Measuring Multidimensional Growth–A Higher-Order IRT Perspective

Steven W. Nydick

Pearson VUE

Chun Wang

University of Minnesota

Xinhui Xiong

CTB/McGraw-Hill

-

Author Note

Paper presented at the Annual Meeting of the American Educational Research Association (AERA) in Philadelphia, PA, April 5–7, 2014.

Abstract

Growth models have generally fallen into two camps: (1) longitudinal item response theory models; and (2) latent growth curve models. Psychometricians recently combined both models into a multilevel model with categorical outcomes, multiple latent traits at several time points, and individual growth parameters. The current study examines an extension of multilevel IRT growth to hierarchical IRT models using the SEM formulation. Conditions varied include correlation between latent traits, items loading on each dimension, and number of simulees. For each condition, item, person, and growth parameters are compared when using one of several model formulations or estimation algorithms. Mplus code is provided.

Measuring Multidimensional Growth–A Higher-Order IRT Perspective

Introduction

Recently, the federal government instituted a program entitled "Race to the Top" to encourage schools to "build data systems that measure student growth and success" (U.S. Department of Education, 2009, p. 2). The easiest method of estimating growth, comparing raw test scores before and after a period of learning, has been criticized by psychometricians as unreliable (e.g., Cronbach & Furby, 1970; Kim-Kang & Weiss, 2008). Researchers have developed models to account for the unreliability of raw scores by positing a latent trait imperfectly measured by observable phenomena that changes over time. These "growth" models have, until recently, fallen into two camps: (1) longitudinal item response theory (IRT) models (e.g., Andersen, 1985; Embretson, 1991; von Davier, Xu, & Carstensen); and (2) latent growth curve models (e.g., Bollen & Curran, 2006; Duncan, Duncan, & Strycker, 2006; Hancock & Lawrence, 2006). Item response theory directly models latent growth through a set of ordinal variables, such as responses to test items, but cannot easily account for varying growth structures or data collection methods. Latent growth curve (LGC) methods separate the measurement and growth parts of a model but do not easily consider ordinal outcome variables.

Recent work on measuring academic growth has combined the IRT measurement model with the LGC second level growth model (e.g., McArdle, 1988) and extended the assessment of growth to multidimensional IRT models (e.g., Hsieh, von Eye, & Maier, 2010). Multidimensional IRT models posit that item responses are probabilistically determined by multiple, possibly correlated, latent traits. Unfortunately, neither unidimensional nor multidimensional IRT models adequately capture the higher-order nature of learning, whereby general ability (such as math aptitude) informs domain-specific abilities (such as algebra, geometry, calculus, or subsets thereof). Unidimensional IRT models measure an overall trait but ignore the sub-traits, whereas multidimensional IRT models ignore the overall trait. To circumvent limitations of prior models, de la Torre and

Song (2009) proposed a higher-order item response theory model (HO-IRT) that captures the overall and domain-specific abilities required in formative learning. By positing a higher-order structure, the HO-IRT model has been shown to measure domain abilities and estimate item parameters better than the typical multidimensional IRT model. The objective of this study is to propose a new IRT model for measuring growth by combining the HO-IRT measurement model with the latent growth curve structure and to compare the multilevel formulation with standard multidimensional IRT models in capturing underlying, domain-specific abilities. Each model was estimated using the SEM formulation (Múthen & Múthen, 1998–2012), and corresponding Mplus code is provided.

Growth Models in IRT

Measurements of ability and achievement should always be tailored to the goals of an assessment. If determining overall ability at the end of an instructional program, such as knowing whether someone is competent to practice medicine, then practitioners would construct a summative assessment. Summative assessments are designed to adequately measure one or more broad domains of knowledge, such as surgical ability, patient care, or biological knowledge. In contrast to summative assessments, formative assessments provide feedback throughout the instructional period. Because formative assessments are generally used to inform instructional decisions, these assessments should adequately measure finer-grained, domain-specific abilities in addition to overall knowledge.

Formative assessments assume a hierarchical structure to knowledge and thus require a hierarchical measurement model. To this end, the higher-order IRT (HO-IRT) model was developed to improve the reliability of numerous sub-scores by introducing a higher-order overall score (de la Torre & Song, 2009). Psychometricians can use the HO-IRT model to simultaneously estimate item parameters, overall ability, and domain-specific abilities. de la Torre and Hong (2010) demonstrated that the HO-IRT model estimates item parameters more accurately than either unidimensional or single-level multidimensional IRT models.

The simplest HO-IRT model contains two levels: (1) a link between a single overall ability and one of several domain abilities; and (2) a probabilistic relationship between each domain abilities and items designed to measure only one domain. Specifically, let θ represent the overall ability underlying responses to test items and ξ represent the higher-order trait. Then one can hypothesize that

$$\theta_{ik} = \lambda_k \xi_i + \epsilon_{ik} \tag{1}$$

where ξ_i is the overall ability of examinee i, θ_{ik} represents domain-specific ability $k \in \{1, \ldots, K\}$ for examinee i, and λ_k indicates the relationship between domain-specific ability and overall ability. Moreover, the probability of examinee i correctly responding to item j on domain k is defined by the following item response function (IRF):

$$p_{jk_j}(\theta_{ik}) = \Pr(Y_{ijk_j}|\theta_{ik}, a_{jk_j}, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp[-a_{jk_j}(\theta_{ik} - b_j)]}, \tag{2}$$

where a_{jk_j}, b_j, and c_j represent the discrimination, difficulty, and pseudo-guessing parameters of the j^{th} item measuring the k^{th} domain.

Extending the HO-IRT model across several time points allows one to detect individual growth by modeling change in the higher-order trait, a principal goal of formative assessment. The general framework for extending the HO-IRT model across several time points is relatively simple. First, propose a set of abilities for each person across all time points, which one can organize in matrix form:

$$\mathbf{\Gamma}_i = \begin{bmatrix} \xi_i^{(1)} & \theta_{i1}^{(1)} & \theta_{i2}^{(1)} & \cdots & \theta_{iK}^{(1)} \\ \xi_i^{(2)} & \theta_{i1}^{(2)} & \theta_{i2}^{(2)} & \cdots & \theta_{iK}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \xi_i^{(T)} & \theta_{i1}^{(T)} & \theta_{i2}^{(T)} & \cdots & \theta_{iK}^{(T)} \end{bmatrix}$$

Second, administer a set of items to each person at each time point that satisfy

parameter identifiability assumptions. Finally, use a model-based estimation method, such as MCMC or EM, to estimate unknown item parameters, estimate person parameters, and determine change. Note that the above ability matrix might be sparse due to some domain-specific abilities only manifesting, and thus being linked to the overall ability, at certain ages. For instance, fractional subtraction ability might not appear until grade 4 or 5 due to students theretofore not learning those skills.

One could facilitate estimation of a longitudinal HO-IRT model by positing a growth trajectory at either the domain or the general ability levels. Because domain abilities relate to the general ability, they would also be carried along by any growth in the general ability. Assume that a person-specific linear relationship exists between time and general ability. Then the general ability for person i at time-point t, $\xi_i^{(t)}$, can be written as a deviation from the person-specific regression line,

$$\xi_i^{(t)} = \pi_{0i} + \pi_{1i} \times (t-1) + \delta_i^{(t)}. \tag{3}$$

These individual intercept and slope parameters can be written as deviations from the overall average intercepts and slopes, or

$$\pi_{0i} = \beta_0 + \nu_{0i} \tag{4}$$

$$\pi_{1i} = \beta_1 + \nu_{1i}. \tag{5}$$

Given $\xi_i^{(t)} = \pi_{0i} + \pi_{1i} \times (t-1) + \delta_i^{(t)}$, a domain-specific ability for person i at time-point t, $\theta_{ik}^{(t)}$, would be predicted to systematically change over time,

$$\begin{aligned}\theta_{ik}^{(t)} = \lambda_k \xi_i^{(t)} + \epsilon_{ik}^{(t)} &= \lambda_k(\pi_{0i} + \pi_{1i} \times (t-1) + \delta_i^{(t)}) + \epsilon_{ik}^{(t)} \\ &= \lambda_k \pi_{0i} + \lambda_k \pi_{1i} \times (t-1) + (\lambda_k \delta_i^{(t)} + \epsilon_{ik}^{(t)}) \\ &= \zeta_{0ki} + \zeta_{1ki} \times (t-1) + \upsilon_{ik}^{(t)}. \end{aligned} \tag{6}$$

Domain abilities relate to individual items by the item response model described in Equation (2) (with $\theta_{ik}^{(t)}$ replacing θ_{ik} and potentially unique items/item parameters at each time-point). Note that the model described in Equations (3)–(6) is restrictive in that domain-specific abilities are only purported to systematically grow via their relationship to the general ability. In the next section, we describe a study designed to determine the most efficient and accurate method of estimating the longitudinal HO-IRT model using popular Mplus software.

Methods and Data

The current study explores Mplus estimation of a longitudinal version of the HO-IRT model in an extensive simulation study. This simulation consisted of testing how Mplus recaptured parameters from two longitudinal models, one with $T = 2$ time points, and one with $T = 4$ time points. The generation of simulees and parameters for each model are slightly different and will be separately explained.

Conditions

We performed two simulations, one with $T = 2$ and one with $T = 4$. Each simulation assumed that simulees had a higher-order ability parameter that changed over time and related (via factor analysis) to several lower-order abilities. These lower-order abilities were then used to predict responses to items via multidimensional item response theory.

For the $T = 2$ simulation, we varied sample size ($N \in \{250, 1000\}$), total number of items loading on each dimension ($J_k \in \{10, 20\}$), total number of dimensions ($K \in \{3, 5\}$),

the correlation between higher-order person parameters ($r \in \{.50, .75\}$, and the loading of the higher-order person parameter on all of the lower-order person parameters ($\lambda \in \{.8, .9\}$).

For the $T = 4$ simulation, we varied the exact same conditions as when $T = 2$, but we imposed a linear form on higher-order growth rather than assuming that higher-order person parameters correlated a particular amount.

In either case, we simulated $R = 25$ replications for each of the 32 conditions (if $T = 2$) or 16 conditions (if $T = 4$) and aggregated results (using the median rather than the mean) across the replications.

Item Parameter Generation

For all simulations, item parameters were generated according to the two-parameter, compensatory, multidimensional IRT model with K dimensions and assuming simple structure. That is, we assumed that the IRT model was of the form

$$p_{jk_j}(\theta_{ik}) = \Pr(Y_{ijk_j}|\theta_{ik}, a_{jk_j}, b_j) = \frac{1}{1 + \exp[-a_{jk_j}(\theta_{ik} - b_j)]}, \tag{7}$$

where k_j indicates the dimension on which item j loads, a_{jk_j} represents the corresponding discrimination parameter, and b_j denotes the item difficulty. If parameterizing the model with threshold rather than difficulty, simply set $d_j = -a_{jk_j}b_j$. Due to simple structure, if J_k items loaded onto each dimension, then the total number of items would be $J = J_k \times K$. The reason that we used a two-parameter rather than three-parameter IRT model is because Mplus cannot currently estimate the lower-asymptote of Equation (2). Moreover, different parameter sets were generated for every single replication within condition.

Person Parameter Generation

Person parameters were generated somewhat differently for each simulation and will thus be separately explained in the following subsections.

$T = 2$. Let $\xi_i^{(t)}$ be the i^{th} persons higher-order ability parameter at time point t. Then $\boldsymbol{\xi}_i = [\xi_i^{(1)}, \xi_i^{(2)}]^T \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\xi}} = [0.0, 0.3]^T, \boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \left[\begin{smallmatrix} 1 & r \\ r & 1 \end{smallmatrix}\right]\right)$. Next, let $\theta_{ik}^{(t)}$ be the i^{th} persons lower-order ability parameter for dimension k at time point t. Then $\theta_{ik}^{(t)} = \lambda \xi_i^{(t)} + \epsilon_{ik}^{(t)}$, where $\epsilon_{ik}^{(t)} \sim N(\mu_\epsilon = 0.0, \sigma_\epsilon^2 = 1 - \lambda^2)$. Notice that this data generation method had each lower-order person parameter at time t relate identically to the corresponding higher-order person parameter. Moreover, these relationships were not assumed to change over time.

$T = 4$. For the $T = 4$ simulation, we first simulated a set of linear model parameters, π_0 and π_1, and then used the linear parameters to generate higher-order abilities. The intercept parameter, π_0, was generated to be normally distributed with mean $\mu_{\pi_0} = 0.0$ and variance $\sigma_{\pi_0}^2 = 0.5$. The slope parameter, π_1, was generated to be normally distributed with mean $\mu_{\pi_1} = 0.25$ and variance $\sigma_{\pi_1}^2 = 0.01$. Given π_{0i} and π_{1i} for person i, higher-order person parameters were then set to

$$\xi_{it} = \pi_{0i} + \pi_{1i} \times (t - 1) + \delta_{it},$$

where $\delta_{it} \sim N(\mu_\delta = 0.0, \sigma_\delta^2 = 0.10)$. Given a set of higher-order person parameters, lower-order parameters were then generated using exactly the same algorithm as in the $T = 2$ simulation.

Overall Simulation

For each simulation, we generated a response matrix of size $N \times (JT)$, where each item was taken by each simulee across all time points. Responses were generated by determining the probability of a correct response using Equation (7) and then setting the response equal to 1 if a random uniform deviate was less than that probability and 0 otherwise. Given a particular condition, a script was then generated so that Mplus could estimate each model using the response matrix. The Mplus code was specified somewhat differently for each simulation. In all cases, we used Mplus's `MONTECARLO` integration routine for the robust maximum likelihood estimator.

$T = 2$. For the $T = 2$ simulation, we estimated item and person parameters in two steps. We first jointly estimated item and person parameters across all time points and then separately estimated item and person parameters at each time point and linked them to the first time point using methods discussed in a different document. The following Mplus code indicates some of the important assumptions for the joint calibration in the `MODEL` section of the Mplus script given $K = 3$ (although see Appendix B for a complete example Mplus script that we used).

```
xi1 BY th1_1-th3_1*.8 (lamb);
xi2 BY th1_2-th3_2*.8 (lamb);
[th1_1-th3_1@0];
[th1_2-th3_2*0.2];
th1_1-th3_1@.5 th1_2-th3_2@.5;
[xi1@0];
[xi2*0.5];
xi1@1;
xi2*1;
xi1-xi2 WITH xi1-xi2;
```

That is, we assumed that the higher-order person parameter at each time point loaded onto all corresponding lower-order person parameters with one loading, λ, which we initialized to .8 for each dimension. Moreover, we assumed that the mean of higher-order and lower-order person parameters at the first time point was equal to 0, but we let the mean of higher-order and lower-order person parameters at the second time point vary. Finally, the variance of domain person parameters was set to 0 for all $t \in \{1, 2\}$ and $k \in \{1, 2, 3, 4, 5\}$, the variance of higher-order person parameters was set to 1 for $t = 1$, and the remaining parameters were free to vary.

The separate calibrations were very similar to the joint calibration with a few exceptions. The following Mplus code indicates some of the important assumptions for each separate calibration in the `MODEL` section of the Mplus script given $K = 3$ (although see Appendix C for a complete example Mplus script that we used).

```
xi1 BY th1_1-th3_1*.8 (lamb);
[th1_1-th3_1@0];
th1_1-th3_1@.5;
[xi1@0];
xi1@1;
```

Thus, for the separate calibration, we fixed all of the person parameter means and variances and assumed that the higher-order person parameter loaded onto the lower-order person parameters with one loading, λ, which we initialized to .8. Unlike the joint calibration, we could not constrain all of the separately calibrated loadings to be the same. We then linked parameters from $t = 2$ to the scale of parameters from $t = 1$ using a standard IRT linking method (described in Appendix A).

$T = 4$. For the $T = 4$ simulation, we simply separately estimated item and person parameters at each time point (using the specifications of the previous subsection) and then linked them onto the first time point using methods discussed Appendix. We then estimated π_0 and π_1 for each person using the `lme4` (Bates, Maechler, Boker, & Walker, 2013) package in `R` (R Core Team, 2013). Specifically, we ran the model

```
lmer( xi ~ time + (time | person) )
```

where `xi` represents a higher-order person parameter and can be predicted given a fixed effect of `time` and a random effect of `time` for each person. The intercept and slope per person was then simply set to the sum of the fixed effects across all people and the random effect of intercept and slope for that person.

Simulation Results

$T = 2$

This section briefly presents results for the conditions when $T = 2$. For all of the tables presented in this section, we calculated the Mean Squared Error (MSE) and correlation between generated and estimated person parameters and then aggregated

statistics within a desired condition.[1] We chose only to display results for person parameters and for those conditions with $K = 3$. These choices were made to simplify presentation. A full set of results are available in an online appendix at http://www.tc.umn.edu/~nydic001/.

INSERT TABLE 1 ABOUT HERE

Table 1 displays the average MSE for different correlation conditions. Correlation in this case simply refers to the specified correlation between $\xi^{(1)}$ and $\xi^{(2)}$. The first two rows of Table 1 indicate the average MSE for those conditions calibrated using a combined method, whereas the bottom two rows of Table 1 indicate the average MSE when parameters were separately calibrated at each time point. When using the combined calibration method, the MSE is typically smaller across person parameters if $r = .75$ as compared to $r = .50$. This difference is most dramatic for the higher-order parameters than the lower-order parameters. If separately calibrating conditions at each time point, then increasing the correlation between $\xi^{(1)}$ and $\xi^{(2)}$ actually results in a slight increase in the MSE for all estimated person parameters. Yet, despite the noticeable, systematic trends, differences between conditions are rather slight.

INSERT TABLE 2 ABOUT HERE

Results from Table 1 are reinforced by those presented in Table 2, which displays the average correlation between true and estimated person parameters for different correlation conditions. The correlation between true and estimated person parameters is slightly higher, on average, if $r = .75$ as compared to $r = .50$ when using combined calibration, and the correlation between true and estimated person parameters is slightly higher when using combined calibration rather than separate calibration. However, few people would consider a .01 difference between conditions meaningful in any sense.

[1] Technically, we calculated the MSE using the equation $\frac{\sum(\hat{\gamma}-\gamma)^2}{N}$, where γ is the desired parameter, which is estimated by $\hat{\gamma}$, and summation was taken across the entire sample of N simulees. These statistics were then aggregated, using the median rather than the mean, across the set of 25 replications and then averaged across the desired conditions.

INSERT TABLE 3 ABOUT HERE

INSERT TABLE 4 ABOUT HERE

Tables 3 and 4 present the average MSE and correlation, respectively, when aggregating across the pre-specified factor loading, λ. Notice that when $\lambda = .90$, then the correlation is slightly improved for all person parameters as compared to $\lambda = .80$, whereas the MSE is improved for the higher-order trait (ξ) but worsened for the lower-order trait (θ). However, the difference between combined and separate calibrations are, as before, rather slight. Unsurprisingly, using a combined calibration method results in slightly smaller MSE (and larger correlations) when comparing true and estimated person parameters. But this difference is often only in the third decimal place.

INSERT TABLE 5 ABOUT HERE

INSERT TABLE 6 ABOUT HERE

The final set of tables, Tables 5 and 6, indicate the MSE and correlation (when comparing true and estimated person parameters) if aggregating among items within dimension. Recall that we generated items with between-item multidimensionality, in that any item could only load on one dimension. If $J_K = 10$, then each θ_k is measured by a total of 10 items unique to that particular trait. If $J_K = 20$, then all person parameters are better estimated than when $J_K = 10$. This finding is unremarkable. When more items load onto a given dimension, one should be able to better estimate the corresponding trait. Yet as before, joint calibration did not drastically improve the estimation accuracy. And this slight improvement certainly does not justify the additional computing time (sometimes on the order of hours for joint calibration rather than minutes for separate calibrations) of Mplus. Results from those conditions with $T = 2$ and $K = 5$ are similar to those presented when $K = 3$, and, thus, will not be presented in this section. However, anyone seeking a full set of results should see http://www.tc.umn.edu/~nydic001/.

The next section presents the MSE and correlation between true and estimated person parameters when generating data to fit a linear model with $T = 4$. Because the joint calibration did not improve the precision of parameter estimates, we chose to estimate all conditions separately at each time point and then link parameters at $T \geq 2$ back to the estimated parameters at $T = 1$ using the methods described in Appendix A.

$T = 4$

This section briefly presents results for the conditions when $T = 4$. In contrast to the previous section, results from this section are only presented in graphical form. This decision was made to ease interpretation. A full set of results are available in an online appendix at http://www.tc.umn.edu/~nydic001/.

INSERT TABLE 1 ABOUT HERE

Figure 1 displays the bias (upper panels) and MSE (lower panels) for various factor loading conditions. The left plot presents the higher-order factor, ξ, whereas the right plot presents the domain factors, θ. Points within a plot indicate the average bias or MSE for a given person parameter at each of the four time points after linking parameters at the upper time points to parameters at the first time point. Average bias was calculated by subtracting the true/generated parameters from the estimated parameters, averaging across all persons, and then taking the median across replications. One can find two, fairly obvious, trends as pertaining to the conditions presented in Figure 1. When $\lambda = .9$, an increase in time also yielded a concurrent increase in bias and MSE for both domain and higher-order traits. Note that λ relates the higher-order ξ to the domain-level θs. This result is entirely an artifact of scaling in estimation. As it turns out, the scale that we imposed on ϵ as being $\sigma^2_\epsilon = 1 - \lambda^2$ did not hold for $\sigma^2_\epsilon \approx 0$. Therefore, the scale of ξ at a given time point was not equal to the scale of θ at that time point, so that the linking applied to θ did not also apply to ξ. Moreover, the linking method yielded negatively biased item-slope parameters and positively biased domain person parameters (though the

item-intercept parameters were still accurately estimated). As shown in Figure 1, this scaling problem does not appear to be noticeable for $\lambda = .8$ (except, perhaps, for domain person parameters at $t \geq 3$). Therefore, the remaining results will be presented solely in terms of correlation between true and estimated parameters rather than MSE/bias. Note that the bias and MSE is small for those conditions with $\lambda = .8$ and noticeably large for those conditions with $\lambda = .9$. Thus, the linking method described in Appendix A does not appear to work well for $\lambda \approx 1$ but does appear to work well for $\lambda = .8$.

INSERT TABLE 2 ABOUT HERE

Figure 2 displays the correlation between true and estimated person parameters for $K = 3$ (upper panels) and $K = 5$ (lower panels) for those conditions presented in Figure 1. One can spot a few trends suggested by Figure 2. First, higher λ always results in a larger correlation between true and estimated parameters than lower λ. The blue triangles are always higher up than the green circles in all of the plots. Second, more factors results in larger correlations between true and estimated person parameters regardless of whether looking at higher-order (left panels) or domain-level (right panels) abilities. Finally, true and estimated person parameters tend to have slightly higher correlations for all factors and conditions as time increases. The latter result is due to larger variability among true person parameters at later values of time. We originally tried a similar study with much lower variability among person parameters at $t = 1$, and the resulting estimates poorly correlated with the true person parameters.

INSERT TABLE 3 ABOUT HERE

Figure 3 presents the correlation between true and estimated person parameters when varying J_K, the number of items loading on each dimension. Unsurprisingly, fewer items loading on a particular domain-level trait resulted in poorer estimates of that trait. The green circles are always below the blue triangles. Interestingly, this effect is slightly relieved

for the higher-order trait if more domain abilities load onto the higher-order trait. Notice that the green dots are slightly closer to the respective blue triangles in the lower left quadrant of Figure 3 than they are in the upper right quadrant (whereas domain abilities are further apart when comparing the same conditions). Thus, accurate estimates of higher-order traits can be obtained even with inaccurate estimates of lower-order traits as long as many lower-order traits load onto the higher-order trait.

INSERT TABLE 4 ABOUT HERE

Finally, consider Figure 4, which presents the correlation between true and estimated person parameters when varying N, the number of simulees. When closely examining Figure 4, one can see that, unlike the previous two plots, the green dots and blue triangles are actually fairly close in all quadrants of Figure 4. Therefore, varying the number of simulees did not have much of an effect on the ultimate estimation precision of person parameters. However, the blue triangles and green dots are still somewhat separated. Therefore, varying the number of simulees did have some effect on the ultimate estimation precision of person parameters. The reason for the small but noticeable effect of N is that sample size directly impacts the accuracy of item parameter estimation, so that poorer item parameter estimates propagates through to the person parameter estimates. This effect is incidental rather than direct, so that the ultimate cost in estimation precision is small.

Conclusion

Many teachers, administrators, and government employees require the measurement of student growth. Teachers can use estimated growth to modify lesson plans based on strategies of improvements. Administrators can use estimated growth to examine school performance and help make budgetary decisions. In either case, one must accurately estimate student growth across several, possibly correlated, ability dimensions. This current paper presents a realistic, formative model of growth across several sub-domains and determines the accuracy and efficiency of estimating the model with Mplus. As shown

in the previous section, Mplus accurately estimates person parameters over time as long as the relationship between the higher and lower order traits is small enough to prevent difficulties in scaling the residual variance. Of course, this scaling problem might be alleviated by either different model constraints, alternative methods of linking parameter estimates across time, or concurrent calibration methods. Yet, as shown for the conditions with $T = 2$, concurrently calibrating all of the parameters at all time points did not provide a large increase in estimation precision. We had already attempted to estimate the complete model for $T = 4$, and Mplus took at least four hours and had difficulty converging. Breaking down the problem into separate calibrations at each time point turned a complex problem, namely estimating item and person parameters at each time point with constraints on the item parameters, the relationships, and the scaling, into several more tractable, higher-order IRT models. Maximum likelihood via the EM-algorithm is known to converge very slowly in many applications (e.g., Meng & van Dyk, 1997), and alternative methods, such as Markov Chain Monte Carlo (MCMC; e.g., Patz & Junker, 1999), might retain tractability in estimation without sacrificing model specification. However, regardless of estimation method, constructing and estimating longitudinal IRT models should improve the measurement of educational outcomes, and thus, provide educators with the tools they need to better help students learn.

References

Anderson, E. B. (1985) Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Baldwin, S. G., Baldwin, P., & Nering, M. L. (2007). *A comparison of IRT equating methods on recovering item parameters and growth in mixed-format tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Douglas Bates, Martin Maechler, Ben Bolker and Steven Walker (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5. http://CRAN.R-project.org/package=lme4

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley-Interscience.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"–or should we? *Psychological Bulletin*, *74*, 68–80.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT model approach. *Applied Psychological Measurement*, *34*, 267–285.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*, 620–639.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (Eds.). (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.

Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Muller (Eds.), *Structural equation modeling: A second course* (pp. 171–196). Greenwood, CT: Information Age Publishing, Inc.

Hsieh, C.-A., von Eye, A. A., & Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: The dynamic association between adolescents' social isolation and engagement with

delinquent peers in the national youth survey. *Multivariate Behavioral Research*, *45*, 508–552.

Kim-Kang, G., & Weiss, D. J. (2008). Adaptive measurement of individual change. *Zeitschrift für Psychologie/Journal of Psychology*, *216*, 49–58.

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

Meng, X.-L., & van Dyk, D. (1997). The EM algorithm–An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B (Methodological)*, *59*, 511–567.

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus user's guide. Sixth edition.* Los Angeles, CA: Muthén & Muthén.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

U. S. Department of Education (2009). *Race to the top program executive summary.* Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318–336.

Table 1

Average Mean Squared Error (MSE) for estimates of ξ and θ aggregating within correlation conditions when generating data from a model with $T = 2$ and $K = 3$. The upper half presents results when calibrating all of the parameters together, whereas the lower half presents results when separately calibrating parameters at each time point.

	r	$\xi^{(1)}$	$\xi^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_1^{(2)}$	$\theta_2^{(2)}$	$\theta_3^{(2)}$
Combined	.50	0.24	0.27	0.40	0.39	0.40	0.42	0.41	0.42
	.75	0.21	0.24	0.38	0.40	0.38	0.40	0.40	0.39
Separate	.50	0.25	0.27	0.38	0.37	0.38	0.41	0.40	0.41
	.75	0.25	0.27	0.39	0.39	0.39	0.43	0.43	0.42

Table 2

Average correlation between true and estimated ξ and θ aggregating within correlation conditions when generating data from a model with $T = 2$ and $K = 3$. The upper half presents results when calibrating all of the parameters together, whereas the lower half presents results when separately calibrating parameters at each time point.

	r	$\xi^{(1)}$	$\xi^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_1^{(2)}$	$\theta_2^{(2)}$	$\theta_3^{(2)}$
Combined	.50	.88	.88	.88	.88	.88	.88	.88	.88
	.75	.89	.89	.88	.88	.89	.89	.89	.89
Separate	.50	.87	.87	.88	.88	.88	.88	.88	.88
	.75	.87	.87	.88	.88	.88	.88	.88	.88

Table 3

Average Mean Squared Error (MSE) for estimates of ξ and θ aggregating within factor loading conditions when generating data from a model with $T = 2$ and $K = 3$. The upper half presents results when calibrating all of the parameters together, whereas the lower half presents results when separately calibrating parameters at each time point.

	λ	$\xi^{(1)}$	$\xi^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_1^{(2)}$	$\theta_2^{(2)}$	$\theta_3^{(2)}$
Combined	.80	0.26	0.29	0.26	0.27	0.26	0.27	0.27	0.26
	.90	0.19	0.22	0.51	0.52	0.51	0.55	0.54	0.55
Separate	.80	0.29	0.31	0.27	0.27	0.27	0.28	0.29	0.29
	.90	0.20	0.23	0.50	0.50	0.50	0.55	0.54	0.54

Table 4

Average correlation between true and estimated ξ and θ aggregating within factor loading conditions when generating data from a model with $T = 2$ and $K = 3$. The upper half presents results when calibrating all of the parameters together, whereas the lower half presents results when separately calibrating parameters at each time point.

	λ	$\xi^{(1)}$	$\xi^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_1^{(2)}$	$\theta_2^{(2)}$	$\theta_3^{(2)}$
Combined	.80	.86	.86	.87	.87	.87	.87	.88	.87
	.90	.90	.91	.89	.89	.89	.89	.89	.89
Separate	.80	.84	.84	.87	.87	.87	.87	.87	.87
	.90	.89	.89	.89	.89	.89	.89	.89	.89

Table 5

Average Mean Squared Error (MSE) for estimates of ξ and θ aggregating within items on each dimension when generating data from a model with $T = 2$ and $K = 3$. The upper half presents results when calibrating all of the parameters together, whereas the lower half presents results when separately calibrating parameters at each time point.

	J_K	$\xi^{(1)}$	$\xi^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_1^{(2)}$	$\theta_2^{(2)}$	$\theta_3^{(2)}$
Combined	10	0.25	0.29	0.44	0.44	0.44	0.46	0.46	0.45
	20	0.19	0.22	0.34	0.34	0.34	0.36	0.36	0.36
Separate	10	0.28	0.31	0.43	0.43	0.43	0.46	0.47	0.46
	20	0.21	0.22	0.34	0.34	0.34	0.37	0.37	0.36

Table 6

Average correlation between true and estimated ξ and θ aggregating within items on each dimension when generating data from a model with $T = 2$ and $K = 3$. The upper half presents results when calibrating all of the parameters together, whereas the lower half presents results when separately calibrating parameters at each time point.

	J_K	$\xi^{(1)}$	$\xi^{(2)}$	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_1^{(2)}$	$\theta_2^{(2)}$	$\theta_3^{(2)}$
Combined	10	.87	.87	.86	.86	.86	.86	.86	.86
	20	.90	.90	.91	.91	.91	.91	.91	.91
Separate	10	.85	.85	.85	.85	.85	.85	.85	.85
	20	.89	.89	.91	.91	.91	.91	.91	.91

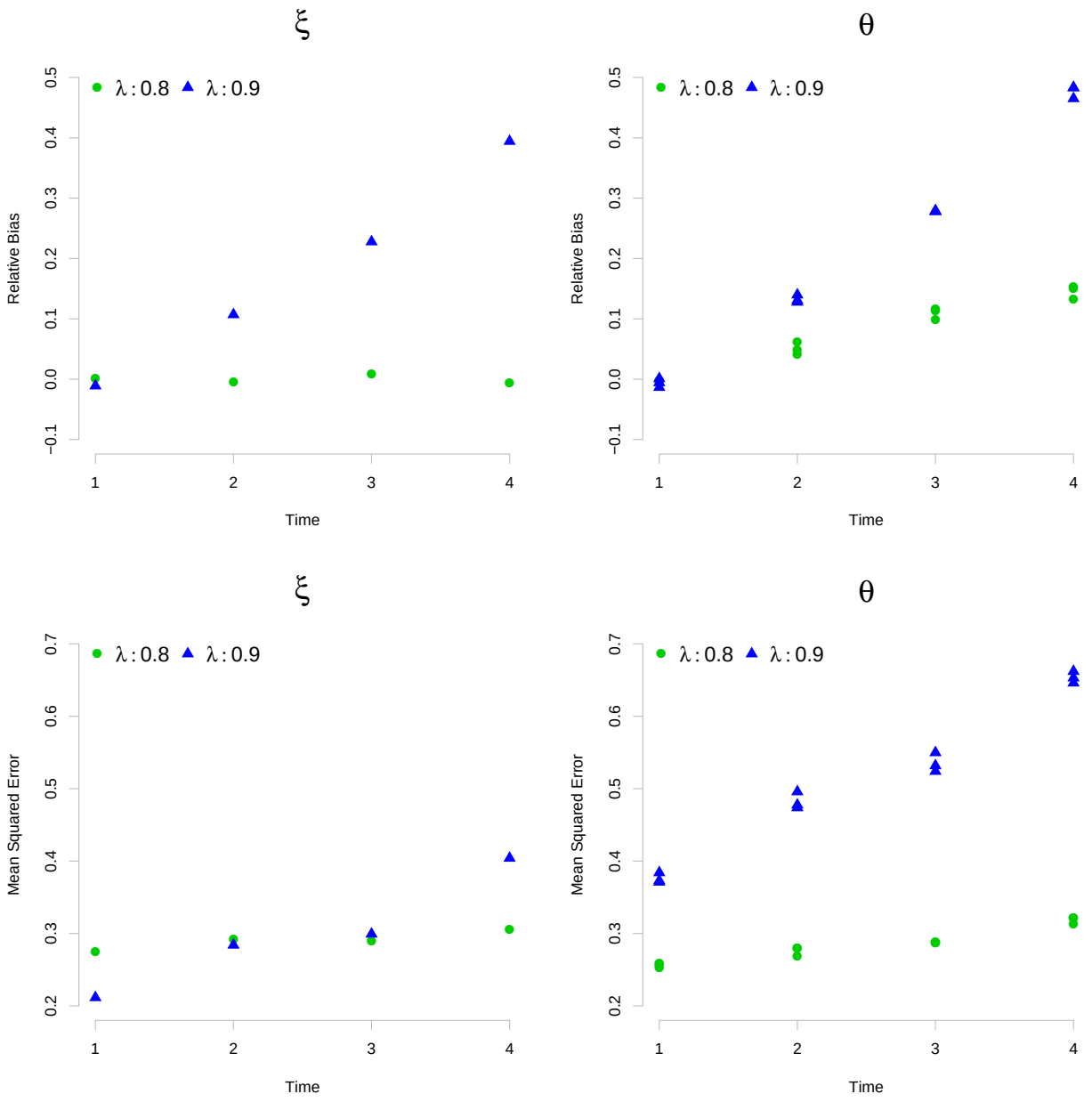

Figure 1. The bias and MSE when comparing true and estimated ξ and θ across time and aggregating within factor loading conditions if generating data from a model with $T = 4$ and $K = 3$. Note that the upper plot displays bias results when subtracting the true value of a person parameter from its estimated value.

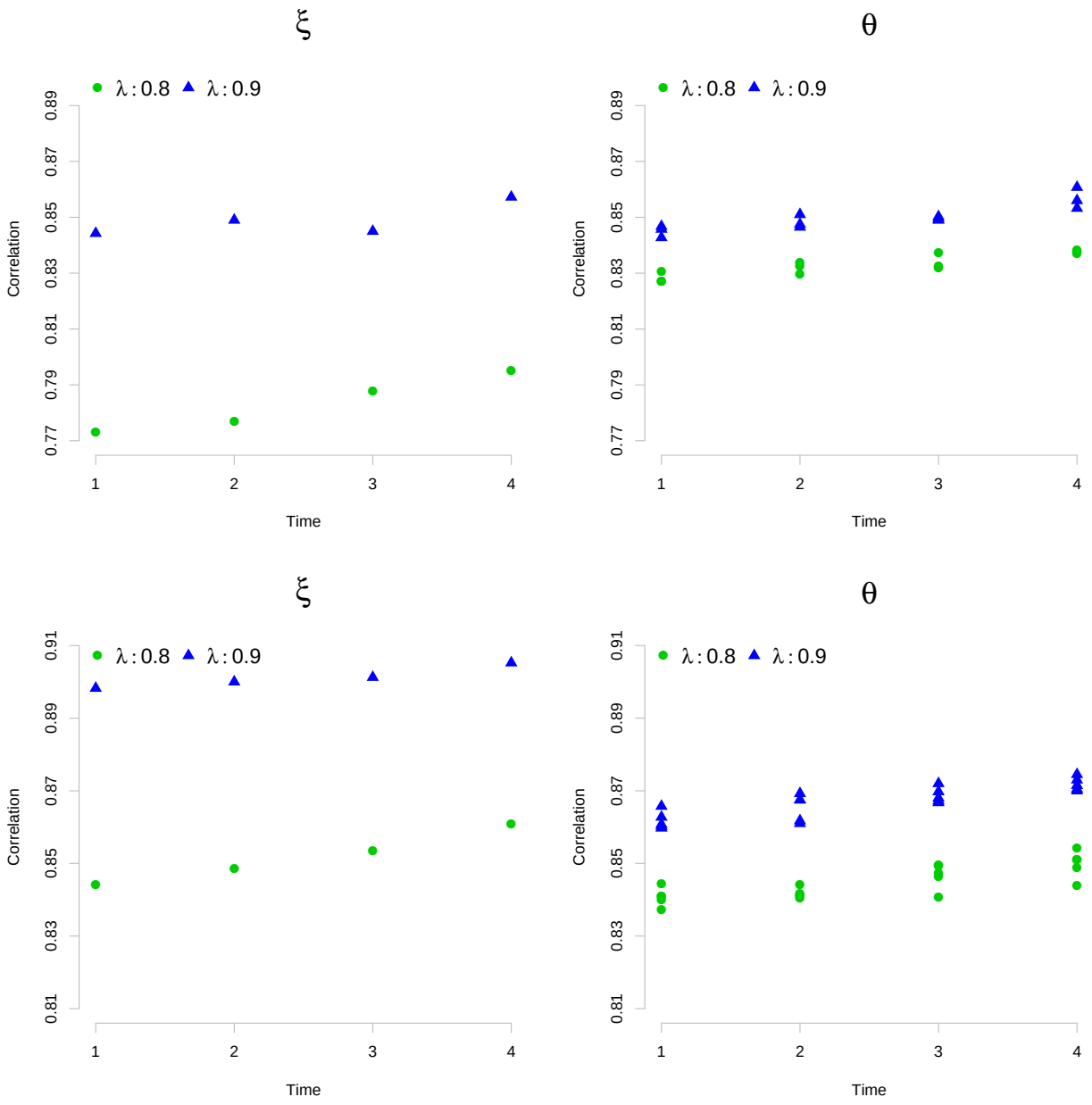

Figure 2. The correlation between true and estimated ξ and θ across time when aggregating within factor loading conditions if generating data from a model with $T = 4$ and $K = 3$ (upper plot) or $K = 5$ (lower plot).

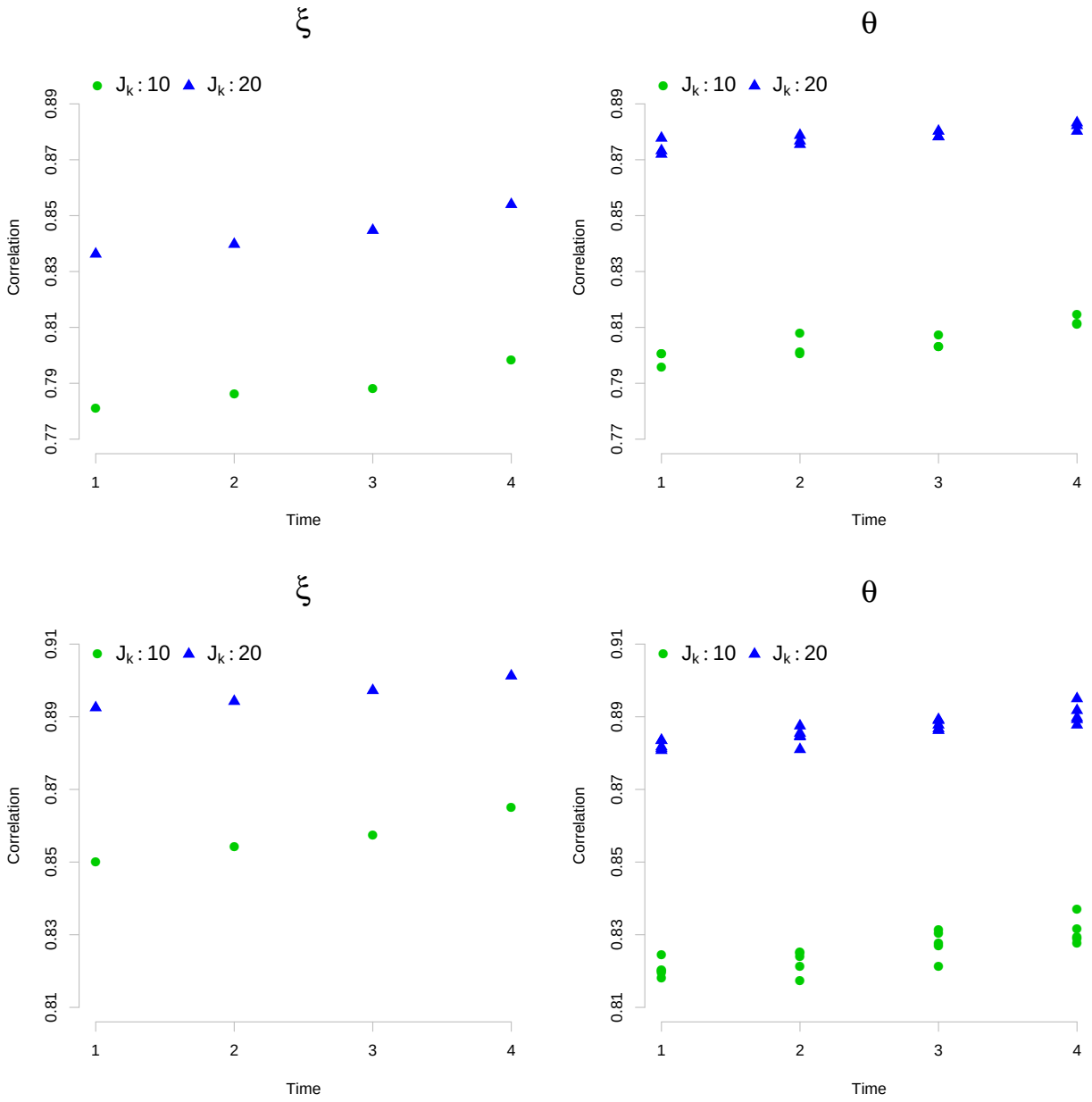

Figure 3. The correlation between true and estimated ξ and θ across time when aggregating within items on each dimension if generating data from a model with $T = 4$ and $K = 3$ (upper plot) or $K = 5$ (lower plot).

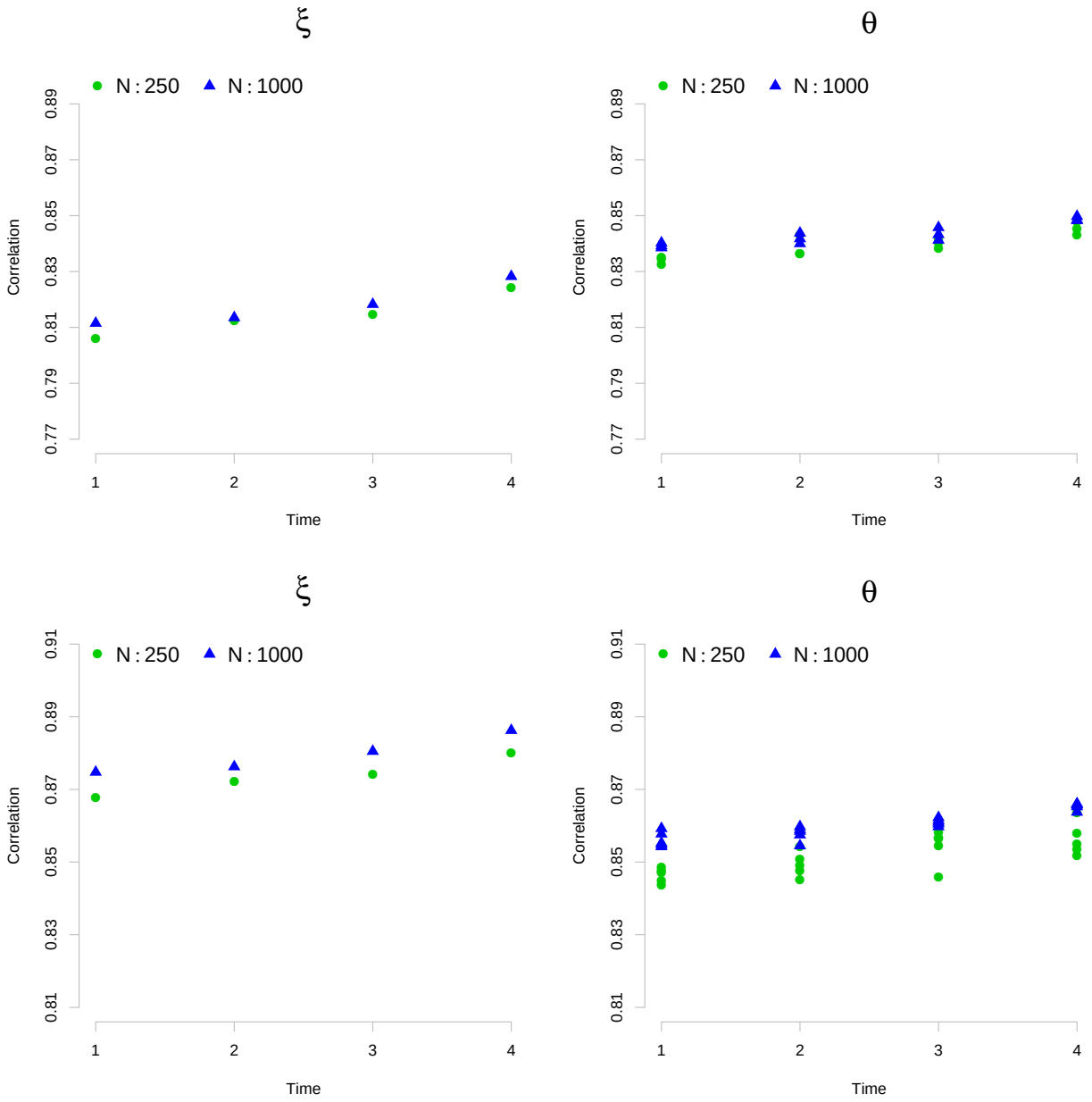

Figure 4. The correlation between true and estimated ξ and θ across time when aggregating within number of persons if generating data from a model with $T = 4$ and $K = 3$ (upper plot) or $K = 5$ (lower plot).

Appendix A

IRT Parameter Linking Equation

Items were linked onto the scale of the first time point using a form of mean/sigma equating. To explain this equating method, define the multi-unidimensional 2PL IRT model as

$$p_{ij} = \Pr(Y_{ijk_j} = 1|\theta_{ik}, a_{jk_j}, d_j) = \frac{1}{1 + \exp[-a_{jk_j}\theta_{ik} - d_j]}, \tag{8}$$

where a_{jk_j} is the j^{th} item discrimination parameter (measuring the latent trait along the k^{th} dimension), θ_{ik} is the k^{th} person parameter for the i^{th} person, and d_j is the j^{th} item threshold parameter. Item threshold parameters can be converted to difficulty parameters by dividing by each corresponding item discrimination parameter, which is the form shown in Equation (7).

Now assume that we have separately estimated parameters of Equation (8) at two time points and wish to link item parameters at the second time point to the scale of item parameters at the first time point. In other words, we have two sets of estimated parameters: $a^{(1)}_{jk_j}(j = 1, \ldots, J; k = 1, \ldots, K)$; $d^{(1)}_j(j = 1, \ldots, J)$; $\theta^{(1)}_{ik}(i = 1, \ldots, N; k = 1, \ldots, K)$ at time point 1, and $a^{(2)}_{jk_j}(j = 1, \ldots, J; k = 1, \ldots, K)$; $d^{(2)}_j(j = 1, \ldots, J)$; $\theta^{(2)}_{ik}(i = 1, \ldots, N; k = 1, \ldots, K)$ at time point 2. If all items loaded onto the same dimension, then one could apply standard IRT linking methods, such as mean/mean, mean/sigma, Stocking-Lord, or Haebara. If all items loaded onto multiple dimensions, then one must also rotate the loadings at the second time point to be oriented in the same direction as the loadings at the first time point. However, because all items load onto only one dimension, the dimensions are well-defined, and one can proceed with standard IRT linking methods.

For the actual linking, we decided to use the mean/sigma method. Mean/sigma has been shown relatively robust to models measuring growth (e.g., Baldwin, Baldwin, & Nering, 2007) and much easier to explain/implement than alternative options. To complete

the mean/sigma linking, we set $t = 1$ as the base distribution. Therefore, $\sigma_1 = \sqrt{\frac{\sum(b_j^{(1)} - \bar{b}_j^{(1)})^2}{J-1}}$, and $\mu_1 = \bar{b}_j^{(1)} = \sum b_j^{(1)}/J$ was assumed the desired standard deviation and mean of the item parameters at time point 2, where $b_j = -d_j/a_{jk_j}$. Next, let $\sigma_2 = \sqrt{\frac{\sum(b_j^{(2)} - \bar{b}_j^{(2)})^2}{J-1}}$ and $\mu_1 = \bar{b}_j^{(2)} = \sum b_j^{(2)}/J$. Then the equating slope would be $\beta_{1|2} = \frac{\sigma_1}{\sigma_2}$, and the equating intercept would be $\alpha_{1|2} = \mu_1 - \beta_{1|2}\mu_2$. Finally,

$$
\begin{aligned}
a_{jk_j}^{(2)\star} &= a_{jk_j}^{(2)}/\beta_{1|2} && (j = 1, \ldots, J; k = 1, \ldots, K) \\
\theta_{ik}^{(2)\star} &= \theta_{ik}^{(2)} \times \beta_{1|2} + \alpha_{1|2} && (n = 1, \ldots, N; k = 1, \ldots, K) \\
d_j^{(2)\star} &= -(b_j^{(2)} \times \beta_{1|2} + \alpha_{1|2}) \times a_{jk_j}^{(2)\star} && (j = 1, \ldots, J)
\end{aligned}
$$

If $t > 2$, then one must link parameters at each time point to the scale of parameters at the first time point by simply replacing the superscript (2) with the appropriate t.

Appendix B

Example Mplus Script for Joint Calibration and $T = 2$

```
 TITLE: Longitudinal Hierarchical IRT Estimation
 DATA: FILE IS hSim11111.dat;
 VARIABLE: NAMES ARE it1_1 it2_1 it3_1 it4_1 it5_1
it6_1 it7_1 it8_1 it9_1 it10_1
it11_1 it12_1 it13_1 it14_1 it15_1
it16_1 it17_1 it18_1 it19_1 it20_1
it21_1 it22_1 it23_1 it24_1 it25_1
it26_1 it27_1 it28_1 it29_1 it30_1
it1_2 it2_2 it3_2 it4_2 it5_2
it6_2 it7_2 it8_2 it9_2 it10_2
it11_2 it12_2 it13_2 it14_2 it15_2
it16_2 it17_2 it18_2 it19_2 it20_2
it21_2 it22_2 it23_2 it24_2 it25_2
it26_2 it27_2 it28_2 it29_2 it30_2;
 CATEGORICAL ARE it1_1 it2_1 it3_1 it4_1 it5_1
it6_1 it7_1 it8_1 it9_1 it10_1
it11_1 it12_1 it13_1 it14_1 it15_1
it16_1 it17_1 it18_1 it19_1 it20_1
it21_1 it22_1 it23_1 it24_1 it25_1
it26_1 it27_1 it28_1 it29_1 it30_1
it1_2 it2_2 it3_2 it4_2 it5_2
it6_2 it7_2 it8_2 it9_2 it10_2
it11_2 it12_2 it13_2 it14_2 it15_2
it16_2 it17_2 it18_2 it19_2 it20_2
it21_2 it22_2 it23_2 it24_2 it25_2
```

```
it26_2 it27_2 it28_2 it29_2 it30_2;
 ANALYSIS: TYPE = GENERAL;
 ESTIMATOR = MLR;
 LINK = LOGIT;
 INTEGRATION = MONTECARLO;
 MODEL: th1_1 BY it1_1* it2_1* it3_1* it4_1* it5_1* (f1 f2 f3 f4 f5)
     it6_1* it7_1* it8_1* it9_1* it10_1* (f6 f7 f8 f9 f10);
 th2_1 BY it11_1* it12_1* it13_1* it14_1* it15_1* (f11 f12 f13 f14 f15)
     it16_1* it17_1* it18_1* it19_1* it20_1* (f16 f17 f18 f19 f20);
 th3_1 BY it21_1* it22_1* it23_1* it24_1* it25_1* (f21 f22 f23 f24 f25)
     it26_1* it27_1* it28_1* it29_1* it30_1* (f26 f27 f28 f29 f30);
 th1_2 BY it1_2* it2_2* it3_2* it4_2* it5_2* (f1 f2 f3 f4 f5)
     it6_2* it7_2* it8_2* it9_2* it10_2* (f6 f7 f8 f9 f10);
 th2_2 BY it11_2* it12_2* it13_2* it14_2* it15_2* (f11 f12 f13 f14 f15)
     it16_2* it17_2* it18_2* it19_2* it20_2* (f16 f17 f18 f19 f20);
 th3_2 BY it21_2* it22_2* it23_2* it24_2* it25_2* (f21 f22 f23 f24 f25)
     it26_2* it27_2* it28_2* it29_2* it30_2* (f26 f27 f28 f29 f30);
 xi1 BY th1_1-th3_1*.8 (lamb);
 xi2 BY th1_2-th3_2*.8 (lamb);
 [it1_1$1 it2_1$1 it3_1$1 it4_1$1 it5_1$1] (f31 f32 f33 f34 f35);
 [it6_1$1 it7_1$1 it8_1$1 it9_1$1 it10_1$1] (f36 f37 f38 f39 f40);
 [it11_1$1 it12_1$1 it13_1$1 it14_1$1 it15_1$1] (f41 f42 f43 f44 f45);
 [it16_1$1 it17_1$1 it18_1$1 it19_1$1 it20_1$1] (f46 f47 f48 f49 f50);
 [it21_1$1 it22_1$1 it23_1$1 it24_1$1 it25_1$1] (f51 f52 f53 f54 f55);
 [it26_1$1 it27_1$1 it28_1$1 it29_1$1 it30_1$1] (f56 f57 f58 f59 f60);
 [it1_2$1 it2_2$1 it3_2$1 it4_2$1 it5_2$1] (f31 f32 f33 f34 f35);
 [it6_2$1 it7_2$1 it8_2$1 it9_2$1 it10_2$1] (f36 f37 f38 f39 f40);
```

```
[it11_2$1 it12_2$1 it13_2$1 it14_2$1 it15_2$1] (f41 f42 f43 f44 f45);
[it16_2$1 it17_2$1 it18_2$1 it19_2$1 it20_2$1] (f46 f47 f48 f49 f50);
[it21_2$1 it22_2$1 it23_2$1 it24_2$1 it25_2$1] (f51 f52 f53 f54 f55);
[it26_2$1 it27_2$1 it28_2$1 it29_2$1 it30_2$1] (f56 f57 f58 f59 f60);
[th1_1-th3_1@0];
[th1_2-th3_2*0.2];
th1_1-th3_1@.5 th1_2-th3_2@.5;
[xi1@0];
[xi2*0.5];
xi1@1;
xi2*1;
xi1-xi2 WITH xi1-xi2;
OUTPUT: TECH1, TECH8;
SAVEDATA: FILE IS hSim11111.sav; SAVE = FSCORES;
PLOT: TYPE = PLOT3;
```

Appendix C

Example Mplus Script for Separate Calibration and $T = 2$

```
 TITLE: Longitudinal Hierarchical IRT Estimation
 DATA: FILE IS hSim11111_t1.dat;
 VARIABLE: NAMES ARE it1_1 it2_1 it3_1 it4_1 it5_1
it6_1 it7_1 it8_1 it9_1 it10_1
it11_1 it12_1 it13_1 it14_1 it15_1
it16_1 it17_1 it18_1 it19_1 it20_1
it21_1 it22_1 it23_1 it24_1 it25_1
it26_1 it27_1 it28_1 it29_1 it30_1;
 CATEGORICAL ARE it1_1 it2_1 it3_1 it4_1 it5_1
it6_1 it7_1 it8_1 it9_1 it10_1
it11_1 it12_1 it13_1 it14_1 it15_1
it16_1 it17_1 it18_1 it19_1 it20_1
it21_1 it22_1 it23_1 it24_1 it25_1
it26_1 it27_1 it28_1 it29_1 it30_1;
 ANALYSIS: TYPE = GENERAL;
 ESTIMATOR = MLR;
 LINK = LOGIT;
 INTEGRATION = MONTECARLO;
 MODEL: th1_1 BY it1_1* it2_1* it3_1* it4_1* it5_1* (f1 f2 f3 f4 f5)
     it6_1* it7_1* it8_1* it9_1* it10_1* (f6 f7 f8 f9 f10);
 th2_1 BY it11_1* it12_1* it13_1* it14_1* it15_1* (f11 f12 f13 f14 f15)
     it16_1* it17_1* it18_1* it19_1* it20_1* (f16 f17 f18 f19 f20);
 th3_1 BY it21_1* it22_1* it23_1* it24_1* it25_1* (f21 f22 f23 f24 f25)
     it26_1* it27_1* it28_1* it29_1* it30_1* (f26 f27 f28 f29 f30);
 xi1 BY th1_1-th3_1*.8 (lamb);
```

```
[it1_1$1 it2_1$1 it3_1$1 it4_1$1 it5_1$1] (f31 f32 f33 f34 f35);
[it6_1$1 it7_1$1 it8_1$1 it9_1$1 it10_1$1] (f36 f37 f38 f39 f40);
[it11_1$1 it12_1$1 it13_1$1 it14_1$1 it15_1$1] (f41 f42 f43 f44 f45);
[it16_1$1 it17_1$1 it18_1$1 it19_1$1 it20_1$1] (f46 f47 f48 f49 f50);
[it21_1$1 it22_1$1 it23_1$1 it24_1$1 it25_1$1] (f51 f52 f53 f54 f55);
[it26_1$1 it27_1$1 it28_1$1 it29_1$1 it30_1$1] (f56 f57 f58 f59 f60);
[th1_1-th3_1@0];
th1_1-th3_1@.5;
[xi1@0];
xi1@1;
  OUTPUT: TECH1, TECH8;
SAVEDATA: FILE IS hSim11111_t1.sav; SAVE = FSCORES;
PLOT: TYPE = PLOT3;
```