JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Winter 2018, Vol. 55, No. 4, pp. 614–634

Calculating Conditional Reliability for Dynamic Measurement Model Capacity Estimates

Daniel McNeish
Arizona State University
Denis Dumas
University of Denver

Dynamic measurement modeling (DMM) is a recent framework for measuring developing constructs whose manifestation occurs after an assessment is administered (e.g., learning capacity). Empirical studies have suggested that DMM may improve consequential validity of test scores because DMM learning capacity estimates were shown to be much less related to demographic factors like examinees' socioeconomic status compared to traditional single-administration item response theory (IRT)–based estimates. Though promotion of DMM has hinged on improved validity, no methods for computing reliability (a prerequisite for validity) have been advanced and DMM is sufficiently different from classical test theory (CTT) and IRT that known methods cannot be directly imported. This article advances one method for computing conditional reliability for DMM so that precision of the estimates can be assessed.

In educational and psychological measurement, a critical distinction can be drawn between the assessment of developed constructs (e.g., mathematics ability, extraversion, and self-efficacy) which are current at the time of testing, and developing constructs (e.g., learning capacity, giftedness, and potential) which lie in the future from the time of testing (Sternberg et al., 2002). Most currently available measurement modeling paradigms (e.g., item response theory [IRT] or confirmatory factor analysis) are capable of quantifying developed constructs, although resultant developed construct (i.e., ability) scores have often been used by substantive researchers to make inferences about developing constructs (Pfeiffer, 2012). In order to do so, an assumption of rank-order preservation must be made, in which the rank order of examinees on a developed construct (e.g., mathematics ability) is assumed to be equal to the rank order of examinees on the developed construct (e.g., mathematics learning capacity). As has been argued in the social sciences over the last century (Du Bois, 2013; Erwin & Worrell, 2012; Feuerstein, Rand, & Hoffman, 1979; Vygotsky, 1962), such a rank-order assumption leads to major consequential validity and fairness issues, especially when interpreting scores from impoverished or socially marginalized populations, because such examinees may not have had the opportunity to develop their abilities, although they retain their capacity for learning in the future.

In order to address this issue, dynamic assessment (DA) methods—which incorporate multiple testing occasions separated by standardized instruction by a clinician—were posited (e.g., Tzuriel, 2001). By charting an examinees' learning trajectory until asymptotic behavior of the learning curve was observed, DA researchers were able to quantify participants' developed ability across time, and an

asymptote for their learning capacity. However, because DA methods require that each participant receive individualized attention from a clinician, such methods are highly costly and therefore have not been widely adopted in the educational setting.

In response to this issue, dynamic measurement modeling (DMM)—which utilizes a nonlinear mixed effects modeling framework to estimate examinee-specific learning capacity asymptotes and is applicable to large-scale longitudinal educational data sets—was created (McNeish & Dumas, 2017). Using such a measurement-of-growth paradigm, DMM is capable of producing examinee-specific learning capacity asymptotes that have been noted to be far less affected by salient examinee demographic differences (e.g., socioeconomic status and race) than are single-time-point ability scores, therefore demonstrating the potential of DMM to improve the consequential validity of psychometric scores (Dumas & McNeish, 2017). However, at this point in the development of DMM, no reliability index for ascertaining the precision of estimates across the capacity score distribution exists. Such an index will be crucial if refinement of the DMM method is to continue.

In this article, we consider previous theoretical conceptualizations of reliability in IRT and how they may be applied to DMM. We then propose an adaptation of the conditional reliability method outlined in Nicewander (2018) using the quantities and information function present in DMM rather than those present in item response models. We conduct a simulation study to show that the proposed method is effective and that it is superior to the rudimentary approaches that have previously been implemented to capture the reliability of DMM estimates. A real-data example is provided to show the calculation of our proposed index and the types of questions it can help to answer. Limitations and future directions are then discussed.

Conceptualizing DMM as a Scoring Model

Though the DMM model specification mirrors a typical growth model, its application has much more in common with IRT models. In typical IRT applications, it is posited that there is some true latent variable that dictates examinee's responses to items that are designed to tap aspects of the ability in question (Thissen & Steinberg, 2009). For instance, if mathematics ability were the construct of interest, IRT theorizes that each examinee has an unobservable value for mathematics ability that increases or decreases the probability of a correct response to particular items. Mathematics items are administered to assess various aspects of mathematics and responses to these items are used to determine each examinee's level of mathematics ability—if many items are answered correctly or if an examinee is able to answer the most difficult items, it stands to reason that the examinee's ability is high. This basic IRT model can be thought of as a latent multivariate regression (e.g., Antal, 2007; de Boeck & Wilson, 2004)—the multiple outcomes are observed items responses and the probability of a correct response is predicted by a single latent predictor, termed ability or θ. Figure 1 expresses the basic IRT model for five items without guessing parameters (e.g., a Rasch or two-parameter logistic [2PL] model) as a path diagram: arrows point from the latent variable to the item response as it is the latent variable that is ultimately predicting the observed item responses. The overarching goal of IRT

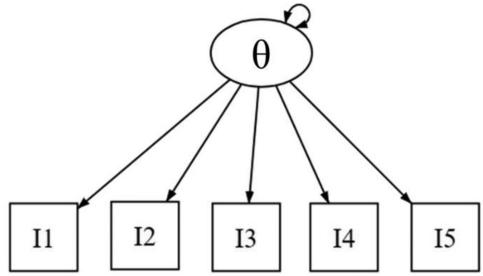

Figure 1. Path diagram of basic five-item IRT model without a guessing parameter. For a Rasch/1PL model, the loadings from θ to each item would be constrained to be equal; for a 2PL model, each loading would be freely estimated. This diagram was created using semdiag program (Mai, Zhang, & Yuan, 2016).

most often lies in obtaining estimates of the latent ability for examinees (Edelen & Reeve, 2007).

Similar to ability in IRT, DMM hypothesizes that each examinee has some unobserved capacity for the domain of interest. Operating on many of the same theoretical principles as IRT, the primary motivation is to obtain estimates of this latent capacity. Given the broader definition of capacity in DMM as contrasted with ability in IRT, the quantity of interest has a different theoretical meaning. Rather than administering items across a range of content areas relevant to the domain of interest to inform ability estimates, DMM makes use of ability scores across time to inform capacity estimates. Conceptually, the capacity in DMM is a higher-order or meta-construct for the IRT abilities: items are scored with IRT to estimate latent abilities, abilities are scored with DMM to estimate capacities. Just as latent abilities are estimated from item responses, capacities are estimated from ability scores. That is, DMM is essentially a scoring model just like IRT; however, instead of scoring items to obtain the latent ability construct, DMM scores abilities to obtain a latent capacity meta-construct.[1]

The similarity between DMM and IRT can be seen visually when comparing path diagrams. Figure 2 shows DMM as a structured latent curve model for data with five waves.[2] Basic nonlinear growth models with upper asymptotes useful for DMM (e.g., Gompertz, exponential, and Michaelis-Menten) tend to have at least three parameters to define growth curve: an initial value, an upper asymptote, and a rate parameter. Each of these growth parameters are featured as a latent variable. Importantly, note that the capacity latent variable in Figure 2 mirrors the role of the ability (θ) latent variable in Figure 1. Even though the model is specified as a growth model in Figure 2 and therefore includes more latent variables and parameters to define the model, at its core, there remains a focal variable of interest—the capacity—which is a latent variable that predicts ability scores and on which each person receives an examinee-specific score.

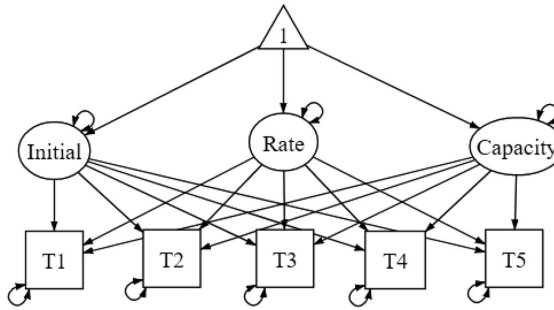

Figure 2. Path diagram of a basic DMM using five waves. When fitting the model in the structural equation modeling framework as a structured latent curve model, the loadings would be constrained to values based on the second partial derivatives with respect to the latent variable of interest. This diagram was created using semdiag program (Mai, Zhang, & Yuan, 2016).

Specifying a DMM

The empirical example upon which previous DMM research was based and that we will use here is the Early Childhood Longitudinal Study-Kindergarten (ECLS-K) mathematics test data available from the National Center for Educational Statistics. These data feature test scores on children at seven waves from kindergarten through Grade 8 (Kindergarten Fall, Kindergarten Spring, Grade 1 Fall, Grade 1 Spring, Grade 3 Spring, Grade 5 Spring, and Grade 8 Spring). Test scores are obtained via three-parameter logistic (3PL) IRT model and scores are vertically scaled so that all scores align on a common scale. Specific details of the variables and inclusion criteria for the subsample we used are described in the Empirical Example section of this article.

The best-fitting functional form for ECLS-K mathematics scale scores is the Michaelis-Menten function—which, for the ith examinee in the data set ($i = 1, \ldots, N$) at the tth time point ($t = 1, \ldots, T$)—can be written as

$$Math_{ti} = \beta_{0i} + \frac{(\beta_{Ai} - \beta_{0i})Time_t}{\beta_{Ri} + Time_t} + d_{ti}, \tag{1}$$

where

$$\begin{aligned} \beta_{0i} &= \gamma_0 + u_{0i}, \\ \beta_{Ai} &= \gamma_A + u_{Ai}, \\ \beta_{Ri} &= \gamma_R + u_{Ri}, \end{aligned} \tag{2}$$

and

$$\begin{aligned} \mathbf{d}_i &\sim MVN(\mathbf{0}, \mathbf{R}), \\ \mathbf{u}_i &\sim MVN(\mathbf{0}, \mathbf{G}). \end{aligned} \tag{3}$$

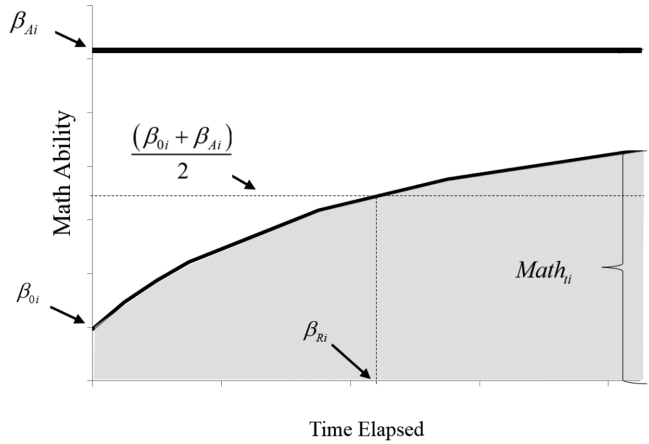

Figure 3. Graphical depiction of parameters in Equations 1 and 2 onto Michaelis-Menten plot for one hypothetical person. β_{Ai} is the capacity for this person, β_{0i} is the initial value for this person, $Math_{ti}$ is the person's ability up to time t, $\frac{(\beta_{0i}+\beta_{Ai})}{2}$ is the ability value that is halfway between the person's initial value and their capacity, and β_{Ri} is the time at which the person's ability reaches $\frac{(\beta_{0i}+\beta_{Ai})}{2}$.

In Equation 1, the three parameters of the Michaelis-Menten function are the initial value (β_{0i}) that captures ability when time is 0 (e.g., Kindergarten Fall in the ECLS-K data), the rate parameter (β_{Ri}), which captures the point in time when ability is halfway between the initial value, and the asymptote (β_{Ai}), which captures the maximum value of the outcome as time approaches infinity (i.e., the capacity); d_{ti} is a residual term that captures the difference between the model implied and observed value at each time point. **R** is a 7 × 7 covariance matrix for the residuals at each time point and **G** is a 3 × 3 random effect covariance matrix for the Michaelis-Menten presented in Equation 1. The normal assumptions of **d** and **u** can be relaxed if necessary.

In Equation 2, each of the three beta parameters in Equation 1 are composed of a population-averaged fixed effect (γ) and a examinee-specific random effect (u_i). These random effects allow each examinee in the data (note the *i* subscripts) to have their own unique growth curve where each u_i captures the difference between the examinee-specific growth parameter and the population-averaged parameter estimate (e.g., large positive values of u_i indicate that examinee *i* in the data is above average for that parameter). Though not explicitly shown in Equations 1 and 2, the random effects are allowed to covary with each other and the residual variances are uniquely estimated at each time point (i.e., no homoskedasticity assumption is made). Figure 3 shows a graphical depiction of how Equations 1 and 2 result in an interpretable nonlinear growth trajectory and capacity asymptote.

Though the Michelis-Menten function was the growth trajectory fit to this data, the general DMM framework is not restricted only to this particular trajectory. Alternative nonlinear growth models that feature asymptote parameters are equally

suitable if they accurately characterize the data of interest. This includes Gompertz, von Bertalanffy, logistic, Richards, or Morgan-Mercer-Flodin curves. Comparisons of these different curves within the DMM framework are provided in McNeish and Dumas (2017).

To tie the model back to the theoretical conceptualization of DMM, the components map as follows:

- Ability: $Math_{ti}$ is the mathematics scale score for the *i*th person in the data at time *t*.
- Capacity: u_{Ai} is the examinee-specific asymptote random effect for the *i*th examinee in the data. Random effects are defined to have a mean of 0, so a u_{Ai} value of 0 means examinee *i* has average capacity (relative to the sample), positive scores mean above average, and negative scores mean below average. The scale does not necessarily come from a standard normal and random effect variances are estimated and are not necessarily equal to 1.

Accounting for Measurement Error

As specified in Equation 1 through Equation 3, the DMM in this format assumes that scaled scores ($Math_{ti}$) are observed variables without measurement error. In actuality, these scores will typically be obtained with some type of scoring model such that the scale scores contain some uncertainty (a 3PL model is used in ECLS-K). This issue is addressed in McNeish and Dumas (2017) where they show how a DMM can be specified as a second-order model that simultaneously incorporates a measurement model and the nonlinear growth model. Specifically, the outcome variable can be scored with a latent variable model at each time point and, pending satisfactory adherence to longitudinal invariance of the scoring model, the nonlinear growth model can be applied to the latent variables at each time point. This type of joint model is sometimes referred to as a curve of factors model (McArdle, 1988).

Though this model has been devised, it does have noticeable weaknesses for use with empirical data. Notably, all DMMs are computationally intensive because (a) the likelihood does not have a close form and demanding approximations are required, (b) the model includes several random effects, and (c) sample sizes tend to be rather large. When vastly increasing the complexity of the model by combining measurement and growth models, the computational demand can become such a burden as to make the model unusable. For instance, in McNeish and Dumas (2017), this second-order DMM with about 2,100 people and seven waves was reported to take several weeks to converge when using only four quadrature points. Therefore, we proceed in this article by treating scale scores are observed. We concede this methodological choice results in some limitations but also wish to acknowledge that computational challenges make this limitation quite difficult to work around at the present time.

Reliability in the DMM Framework

Though DMM was advanced to address validity issues that can plague single-administration assessment, reliability is a prerequisite to establishing validity of any

kind because a method that cannot produce consistent and stable results inevitably has dubious validity (e.g., Lissitz, 2009). The focus of the remainder of this article is on discussing how to compute reliability within the DMM framework because DMM's status as a scoring model with unique longitudinally related quantities means that such computations diverge from previous formulas used in traditional contexts with binary items and single administration measurement.

Classical Methods

In previous applications of DMM to empirical data, reliability evidence has been established using classical test theory (CTT) methods. For example, using the ECLS-K mathematics test score data, Dumas and McNeish (2017) fit a DMM to scores from Kindergarten Spring through Grade 8 Spring. To establish reliability, a reduced model was fit that spanned only from Kindergarten Spring to Grade 5 Spring and the examinee-specific capacity estimates of the two models were correlated. The resulting correlation was $r = .93$, which was taken as evidence that the examinee-specific capacity estimates were stable and not merely statistical noise produced via empirical Bayes shrinkage (i.e., that the examinee-specific estimates were not random draws from a distribution centered around the capacity fixed effect).

Though effective as preliminary evidence, such a full-reduced model reliability method has notable shortcomings. First, similar to IRT, DMM is a latent variable method meaning that CTT methods of assessing reliability tend to oversimplify the true precision of the estimates. As with IRT, a single index of reliability for all examinees is likely inadequate to capture the precision of the capacity estimates in DMM because the precision changes across the distribution. Thus, a single index method like a correlation between estimates from two models neglects this varying degree of precision. Second, even if ignoring the issue of varying levels of precision across the capacity distribution, the number of testing occasions tends to be rather small in general for DMM analyses because the equivalent of "items" in the DMM scoring model are scale scores from an entire test rather than individual items. The method of correlating estimates from a full model and a reduced model requires that a model be fit to data consisting of fewer waves (i.e., a reduced model). This can lead to weaker estimates of the examinee-specific capacities, especially if the $T - 1$th time point is much earlier temporally than the Tth time point. This is prevalent in the ECLS-K data where the $T - 1$th time point is collected 3 years earlier than the Tth time point. In such situations, the precision of the examinee-specific estimates will be diminished as data are further removed from the time where growth begins to demonstrate asymptotic behavior. Furthermore, if data have relatively few waves to begin with, reducing the number of available waves can hinder the ability of the model to adequately capture the growth trajectory of the data. That is, the asymptotic portion of the growth function may no longer be apparent at the $T - 1$th time point, which would undoubtedly have an adverse effect on both the fixed effect estimates and the random effect predictions. Alternatively, reducing the waves may make the growth function appear linear, precluding a model that imposes an asymptote: the key parameter of interest in DMM. Perhaps most critical, in the case of data with only three or four waves, the reduced model approach may not be possible at all.

Therefore, full-reduced model correlations serve as a reasonable starting point for estimating reliability coefficients, but fail to incorporate important nuances of what a DMM analysis offers. As will be shown in a forthcoming simulation, properties of such a method are rather poor and leave much to be desired.

Conditional Reliability

In latent variable measurement models, precision of scores is more commonly assessed via an information function that spans the score distribution. Though information provides a more nuanced assessment of the precision of scores, it can be difficult to interpret absolutely in substantive applications because the scale of information has no upper bound, unlike classical test theory reliability, which is only defined over the [0,1] interval (Markon, 2013). Information also has the shortcoming in that it does not offer a global index that succinctly summarizes precision over the distribution of score estimates (Nicewander, 2018).

Nicewander (2018) notes that an index of conditional reliability rather than information would be more readily interpretable than information for tests scored with a latent variable measurement model. Although intuitively appealing, Nicewander notes that deriving such an index in the IRT context can be difficult, which is partially attributable to the logistic link function often used in IRT because this complicates the calculation of error variance featured in the classical reliability formula $\rho_{XX'} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$, where $\rho_{XX'}$ is the estimated reliability, σ_E^2 is the error variance of the scores, and σ_X^2 is the variance of the observed scores. This issue can be circumvented by treating observed scores as true scores as in the method proposed by Raju, Price, Oshima, and Nering (2007). To calculate the conditional reliability of IRT-based ability estimates θ, Raju et al. (2007) propose

$$\rho_{\hat{\theta}\hat{\theta}'} = \frac{\sigma_{\hat{\theta}}^2 - E\left(I(\theta)^{-1}\right)}{\sigma_{\hat{\theta}}^2}, \tag{4}$$

where $\hat{\theta}$ is an ability estimate for an examinee, $I(\theta)$ is the total test information function at θ, and $\sigma_{\hat{\theta}}^2$ is the variance of $\hat{\theta}$. This derivation is based on relations through CTT and results in the largest reliability estimates being in the tails of the score distribution as opposed to the typical behavior of information functions where maximal precision occurs near the center of the distribution (Nicewander, 2018). In contrast to the method proposed in Equation 4, Nicewander derives conditional reliability for tests consisting of binary responses for fixed-length assessments such that

$$\rho_{XX'}|\theta = \frac{I(X,\theta)}{I(X,\theta)+1}, \tag{5}$$

where X is a particular observed score and θ is an estimated latent ability. Nicewander (2018) shows that the conditional reliability formula in Equation 5 follows the expected pattern whereby precision is greatest in the center of the score distribution while also being on a scale that is bounded within the [0,1] interval. Nicewander further notes that these conditional reliabilities can also be aggregated to yield an

approximate marginal reliability as a global index of precision through Gaussian quadrature that sums the conditional reliabilities over equally spaced intervals of θ:

$$\bar{\rho}_{XX'} = \sum_{k=1}^{K} \varphi^*(\theta_k)(\rho_{XX'}|\theta_k), \tag{6}$$

where $\bar{\rho}_{XX'}$ is the approximate marginal reliability, K is the number of quadrature points, $\varphi^*(\theta_k)$ are normal density weights where $\varphi^*(\theta) = \frac{1}{\sqrt{2\pi}} e^{-.5\theta_k^2}$ (standardized so that $\sum \varphi^* = 1$), and $(\rho_{XX'}|\theta_k)$ is the conditional reliability over the kth interval. Though conceptually similar to CTT reliability, Nicewander (2018) notes that this integration will not be equal to CTT reliability because the former is a mean of ratios, whereas the latter is a ratio of means.

Though this recent work on conditional reliability creates an important foundation, DMM deviates in a meaningful way: capacity estimates in DMM are not based on binary items but rather on interval level scaled test scores. As a result, the issue concomitant with calculating the error variance differs from the typical case addressed by Nicewander (2018) and Raju et al. (2007). Additionally, the derivations in Nicewander (2018) are based on a binominal-based information function, meaning that this computation cannot be directly ported to the context of DMM capacity estimates.

To compute conditional reliability in DMM, the appropriate information function is required. Fortunately, when nonlinear mixed effects models are estimated with maximum likelihood, Fisher information for the examinee-specific random effect estimates is known, albeit indirectly because the primary interest of the model is the sampling variability of these parameters, which is equal to the inverse of Fisher information. The computation of this quantity is known from Booth and Hobert (1998) and easily computed in software, such as SAS PROC NLMIXED. Namely, the covariance matrix for examinee-specific random effects evaluated at a vector of fixed effect estimates $\hat{\gamma}$ and a vector of examinee-specific random effect estimates $\hat{\mathbf{u}}_i$ is

$$\mathbf{P}|\hat{\boldsymbol{\gamma}}, \hat{\mathbf{u}}_i = \begin{bmatrix} \hat{\mathbf{H}}^{-1} & \hat{\mathbf{H}}^{-1}\left(\frac{\partial \hat{\mathbf{u}}_i}{\partial \hat{\boldsymbol{\gamma}}}\right)^{\mathrm{T}} \\ \left(\frac{\partial \hat{\mathbf{u}}_i}{\partial \hat{\boldsymbol{\gamma}}}\right)\hat{\mathbf{H}}^{-1} & \boldsymbol{\Omega}^{-1} + \left(\frac{\partial \hat{\mathbf{u}}_i}{\partial \hat{\boldsymbol{\gamma}}}\right)\hat{\mathbf{H}}^{-1}\left(\frac{\partial \hat{\mathbf{u}}_i}{\partial \hat{\boldsymbol{\gamma}}}\right)^{\mathrm{T}} \end{bmatrix}, \tag{7}$$

where $\hat{\mathbf{H}}^{-1}$ is the inverse of the Hessian matrix of the fixed effects (the second partial derivatives of the log-likelihood function), $\frac{\partial \hat{\mathbf{u}}_i}{\partial \hat{\boldsymbol{\gamma}}}$ is the derivative of the $\hat{\mathbf{u}}_i$ vector evaluated at $(\hat{\boldsymbol{\gamma}},\hat{\mathbf{u}}_i)$, and $\boldsymbol{\Omega}^{-1}$ is the Hessian matrix of the examinee-specific random effects. An important relation to keep in mind is that observed Fisher information is equal to the negative inverse of the Hessian matrix, $I(\hat{\boldsymbol{\tau}}) = -\frac{\partial^2}{\partial \boldsymbol{\tau}_i \partial \boldsymbol{\tau}_j} l(\boldsymbol{\tau}) \mid_{\boldsymbol{\tau}=\hat{\boldsymbol{\tau}}}$ for $\boldsymbol{\tau}$ a vector of arbitrary parameters.

More importantly, the known computation of information is relevant because it facilitates the computation of conditional reliability using traditional formulas without additional derivations. That is, Nicewander (2018) defines conditional reliability as

$$\rho_{XX'} = \frac{\sigma^2(\lambda_\theta)}{\sigma^2(\lambda_\theta) + \sigma^2(X|\theta)} \tag{8}$$

for $\lambda_\theta = E(X|\theta)$, the expected true score at a fixed value of θ. Rather than derive $\sigma^2(\lambda_\theta)$ as was the approach in Nicewander (2018), Equation 8 can be rearranged such that

$$\rho_{XX'} = 1 - \frac{\sigma_e^2}{\sigma_e^2 + \sigma^2(X|\theta)}, \tag{9}$$

where σ_e^2 is the error variance, which is equal to the lower right element of $\mathbf{P}|\hat{\boldsymbol{\gamma}}, \hat{\mathbf{u}}_i$: $\boldsymbol{\Omega}^{-1} + (\frac{\partial \hat{\mathbf{u}}_i}{\partial \hat{\boldsymbol{\gamma}}})\hat{\mathbf{H}}^{-1}(\frac{\partial \hat{\mathbf{u}}_i}{\partial \hat{\boldsymbol{\gamma}}})^{\mathrm{T}}$. In the DMM context, the total value of the denominator is also straightforward to obtain—it is the variance of the asymptote random effects u_{Ai} defined in Equation 3. With values for each term defined, the conditional reliability for DMM capacity estimates can be computed as

$$\rho_{XX'|\,\beta_{Ai}} = 1 - \frac{Var(\hat{\beta}_{Ai})}{Var(u_{Ai})}, \tag{10}$$

where $\hat{\beta}_{Ai}$ is the examinee-specific capacity estimate, $Var(\hat{\beta}_{Ai})$ is the sampling variability for the ith examinee's capacity estimate, and $Var(u_{Ai})$ is the variance of the capacity random effects, which is an estimated parameter in the model. Note that this equation only depends on the quantities concerning the asymptote parameter, so it is unaffected by the growth trajectory and is not restricted to the Michaelis-Menten trajectory used in the example data. Similar to the derivation in Nicewander (2018), Equation 10 computes the precision of the examinee-specific capacities across the distribution of scores on an interpretable metric.

Simulation

To demonstrate that the proposed method can reasonably recover the reliability from DMM capacity scores and that it improves upon the full-reduced model reliability method that has previously been implemented in the DMM framework, we conducted a simulation study where the true reliability is known. To ensure that values are representative of real data, we based our data generation model on the parameter estimates from the reading assessment scores within the ECLS-K data set. The ECLS-K reading assessment focuses on familiarity with letters and phonemes at early waves (Rock & Pollack, 2002) and moves to more complex items based on passages that test interpretation, comprehension, and critical thinking (Najarian, Pollack, & Sorongon, 2009). Internal consistency reliability of the latent ability at each time point for the reading assessment is high and ranges from .87 (Grade 8 Spring) to .96 (Grade 1 Fall and Spring). Writing skills are not included on the reading assessment.

To obtain the population values, we fit a Michaelis-Menten model with random effects for the intercept, asymptote, and midpoint to reading scores at all seven waves using an unstructured random effect covariance structure and a homogeneous diagonal residual error structure to obtain these parameter estimates. This created the following population model:

$$\begin{aligned} y_{ti} &= \beta_{0i} + \frac{(\beta_{Ai} - \beta_{0i})Time_t}{\beta_{Ri} + Time_t} + d_{ti} \\ &\text{for } t = 0,\ 0.5,\ 1.0,\ 1.5,\ 3.5,\ 5.5,\ 8.5, \end{aligned} \tag{11a}$$

$$\begin{aligned}\beta_{0i} &= 29.58 + u_{0i},\\ \beta_{Ai} &= 288.40 + u_{Ai},\\ \beta_{Ri} &= 6.27 + u_{Ri},\end{aligned} \tag{11b}$$

$$\begin{aligned}\mathbf{d}_i &\sim MVN\left(\mathbf{0}, 167.07\mathbf{I}_7\right),\\ \boldsymbol{u}_i &\sim MVN\left(\mathbf{0}, \begin{bmatrix} 44.11 & & \\ -55.23 & 1504.39 & \\ -10.84 & 32.53 & 3.90 \end{bmatrix}\right).\end{aligned} \tag{11c}$$

The nature of reliability in the DMM framework and its reliance on the random effect Hessian matrix make it difficult to specify a true value strictly through setting parameter values in a simulation study. Therefore, generated data came from one super-population and we induced interreplication variability by sampling randomly with replacement from this super-population. This facilitated specification of a population value by which each sample could be compared in order to determine if the true reliability was recovered.

Two sample sizes are used in the simulation: 2,145 and 1,000. The 2,145 sample size was selected because that is the actual number of examinees in the ECLS-K data with reading assessment data at all seven waves. The 1,000 sample size condition was selected to explore properties of each reliability method when the sample size was smaller and the model more difficult to estimate. Each condition features 200 samples from the super-population. Sample size affects the precision of the random effect Hessian matrix, so the population reliability value is different in each sample size condition even though the super-population model parameters are the same. In the $N = 2{,}145$ condition, population reliability is .575. In the $N = 1{,}000$ condition, the population reliability is .439. To each sample, we fit the same Michaelis-Menten model to the data so that there were no misspecifications present. We then calculated marginal reliability for each replication using the previously outlined method. We calculated reliability using the full-reduced model method used in Dumas and McNeish (2017) as a basis of comparison. To obtain this estimate, we fit a model to a reduced model featuring only the first six waves and correlated the examinee-specific asymptotes of this reduced model to the examinee-specific asymptotes from the full seven-wave model.

We will inspect the distributional properties of the marginal reliability and full-reduced model method to determine how well they estimate the population value of reliability. Additionally, the relative bias of each reliability index will be calculated by $(\frac{\hat{\bar{\rho}}_{XX'} - \bar{\rho}_{XX'}}{\bar{\rho}_{XX'}})$. All analysis are conducted in SAS using PROC NLMIXED and models were estimated with Gaussian quadrature with three quadrature points and double dogleg optimization to keep computational times manageable given the complexity of the model and the sample sizes involved. The integration over the conditional reliability distribution required in the computation of marginal reliability will be done with 200 quadrature points.

Results

Table 1 shows the evaluation results for the each method and sample size condition across the 200 samples. For both sample size conditions, the marginal reliability

Table 1
Simulated Results from 200 Samples Based on ECLS-K Reading Score Model

Sample Size	Outcome	Marginal Reliability	Full-Reduced Model Reliability
1,000	Population Value		.439
	Mean Estimate	.462	.419
	Median Estimate	.419	.476
	95% Limits	[.280,.691]	[−.050, .833]
	Mean Relative Bias	5.17%	4.61%
	Median Relative Bias	−4.54%	8.37%
	Convergence	99.0%	92.0%
2,145	Population Value		.575
	Mean Estimate	.583	.430
	Median Estimate	.608	.502
	95% Limits	[.408, .670]	[−.049, .879]
	Mean Relative Bias	1.33%	−25.2%
	Median Relative Bias	5.58%	−12.7%
	Convergence	97.5%	91.5%

method that we propose in this article yielded estimates that are reasonably close to the population value. In the $N = 1{,}000$ condition, the mean and median marginal reliability estimates of .462 and .419, respectively, were quite close to the population value of .439. This resulted in relative small mean and median relative bias values of 5.17 % and –4.54%, respectively. In the $N = 2{,}145$ condition, the mean and median marginal reliability estimates of .583 and .608, respectively, were quite close to the population value of .575. This resulted in relative small mean and median relative bias values of 1.33% and 5.58%, respectively. All relative bias values for the marginal reliability method were within acceptable limits suggested by Hoogland and Boomsma (1998) and Flora and Curran (2004) for being considered to be within sampling error of the population value (these sources suggest that the magnitude of relative bias should be 10% or less). The 95% limits also showed that, while there was some variability (and that variability was higher with smaller samples), the values tended to be in the general vicinity of the population value.

On the other hand, the formerly utilized full-reduced model method performed rather poorly in the $N = 2{,}145$ condition. The mean and median estimates across all 200 samples were rather far below the population value of .575 at .430 and .502, respectively. This was reflected in mean and median relative bias values (–25.2% and –12.7%, respectively) that fell outside generally acceptable ranges. The most telling value of the inappropriateness of this method was the 95% limits, which essentially spanned the entire support of the index and even included negative values. As another disadvantage, because the full-reduced model method also requires estimation of two computationally intensive DMMs, the convergence rate is also lower than the convergence of the marginal reliability method, which only requires estimating the model once.

Though the mean, median, and relative bias values appeared to be acceptable for the full-reduced model method in the $N = 1{,}000$ condition, this is an artifact of the

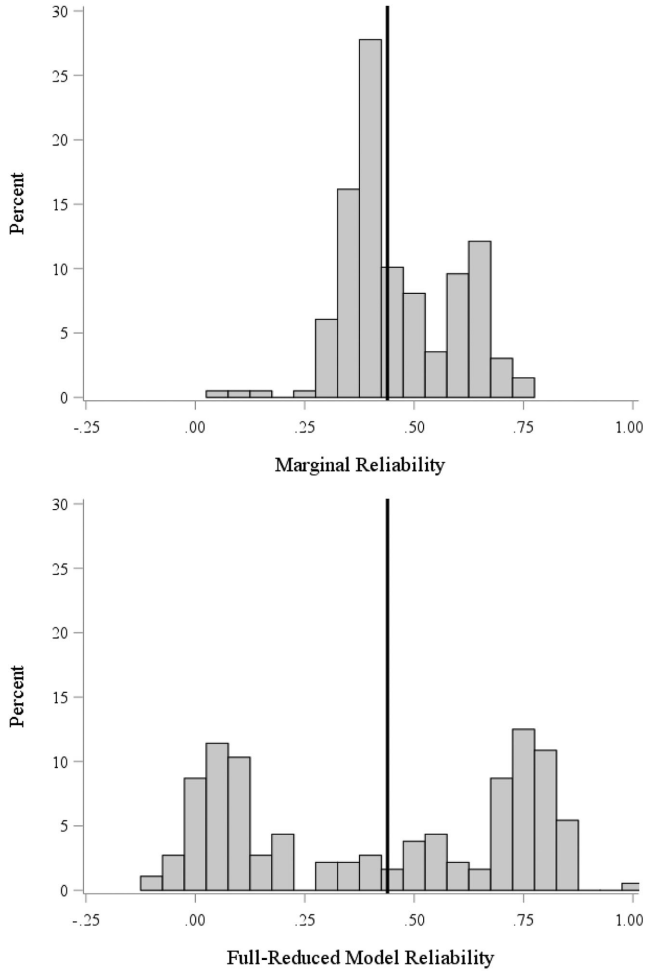

Figure 4. Histogram of marginal reliablity method (top panel) and full-reduced model reliability method (bottom panel) across 200 samples for the $N = 1{,}000$ sample size condition. The population value of .439 is represented by the black vertical line.

population value and a highly variable distribution. To demonstrate visually, Figure 4 shows the histogram of reliability estimates across the 200 samples for each method in the $N = 1{,}000$ condition. The marginal reliability values in the top panel have some variability, but the values are most clearly huddled near the population value (represented by the vertical black line). For the full-reduced model method, very few values are near the population value and the most frequent values are actually near .00 and .75, which happens to average out near the population value.

Figure 5 shows the same plot for the $N = 2{,}145$ condition. The results are similar in that the marginal reliability values mostly fall near the population value, whereas the full-reduced model estimates are widely spread over the entire range of possible

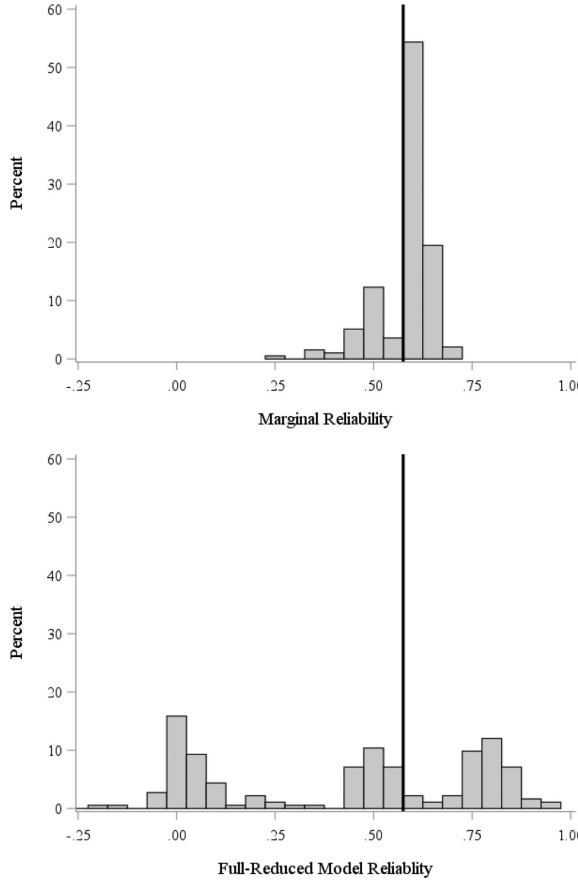

Figure 5. Histogram of marginal reliablity method (top panel) and full-reduced model reliability method (bottom panel) across 200 samples for the $N = 2{,}145$ sample size condition. The population value of .575 is represented by the black vertical line.

values. Overall, these results seem to provide some evidence that (a) the proposed marginal reliability method is capable of consistently recovering the population reliability value and (b) previously suggested full-reduced model methods do not appear to be appropriate for summarizing reliability in the DMM framework. Therefore, the marginal reliability method proposed in this article appears to be a useful addition to educational measurement using a DMM paradigm.

Empirical Example

Data Source

The sample consists of the 1998 ECLS-K public use data. We fit a DMM to all seven waves (Kindergarten Fall to Grade 8 Spring) for the 1,949 examinees who had complete data at all seven waves. Assessments at each wave are scored with a

3PL IRT model and then vertically scaled such that all waves use a single metric and scores at different waves can be directly compared. Content on the mathematics assessment focuses on number sense and basic operations at earlier waves (Rock & Pollack, 2002) and emphasizes algebra, patterns, functions, and data analysis by Grade 8 (Najarian et al., 2009). Internal consistency reliability of the latent ability at each time point for the mathematics assessment ranges from .91 (Kindergarten Fall) to .95 (Grade 3 and Grade 4 Spring; Najarian et al., 2009). The sample was 49% male, 72.2% white, 11.2% black/African American, 12.3% Latino/Hispanic, 4.3% Asian/Pacific Islander, and 22.3% qualified for free or reduced lunch. Table 2 shows the sample correlation matrix, standard deviations, minimum and maximum values, and the mean of the scale scores at all seven waves. Though missing data methods (e.g., full information estimation and imputation) for such data exist, we elected to only use data that were observed so that we did not need to rely on missing data assumptions. Additionally, the computationally intensive nature of the analysis would be greatly magnified employing such methods (e.g., multiple imputations would require estimating each model at least five times).

Analysis Details

We used PROC NLMIXED in SAS 9.3 using maximum likelihood estimation via Gaussian quadrature with 18 quadrature points and double dogleg optimization to estimate the model. For this example, we treat the ECLS-K scale scores as if they are observed scores without error. The variance of the seven scale scores appeared to grow over time, so we compared the Bayesian information criterion (BIC) values for a model that used a homogeneous residual error structure (where an equality constraint is placed on all seven residual variances) to a model that used a heterogeneous residual structure (where each residual variance is freely estimated). The BIC of the heterogeneous structure model (BIC = 80,445) was lower than the BIC of the homogeneous structure model (BIC = 106,199) and was used as the final model. Estimates for the final model are shown in Table 3. Figure 6 shows a plot of the empirical sample means against the model-implied Michaelis-Menten prediction line; the close correspondence of the empirical means to the line provides evidence that Michaelis-Menten is a reasonable choice for these data.

Conditional Reliability

To calculate the reliability of each examinee's score, we used the previously discussed formula such that $\rho_{XX'}|\hat{\beta}_{Ai} = 1 - \frac{Var(\hat{\beta}_{Ai})}{Var(u_{Ai})}$. The total variance of the capacity random effects in the model was 500.85 and the standard error of prediction for the examinee-specific capacities ranged from 10.2 to 14.7 (these values must be squared prior to entering the conditional reliability formula). Figure 7 shows a plot of the conditional reliability across the range of the capacity distribution along with the 95% confidence interval of the conditional reliability. Figure 7 largely mimics an IRT test information plot—reliability is highest near the center of the distribution, whereas reliability dips at both tails of the range. The general shape of the plot also resembles plots reported in Nicewander (2018) based on mathematics scores from

Table 2
Descriptive Statistics for ECLS-K Mathematics Scale Scores

	Fall Kindergarten	Spring Kindergarten	Fall Grade 1	Spring Grade 1	Spring Grade 3	Spring Grade 5	Spring Grade 8
Fall Kindergarten	9.34	–	–	–	–	–	–
Spring Kindergarten	.81	12.04	–	–	–	–	–
Fall Grade 1	.68	.75	13.88	–	–	–	–
Spring Grade 1	.70	.76	.73	17.64	–	–	–
Spring Grade 3	.65	.70	.67	.76	22.94	–	–
Spring Grade 5	.61	.65	.63	.72	.85	21.98	–
Spring Grade 8	.54	.58	.58	.65	.78	.84	19.53
Min	12.04	16.00	.20	21.92	46.93	50.87	71.80
Max	93.23	113.00	99.40	132.49	164.22	170.66	172.20
Mean	28.48	39.70	46.58	65.92	104.86	129.56	146.08

Note: Diagonal values are the standard deviations at each time point; off-diagonal entries are correlations.

Table 3
DMM Estimates for Full Model Using ECLS-K Mathematics Scores

Parameter Name	Parameter	Estimate
	Fixed Effects	
Initial Value	γ_0	26.36
Capacity	γ_A	255.53
Rate	γ_R	7.39
	Random Effect Variances and Correlations	
Var (Initial Value)	g_{00}	71.89
Var (Capacity)	g_{AA}	500.85
Var (Rate)	g_{RR}	2.05
Corr (Initial Value, Capacity)	$r(u_{0i}, u_{Ai})$	−.06
Corr (Initial Value, Rate)	$r(u_{0i}, u_{Ri})$	−.31
Corr (Capacity, Rate)	$r(u_{Ai}, u_{Ri})$	−.36
	Model Fit	
–2Log-Likelihood	80,323	
Number of Parameters	16	
BIC	80,445	

Note: Correlations between the random effects are reported rather than the covariances for ease of interpretation. *Cov* (Initial Value, Capacity) = –11.25, *Cov* (Initial Value, Rate) = –3.73, *Cov* (Capacity, Rate) = –11.41.

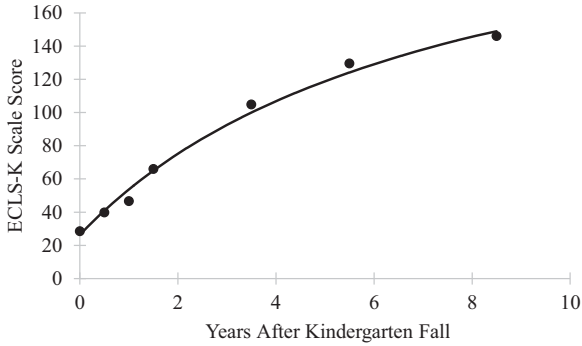

Figure 6. ECLS-K mathematics empirical sample means plotted against the Michaelis-Menten model-implied marginal curve. The close correspondence of the Michealis-Menten curve to the empirically observed sample means implies the suitability of the Michealis-Menten function to model the growth of ECLS-K scores.

Pacific Metrics assessments. The 95% confidence interval of the reliability estimates is also largest at the tails and narrowest near the central tendency. Using 200 quadrature points, the marginal reliability of the capacity estimates is $\bar{\rho}_{XX'} = .678$. Taking the sample dependent mean of conditional reliabilities across the 1,949 examinees yields a slightly higher value of $\overline{(\rho_{XX'}|\beta_{Ai})} = .700$.

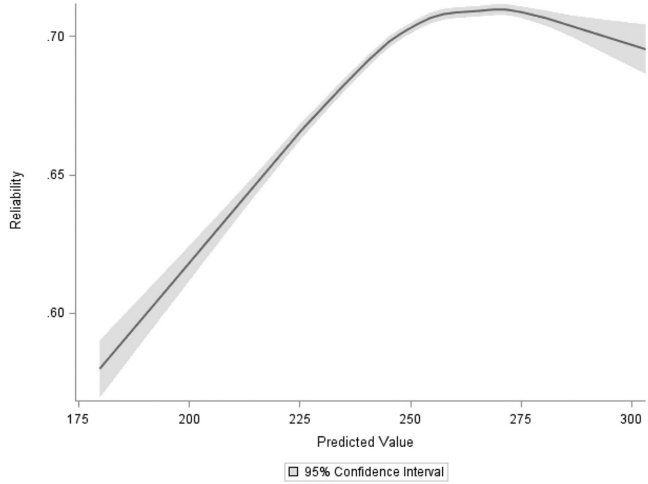

Figure 7. Plot of conditional reliability for ECLS-K capacity estimates with 95% confidence interval shown in gray. Reliability tends to be the highest near the mean capacity ($\alpha_U = 255.53$). Reliability is notably less than at the mean for lower scores; reliability is only slightly less than at the mean for higher scores though the confidence interval is also wider.

Discussion

From a substantive perspective, although reliability values near .70 seem at the lower end of the spectrum of acceptable reliability, note that the precision of the examinee-specific random effects is dependent on how far the last wave is from the asymptotic behavior of growth trajectory. In these data, the Grade 8 data are still rather far from the asymptote as evidenced by the rate parameter being estimated to be only one grade earlier ($\hat{\boldsymbol{\gamma}}_R = 7.39$). For example, if the model were fit to only the first six waves (Kindergarten spring to Grade 5 Spring), the marginal reliability based on 200 quadrature points is estimated to be only .424. Thus, it is reasonable to assume that the conditional reliability of the examinee-specific capacity estimates would increase if additional data were collected after Grade 8 Spring or if a different domain whose asymptotic behavior occurred earlier was being assessed. For example, in the DA tradition, asymptotic behavior can be observed within a few days of instruction when the construct of interest is fluid reasoning ability (Tzuriel & Caspi 2017).

From a conceptual standpoint, reliability may be expected to be lower for DMM capacity estimates than for traditional IRT ability estimates or scale scores. Measuring capacity is an ambitious task to undertake because the underlying meta-construct has not yet fully developed when the assessments are given (i.e., it lies in the future from the time of testing). Therefore, less precise estimates would be expected in such a context, regardless of how strong the model is, because the meta-construct itself remains in flux. At this point in this line of inquiry, the exact factors that most influence the conditional reliability of DMM capacity asymptotes—whether they be

aspects of the construct of interest, the test items themselves, the number of testing waves, or aspects of the DMM modeling framework—are not definitely known. Conceptually, reliability would be expected to be lower with fewer waves, smaller samples, and when the last wave is far from asymptotic behavior but the precise effect or the magnitude of these aspects is not quantifiable from the current study. Future research to uncover such influential factors is of paramount importance to the usefulness of DMM in an applied setting. However, such ongoing and future work would be impossible without an established conditional reliability index for DMM capacity estimates. The goal of this article was to advance a modern method for computing reliability of DMM capacity estimates and our simulation was aimed at demonstrating the efficacy of the proposed method. In this way, the current work is an integral part of developing the DMM methodology, as well as building the field's psychological understanding of learning capacity.

Notes

[1]Throughout this section, we focus on the latent trait perspective to measurement such that the interest lies in obtaining quantitative scores for people because, similar to IRT, DMM falls into this category of models. However, there is a growing literature within the diagnostic framework for measurement where the goal of the model is to assign people to latent qualitative classes based on their performance (Rupp & Templin, 2008). Models for learning in the diagnostic framework have received much attention in the recent literature and may be important to consider alongside or as an alternative to DMM (e.g., Chen, Culpepper, Wang, & Douglas, 2018; Kaya & Leite, 2016; Li, Cohen, Bottge, & Templin, 2015; Wang, Zhang, Douglas, & Culpepper, 2018).

[2]Though a DMM can be fit as a structured latent curve model in the structural equation modeling framework, in order to estimate the model, some linearization is required, which can affect the examinee-specific estimates (e.g., Harring & Blozis, 2016). Because these examinee-specific parameters are the primary focus of the analysis, it is best to fit the model as a nonlinear mixed effects model because no linearization is required in the estimation of models in this framework. However, the structural equation modeling framework allows for a more interpretable path diagram, which is of interest here to demonstrate the relation between DMM and IRT.

References

Antal, T. (2007). *On the latent regression model of item response theory*. (Research report). Princeton, NJ: Educational Testing Service.

Booth, J. G., & Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, *93*, 262–272.

Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2018). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement*, *42*, 5–23.

de Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Du Bois, W. E. B. (2013). *W. E. B. DuBois on sociology and the Black community*. Chicago, IL: University of Chicago Press. (Original essay published 1920).

Dumas, D. & McNeish, D. (2017). Dynamic measurement modeling: Using nonlinear growth models to estimate student learning capacity. *Educational Researcher*, *46*, 284–292.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*, 5–18.

Erwin, J. O., & Worrell, F. C. (2012). Assessment practices and the underrepresentation of minority students in gifted and talented education. *Journal of Psychoeducational Assessment*, *30*, 74–87.

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques*. Baltimore, MD: University Park Press.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Harring, J. R., & Blozis, S. A. (2016). A note on recurring misconceptions when fitting nonlinear mixed models. *Multivariate Behavioral Research*, *51*, 805–817.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research*, *26*, 329–367.

Kaya, Y., & Leite, W. L. (2016). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling an evaluation of model performance. *Educational and Psychological Measurement*, *77*, 369–388.

Li, F., Cohen, A., Bottge, B., & Templin, J. (2015). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement*, *76*, 181–204.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing.

Mai, Y., Zhang, Z., & Yuan, K. (2016). An online interface for drawing path diagrams for structural equation modeling. Retrieved June 21, 2018, from http://semdiag.psychstat.org/

Markon, K. E. (2013). Information utility: Quantifying the total psychometric information provided by a measure. *Psychological Methods*, *18*, 15–35.

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

McNeish, D., & Dumas, D. (2017). Non-linear growth models as psychometric models: A second-order growth curve model for measuring potential. *Multivariate Behavioral Research*, *52*, 61–85.

Najarian, M., Pollack, J. M., & Sorongon, A. G. (2009). *Early childhood longitudinal study, Kindergarten class of 1998–99 (ECLS-K), psychometric report for the eighth grade (NCES 2009–002)*. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2009/2009002.pdf

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods*, *23*(2), 351–362.

Pfeiffer, S. I. (2012). Current perspectives on the identification and assessment of gifted students. *Journal of Psychoeducational Assessment*, *30*, 3–9.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, *31*, 169–180.

Rock, D. A., & Pollack, J. M. (2002). *Early childhood longitudinal study-kindergarten class of 1998–1999 (ECLS-K), psychometric report for kindergarten through first grade (NCES 2002–05)*. Washington, DC: Department of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2002/200205.pdf

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219–262.

Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C., … Bundy, D. A. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence*, *30*, 141–162.

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. E Millsap & A. Maydue-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 148–177). Thousand Oaks, CA: Sage.

Tzuriel, D. (2001). *Dynamic assessment of young children*. New York, NY: Kluwer Academic.

Tzuriel, D., & Caspi, R. (2017). Intervention for peer mediation and mother–child interaction: The effects on children's mediated learning strategies and cognitive modifiability. *Contemporary Educational Psychology*, *49*, 302–323.

Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Wang, S., Zhang, S., Douglas, J., & Culpepper, S. A. (2018). Using response times to assess learning progress: A joint model for responses and response times. *Measurement: Interdisciplinary Research and Perspectives*, *16*, 45–58.

Authors

DANIEL MCNEISH is Assistant Professor in the Quantitative Area of the Psychology Department at Arizona State University, PO Box 871104, Tempe, AZ, USA 85287; dmcneish@asu.edu. His primary research interests include latent variable models, models for dependent data, and small sample analysis.

DENIS DUMAS is Assistant Professor of Research Methods and Statistics at the University of Denver, 1999 E. Evans Ave., Denver, CO 80210; denis.dumas@du.edu. His primary research interests include the identification and estimation of measurement parameters that indicate student learning and growth in an academic domain.