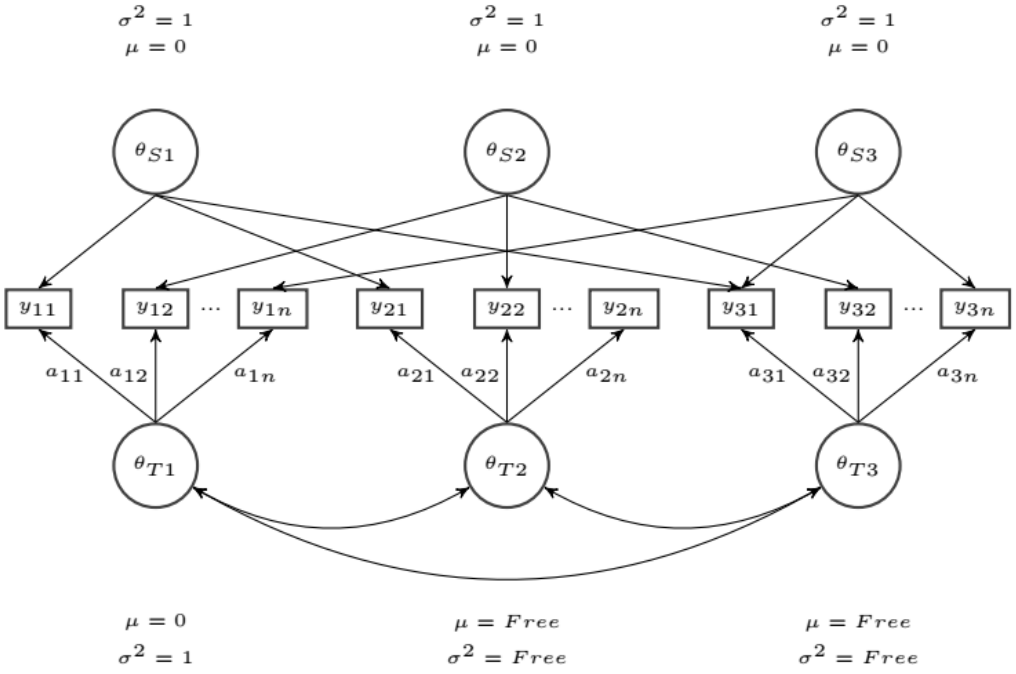


Figure 1. Path diagrams for the four IRT calibration models considered in this study. Each box y_{tij} represents an observed item response at time *t* from individual *i* to item *j*. For parsimony within this figure, we leave off the *i* subscript, so the notation within each observed item above is y_{tj}. For the two MIRT models, measurement invariance constraints are applied so that $a_{1j} = a_{2j} = a_{3j}$.

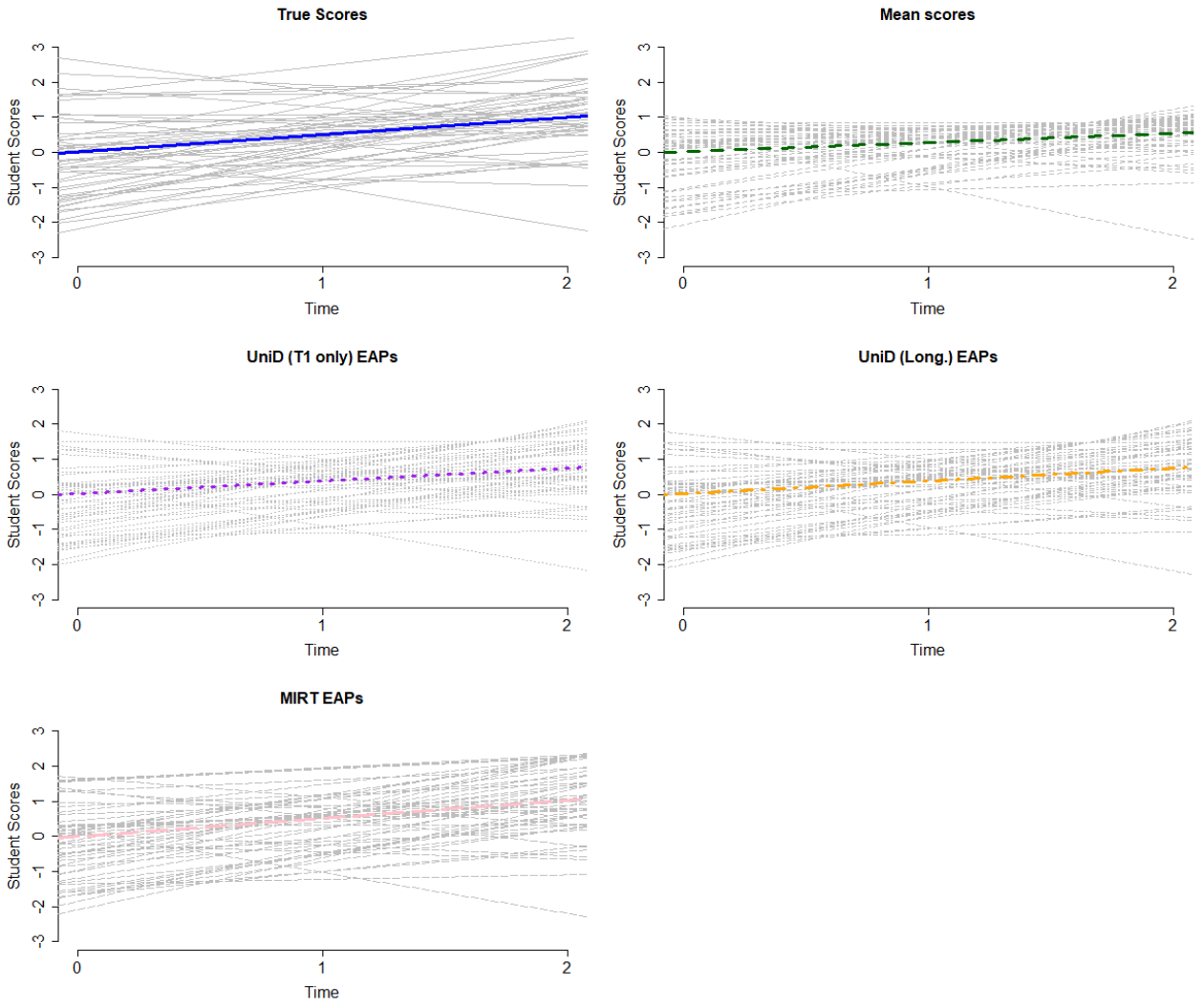


Figure 2. Average trajectory and variability of trajectories across 200 randomly selected simulees from the J=4, N=2000, gain=.5 condition.

Appendix

Table A1

Full Simulation Study 1 Results Across Conditions – Easy Items

Condition	Parameter	Pop.	True Scores	Mean Scores	Cross-sectional	Long. UniD.	Long. MIRT
N=500, True gain=0, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0, J=4 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0, J=4 items	Intercept Variance	0.47	0.47	0.46	0.28	0.27	0.27
N=500, True gain=0, J=4 items	Slope Variance	0.08	0.08	0.08	0.05	0.05	0.05
N=500, True gain=0, J=4 items	Covariance	-0.07	-0.07	-0.07	-0.04	-0.04	-0.04
N=500, True gain=0.2, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0.2, J=4 items	Slope Mean	0.20	0.20	0.20	0.14	0.15	0.15
N=500, True gain=0.2, J=4 items	Intercept Variance	0.47	0.47	0.46	0.28	0.28	0.28
N=500, True gain=0.2, J=4 items	Slope Variance	0.08	0.08	0.08	0.04	0.05	0.05
N=500, True gain=0.2, J=4 items	Covariance	-0.07	-0.07	-0.07	-0.06	-0.05	-0.05
N=500, True gain=0.5, J=4 items	Intercept Mean	0.00	0.00	0.00	0.02	0.01	0.01
N=500, True gain=0.5, J=4 items	Slope Mean	0.50	0.50	0.50	0.28	0.33	0.33
N=500, True gain=0.5, J=4 items	Intercept Variance	0.47	0.47	0.46	0.27	0.28	0.28
N=500, True gain=0.5, J=4 items	Slope Variance	0.08	0.08	0.08	0.04	0.05	0.05
N=500, True gain=0.5, J=4 items	Covariance	-0.07	-0.07	-0.07	-0.08	-0.07	-0.08
N=500, True gain=0, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0, J=8 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0, J=8 items	Intercept Variance	0.47	0.47	0.46	0.24	0.35	0.35
N=500, True gain=0, J=8 items	Slope Variance	0.08	0.08	0.08	0.04	0.06	0.06
N=500, True gain=0, J=8 items	Covariance	-0.07	-0.07	-0.07	-0.04	-0.06	-0.06
N=500, True gain=0.2, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0.2, J=8 items	Slope Mean	0.20	0.20	0.20	0.13	0.17	0.17
N=500, True gain=0.2, J=8 items	Intercept Variance	0.47	0.47	0.46	0.23	0.35	0.35

N=500, True gain=0.2, J=8 items	Slope Variance	0.08	0.08	0.08	0.04	0.06	0.06
N=500, True gain=0.2, J=8 items	Covariance	-0.07	-0.07	-0.07	-0.05	-0.06	-0.06
N=500, True gain=0.5, J=8 items	Intercept Mean	0.00	0.00	0.00	0.02	0.01	0.01
N=500, True gain=0.5, J=8 items	Slope Mean	0.50	0.50	0.50	0.26	0.37	0.38
N=500, True gain=0.5, J=8 items	Intercept Variance	0.47	0.47	0.46	0.22	0.33	0.35
N=500, True gain=0.5, J=8 items	Slope Variance	0.08	0.08	0.08	0.04	0.06	0.06
N=500, True gain=0.5, J=8 items	Covariance	-0.07	-0.07	-0.07	-0.07	-0.08	-0.09
N=1000, True gain=0, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0, J=4 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0, J=4 items	Intercept Variance	0.47	0.47	0.46	0.28	0.27	0.27
N=1000, True gain=0, J=4 items	Slope Variance	0.08	0.08	0.07	0.05	0.04	0.04
N=1000, True gain=0, J=4 items	Covariance	-0.07	-0.07	-0.06	-0.04	-0.04	-0.04
N=1000, True gain=0.2, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0.2, J=4 items	Slope Mean	0.20	0.20	0.20	0.14	0.15	0.15
N=1000, True gain=0.2, J=4 items	Intercept Variance	0.47	0.47	0.46	0.28	0.28	0.28
N=1000, True gain=0.2, J=4 items	Slope Variance	0.08	0.08	0.07	0.04	0.04	0.04
N=1000, True gain=0.2, J=4 items	Covariance	-0.07	-0.07	-0.06	-0.06	-0.05	-0.05
N=1000, True gain=0.5, J=4 items	Intercept Mean	0.00	0.00	0.00	0.02	0.01	0.01
N=1000, True gain=0.5, J=4 items	Slope Mean	0.50	0.50	0.50	0.28	0.33	0.33
N=1000, True gain=0.5, J=4 items	Intercept Variance	0.47	0.47	0.46	0.26	0.28	0.28
N=1000, True gain=0.5, J=4 items	Slope Variance	0.08	0.08	0.07	0.04	0.04	0.04
N=1000, True gain=0.5, J=4 items	Covariance	-0.07	-0.07	-0.06	-0.08	-0.07	-0.07
N=1000, True gain=0, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0, J=8 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0, J=8 items	Intercept Variance	0.47	0.47	0.46	0.23	0.33	0.33
N=1000, True gain=0, J=8 items	Slope Variance	0.08	0.08	0.07	0.04	0.05	0.05
N=1000, True gain=0, J=8 items	Covariance	-0.07	-0.07	-0.06	-0.03	-0.04	-0.04
N=1000, True gain=0.2, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0.2, J=8 items	Slope Mean	0.20	0.20	0.20	0.13	0.17	0.17
N=1000, True gain=0.2, J=8 items	Intercept Variance	0.47	0.47	0.46	0.22	0.34	0.34
N=1000, True gain=0.2, J=8 items	Slope Variance	0.08	0.08	0.07	0.03	0.05	0.05

N=1000, True gain=0.2, J=8 items	Covariance	-0.07	-0.07	-0.06	-0.04	-0.05	-0.05
N=1000, True gain=0.5, J=8 items	Intercept Mean	0.00	0.00	0.00	0.02	0.01	0.01
N=1000, True gain=0.5, J=8 items	Slope Mean	0.50	0.50	0.50	0.26	0.37	0.38
N=1000, True gain=0.5, J=8 items	Intercept Variance	0.47	0.47	0.46	0.22	0.32	0.34
N=1000, True gain=0.5, J=8 items	Slope Variance	0.08	0.08	0.07	0.03	0.05	0.05
N=1000, True gain=0.5, J=8 items	Covariance	-0.07	-0.07	-0.06	-0.06	-0.07	-0.08
N=2000, True gain=0, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0, J=4 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0, J=4 items	Intercept Variance	0.47	0.47	0.48	0.29	0.28	0.28
N=2000, True gain=0, J=4 items	Slope Variance	0.08	0.08	0.08	0.05	0.05	0.05
N=2000, True gain=0, J=4 items	Covariance	-0.07	-0.07	-0.07	-0.04	-0.04	-0.04
N=2000, True gain=0.2, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0.2, J=4 items	Slope Mean	0.20	0.20	0.20	0.14	0.15	0.14
N=2000, True gain=0.2, J=4 items	Intercept Variance	0.47	0.47	0.48	0.28	0.29	0.28
N=2000, True gain=0.2, J=4 items	Slope Variance	0.08	0.08	0.08	0.05	0.05	0.05
N=2000, True gain=0.2, J=4 items	Covariance	-0.07	-0.07	-0.07	-0.06	-0.05	-0.05
N=2000, True gain=0.5, J=4 items	Intercept Mean	0.00	0.00	0.00	0.02	0.01	0.01
N=2000, True gain=0.5, J=4 items	Slope Mean	0.50	0.50	0.50	0.28	0.33	0.32
N=2000, True gain=0.5, J=4 items	Intercept Variance	0.47	0.47	0.48	0.27	0.28	0.29
N=2000, True gain=0.5, J=4 items	Slope Variance	0.08	0.08	0.08	0.04	0.05	0.05
N=2000, True gain=0.5, J=4 items	Covariance	-0.07	-0.07	-0.07	-0.08	-0.07	-0.08
N=2000, True gain=0, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0, J=8 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0, J=8 items	Intercept Variance	0.47	0.47	0.48	0.24	0.35	0.35
N=2000, True gain=0, J=8 items	Slope Variance	0.08	0.08	0.08	0.04	0.06	0.06
N=2000, True gain=0, J=8 items	Covariance	-0.07	-0.07	-0.07	-0.04	-0.05	-0.05
N=2000, True gain=0.2, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0.2, J=8 items	Slope Mean	0.20	0.20	0.20	0.12	0.16	0.16
N=2000, True gain=0.2, J=8 items	Intercept Variance	0.47	0.47	0.48	0.23	0.35	0.35
N=2000, True gain=0.2, J=8 items	Slope Variance	0.08	0.08	0.08	0.04	0.06	0.06
N=2000, True gain=0.2, J=8 items	Covariance	-0.07	-0.07	-0.07	-0.05	-0.06	-0.06

N=2000, True gain=0.5, J=8 items	Intercept Mean	0.00	0.00	0.00	0.02	0.01	0.01
N=2000, True gain=0.5, J=8 items	Slope Mean	0.50	0.50	0.50	0.26	0.37	0.37
N=2000, True gain=0.5, J=8 items	Intercept Variance	0.47	0.47	0.48	0.22	0.32	0.34
N=2000, True gain=0.5, J=8 items	Slope Variance	0.08	0.08	0.08	0.03	0.05	0.06
N=2000, True gain=0.5, J=8 items	Covariance	-0.07	-0.07	-0.07	-0.07	-0.07	-0.08

Table A2

Full Simulation Study 1 Results Across Conditions – Mixed Item Difficulties

Condition	Parameter	Pop.	True Scores	Mean Scores	Cross-sectional	Long. UniD.	Long. MIRT
N=500, True gain=0, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0, J=4 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.01
N=500, True gain=0, J=4 items	Intercept Variance	0.47	0.46	0.23	0.30	0.30	0.45
N=500, True gain=0, J=4 items	Slope Variance	0.08	0.08	0.04	0.05	0.05	0.08
N=500, True gain=0, J=4 items	Covariance	-0.07	-0.06	-0.03	-0.04	-0.04	-0.06
N=500, True gain=0.2, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0.2, J=4 items	Slope Mean	0.20	0.20	0.14	0.16	0.16	0.21
N=500, True gain=0.2, J=4 items	Intercept Variance	0.47	0.46	0.23	0.30	0.30	0.46
N=500, True gain=0.2, J=4 items	Slope Variance	0.08	0.08	0.04	0.05	0.05	0.08
N=500, True gain=0.2, J=4 items	Covariance	-0.07	-0.06	-0.04	-0.04	-0.04	-0.06
N=500, True gain=0.5, J=4 items	Intercept Mean	0.00	0.00	0.01	0.00	0.00	0.00
N=500, True gain=0.5, J=4 items	Slope Mean	0.50	0.50	0.32	0.37	0.38	0.52
N=500, True gain=0.5, J=4 items	Intercept Variance	0.47	0.46	0.23	0.26	0.29	0.46
N=500, True gain=0.5, J=4 items	Slope Variance	0.08	0.08	0.03	0.04	0.05	0.08
N=500, True gain=0.5, J=4 items	Covariance	-0.07	-0.06	-0.05	-0.04	-0.05	-0.06
N=500, True gain=0, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0, J=8 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0, J=8 items	Intercept Variance	0.47	0.46	0.23	0.36	0.36	0.48
N=500, True gain=0, J=8 items	Slope Variance	0.08	0.08	0.04	0.07	0.07	0.09
N=500, True gain=0, J=8 items	Covariance	-0.07	-0.06	-0.04	-0.05	-0.05	-0.07
N=500, True gain=0.2, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=500, True gain=0.2, J=8 items	Slope Mean	0.20	0.20	0.14	0.17	0.18	0.21
N=500, True gain=0.2, J=8 items	Intercept Variance	0.47	0.46	0.23	0.35	0.36	0.48
N=500, True gain=0.2, J=8 items	Slope Variance	0.08	0.08	0.04	0.06	0.06	0.09
N=500, True gain=0.2, J=8 items	Covariance	-0.07	-0.06	-0.04	-0.05	-0.06	-0.07

N=500, True gain=0.5, J=8 items	Intercept Mean	0.00	0.00	0.01	0.00	0.00	0.00
N=500, True gain=0.5, J=8 items	Slope Mean	0.50	0.50	0.31	0.39	0.41	0.52
N=500, True gain=0.5, J=8 items	Intercept Variance	0.47	0.46	0.23	0.29	0.33	0.48
N=500, True gain=0.5, J=8 items	Slope Variance	0.08	0.08	0.04	0.06	0.06	0.09
N=500, True gain=0.5, J=8 items	Covariance	-0.07	-0.06	-0.05	-0.05	-0.06	-0.07
N=1000, True gain=0, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0, J=4 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0, J=4 items	Intercept Variance	0.47	0.46	0.23	0.30	0.30	0.46
N=1000, True gain=0, J=4 items	Slope Variance	0.08	0.08	0.04	0.05	0.05	0.07
N=1000, True gain=0, J=4 items	Covariance	-0.07	-0.06	-0.03	-0.04	-0.04	-0.06
N=1000, True gain=0.2, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0.2, J=4 items	Slope Mean	0.20	0.20	0.14	0.16	0.16	0.21
N=1000, True gain=0.2, J=4 items	Intercept Variance	0.47	0.46	0.23	0.30	0.30	0.46
N=1000, True gain=0.2, J=4 items	Slope Variance	0.08	0.08	0.04	0.05	0.05	0.07
N=1000, True gain=0.2, J=4 items	Covariance	-0.07	-0.06	-0.04	-0.04	-0.04	-0.06
N=1000, True gain=0.5, J=4 items	Intercept Mean	0.00	0.00	0.01	0.00	0.00	0.00
N=1000, True gain=0.5, J=4 items	Slope Mean	0.50	0.50	0.31	0.36	0.38	0.51
N=1000, True gain=0.5, J=4 items	Intercept Variance	0.47	0.46	0.23	0.27	0.29	0.47
N=1000, True gain=0.5, J=4 items	Slope Variance	0.08	0.08	0.03	0.04	0.05	0.08
N=1000, True gain=0.5, J=4 items	Covariance	-0.07	-0.06	-0.05	-0.05	-0.05	-0.06
N=1000, True gain=0, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0, J=8 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0, J=8 items	Intercept Variance	0.47	0.46	0.23	0.36	0.36	0.46
N=1000, True gain=0, J=8 items	Slope Variance	0.08	0.08	0.04	0.06	0.06	0.07
N=1000, True gain=0, J=8 items	Covariance	-0.07	-0.06	-0.03	-0.05	-0.05	-0.06
N=1000, True gain=0.2, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=1000, True gain=0.2, J=8 items	Slope Mean	0.20	0.20	0.14	0.17	0.17	0.20
N=1000, True gain=0.2, J=8 items	Intercept Variance	0.47	0.46	0.23	0.35	0.35	0.47
N=1000, True gain=0.2, J=8 items	Slope Variance	0.08	0.08	0.04	0.06	0.06	0.08
N=1000, True gain=0.2, J=8 items	Covariance	-0.07	-0.06	-0.04	-0.05	-0.05	-0.06
N=1000, True gain=0.5, J=8 items	Intercept Mean	0.00	0.00	0.01	0.00	0.00	0.00

N=1000, True gain=0.5, J=8 items	Slope Mean	0.50	0.50	0.31	0.39	0.41	0.51
N=1000, True gain=0.5, J=8 items	Intercept Variance	0.47	0.46	0.23	0.29	0.33	0.47
N=1000, True gain=0.5, J=8 items	Slope Variance	0.08	0.08	0.03	0.05	0.05	0.08
N=1000, True gain=0.5, J=8 items	Covariance	-0.07	-0.06	-0.05	-0.05	-0.05	-0.06
N=2000, True gain=0, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0, J=4 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0, J=4 items	Intercept Variance	0.47	0.48	0.24	0.31	0.31	0.46
N=2000, True gain=0, J=4 items	Slope Variance	0.08	0.08	0.04	0.05	0.05	0.08
N=2000, True gain=0, J=4 items	Covariance	-0.07	-0.07	-0.03	-0.04	-0.04	-0.06
N=2000, True gain=0.2, J=4 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0.2, J=4 items	Slope Mean	0.20	0.20	0.14	0.16	0.16	0.20
N=2000, True gain=0.2, J=4 items	Intercept Variance	0.47	0.48	0.24	0.30	0.31	0.46
N=2000, True gain=0.2, J=4 items	Slope Variance	0.08	0.08	0.04	0.05	0.05	0.08
N=2000, True gain=0.2, J=4 items	Covariance	-0.07	-0.07	-0.04	-0.05	-0.05	-0.06
N=2000, True gain=0.5, J=4 items	Intercept Mean	0.00	0.00	0.01	0.00	0.00	0.00
N=2000, True gain=0.5, J=4 items	Slope Mean	0.50	0.50	0.31	0.36	0.38	0.51
N=2000, True gain=0.5, J=4 items	Intercept Variance	0.47	0.48	0.24	0.27	0.29	0.46
N=2000, True gain=0.5, J=4 items	Slope Variance	0.08	0.08	0.04	0.05	0.05	0.08
N=2000, True gain=0.5, J=4 items	Covariance	-0.07	-0.07	-0.05	-0.05	-0.05	-0.06
N=2000, True gain=0, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0, J=8 items	Slope Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0, J=8 items	Intercept Variance	0.47	0.48	0.24	0.37	0.37	0.48
N=2000, True gain=0, J=8 items	Slope Variance	0.08	0.08	0.04	0.06	0.06	0.08
N=2000, True gain=0, J=8 items	Covariance	-0.07	-0.07	-0.04	-0.06	-0.06	-0.07
N=2000, True gain=0.2, J=8 items	Intercept Mean	0.00	0.00	0.00	0.00	0.00	0.00
N=2000, True gain=0.2, J=8 items	Slope Mean	0.20	0.20	0.13	0.17	0.17	0.20
N=2000, True gain=0.2, J=8 items	Intercept Variance	0.47	0.48	0.24	0.36	0.36	0.48
N=2000, True gain=0.2, J=8 items	Slope Variance	0.08	0.08	0.04	0.06	0.06	0.08
N=2000, True gain=0.2, J=8 items	Covariance	-0.07	-0.07	-0.04	-0.06	-0.06	-0.07
N=2000, True gain=0.5, J=8 items	Intercept Mean	0.00	0.00	0.01	0.00	0.00	0.00
N=2000, True gain=0.5, J=8 items	Slope Mean	0.50	0.50	0.31	0.39	0.41	0.51

N=2000, True gain=0.5, J=8 items	Intercept Variance	0.47	0.48	0.23	0.30	0.34	0.48
N=2000, True gain=0.5, J=8 items	Slope Variance	0.08	0.08	0.04	0.05	0.06	0.08
N=2000, True gain=0.5, J=8 items	Covariance	-0.07	-0.07	-0.05	-0.05	-0.06	-0.07

Table A3

Growth Mindset Items from the District Survey

Item Wording
My intelligence is something that I can't change very much.
Challenging myself won't make me any smarter.
There are some things I am not capable of learning.
If I am not naturally smart in a subject, I will never do well in it.

Note. All items are on a 5-point Likert (Agree or disagree) scale.