Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: http://www.tandfonline.com/loi/hmbr20

A Hierarchical Rater Model for Longitudinal Data

Jodi M. Casabianca, Brian W. Junker, Ricardo Nieto & Mark A. Bond

To cite this article: Jodi M. Casabianca, Brian W. Junker, Ricardo Nieto & Mark A. Bond (2017) A Hierarchical Rater Model for Longitudinal Data, Multivariate Behavioral Research, 52:5, 576-592, DOI: 10.1080/00273171.2017.1342202

To link to this article: https://doi.org/10.1080/00273171.2017.1342202

MULTIVARIATE BEHAVIORAL RESEARCH
2017, VOL. 52, NO. 5, 576–592
https://doi.org/10.1080/00273171.2017.1342202

Routledge
Taylor & Francis Group

A Hierarchical Rater Model for Longitudinal Data

Jodi M. Casabianca[a], Brian W. Junker[b], Ricardo Nieto[a], and Mark A. Bond[a]

[a]The University of Texas at Austin; [b]Carnegie Mellon University

ABSTRACT

Research studies in psychology and education often seek to detect changes or growth in an outcome over a duration of time. This research provides a solution to those interested in estimating latent traits from psychological measures that rely on human raters. Rater effects potentially degrade the quality of scores in constructed response and performance assessments. We develop an extension of the hierarchical rater model (HRM), which yields estimates of latent traits that have been corrected for individual rater bias and variability, for ratings that come from longitudinal designs. The parameterization, called the longitudinal HRM (L-HRM), includes an autoregressive time series process to permit serial dependence between latent traits at adjacent timepoints, as well as a parameter for overall growth. We evaluate and demonstrate the feasibility and performance of the L-HRM using simulation studies. Parameter recovery results reveal predictable amounts and patterns of bias and error for most parameters across conditions. An application to ratings from a study of character strength demonstrates the model. We discuss limitations and future research directions to improve the L-HRM.

KEYWORDS

Autoregressive; hierarchical rater model; item response theory; latent trait estimation; longitudinal; ratings; time series; trends

With the promotion of higher order skills in the assessment of individuals, there has been a surge in the use of ratings of rich response formats. For example, state K-12 accountability tests rely on constructed response items to assess writing and reasoning ability (e.g. written composition on the State of Texas Assessment of Academic Readiness test), observations of teachers in the classroom are used to inform professional development activity (Hill & Grossman, 2013) as well as assess teachers' effectiveness for research outcomes or human resources decisions (e.g. the Measures of Effective Teaching [MET] Project; see Kane & Staiger, 2012), and ratings of student behavior given by their parents and teachers are used in research settings as outcomes (e.g. see Institute of Education Sciences [IES], 2010; Stanford Center on Adolescence, 2014). These scenarios in the education landscape demonstrate the need for advanced psychometric models that account for rater effects, or the error introduced to scores by the human rating process.

The need to account for rater effects naturally extends to the longitudinal context where we seek to evaluate growth due to some intervention or naturally occurring changes over time. For example, in an IES-funded intervention called Social and Character Development (SACD; IES, 2010), teachers and parents completed surveys to rate children at five timepoints to evaluate their changes in behavior and determine if an intervention to improve character development was effective. This is an instance where a longitudinal model is useful to estimate traits or detect growth over time, but it is also necessary to account for rater effects introduced by the rating process. The structure of the data in these instances is complex. There are many examinees being rated by multiple raters over several timepoints. It is likely that not all examinees are rated by all raters, but it is also likely that subsets of the rater pool will observe and evaluate the same work. The goals in these instances are usually very simple: we want to evaluate individual examinees at multiple timepoints with estimates of latent traits that have been refined for rater biases and we may also want to evaluate the effectiveness of an intervention by estimating an overall trend in growth. In another study of character traits called the Character Development in Adolescent Project (CDAP; Stanford Center on Adolescence, 2014), we may wish to examine changes when there is no intervention. We will use this study as an example to motivate and demonstrate the model described here.

Rater drift, or changes in how raters use the scoring rubric over time, has earned a lot of attention in the literature in recent years (see, for example, Harik et al., 2009; Leckie & Baird, 2011; Myford & Wolfe, 2009; Wilson & Hoskens, 2001). Unfortunately, in most instances including the aforementioned SACD intervention, we lack the ability to simultaneously evaluate

CONTACT Jodi M. Casabianca jcasabianca@ets.org Educational Testing Service, 660 Rosedale Road, MS T-03, Princeton, NJ 08541, USA.
Supplemental material for this article can be accessed at www.tandfonline.com/hmbr

changes in raters and changes in traits. This is true because the examinees' behaviors/responses occur on the same timeline as the raters' evaluations. Decoupling trends for examinees and raters is possible with the proper rating design where observations of examinees are rated on a different timeline than the actual behavior. This is possible, for example, if observations are collected and then rated out of sequence (see e.g. Casabianca, Lockwood, & McCaffrey, 2015).

Considering the complexity of the data structure and the goal of modeling changes in traits based on a collection of longitudinal ratings, a carefully parameterized item response theory (IRT) model is a practical solution. Such a model should: include parameters for rater effects, appropriately treat multiple ratings of the same work, and provide estimates of latent traits over multiple timepoints as well as an estimate of overall growth. There are several IRT models for ratings that could be adapted to suit these needs; however, our research focuses on an extension of the hierarchical rater model (HRM; Casabianca, Junker, & Patz, 2016; Patz, 1996; Patz, Junker, Johnson, & Mariano, 2002) to fit this need. The HRM is a multilevel IRT model that includes parameters for rater bias and rater variability. Its hierarchical structure nests the observed ratings given by multiple raters for a particular item response within a per-item latent variable. Ignoring this nesting structure and including all observed ratings as responses to individual items leads to trait estimates with falsely low standard errors (as shown in Patz, 1996). The final piece of the solution is the longitudinal component. The addition of a growth trend and autoregressive component to the basic parameterization of the HRM is the focus of this paper. Thus, in this article, we present the longitudinal hierarchical rater model (L-HRM) to model changes in latent traits over time as measured by ratings collected in longitudinal designs. In the following sections, we discuss rater models and longitudinal models for ratings data and in general. We then (re)introduce the basic HRM as described in Patz et al. (2002) and then introduce the L-HRM. To evaluate the feasibility of this extension, we present results from a real data illustration and two simulation studies, and then discuss further extensions and limitations.

Psychometric models for ratings

Rater effects have the potential to introduce substantial error into scores, thereby degrading their precision and impacting the subsequent uses of those scores whether it be for the evaluation of an intervention in a research context or a decision about an applicant or employee in a practical context. There are several types of rater effects (see Wolfe [2014] for an in-depth discussion of rater effects in education), but most commonly the literature discusses *rater bias*, a rater's tendency to score higher (leniency) or lower (severity) on average (Hoyt, 2000; Hoyt & Kerns, 1999; Kingsbury, 1922). Considering rater effects and then mitigating rater error are important steps in the measurement of individuals.

There is no shortage of cross-sectional IRT models for ratings data that incorporate parameters to account for, and study, raters and the rating process. However, only a handful of psychometric rater models account for the dependencies brought about by multiple ratings of the same work. These include the HRM (Casabianca et al., 2016; Patz, 1996; Patz et al., 2002); Verhelst and Verstralen's (2001) IRT model for multiple raters, Wilson and Hoskens' (2001) rater bundle model, and more recently, DeCarlo, Kim, and Johnson's (2011) variant of the HRM, called HRM-signal detection theory (HRM-SDT). Other models, such as the IRT Facets model (Linacre, 1989) ignore the nesting structure of ratings of the same work and consider ratings from multiple raters rating the same work (or behavior) as additional information that should contribute to the trait estimation thereby leading to a downward bias in the standard errors of the latent traits (see Mariano, 2002). Unfortunately exhaustive comparisons of all of these rater models have not been conducted. We further discuss and develop the HRM and not any of the other models described here that account for dependencies for a couple of reasons. First, the HRM has already been expanded (see Mariano & Junker, 2007) and we are further building on that expanded framework. Furthermore, despite its ability to provide the analysis of a variety of rater effects and richer diagnostic information about raters, the HRM-SDT involves an even larger parameter space which may prove troublesome with many added parameters such as in this context. Our primary focus here is to estimate serially dependent traits while accounting for rater effects.

Generalizability ("G") study models (Brennan, 2001) are traditionally used to evaluate the variance decomposition in a ratings scenario in order to determine which components contribute substantial error to ratings. The correspondence between G study models and the HRM is noteworthy (Patz et al., 2002). Much like the HRM, G study models incorporate unobservable parameters representing the quality of an examinee's response to an item. Values of these parameters reflect the scores on a particular item as if it were measured with perfect reliability and without bias. Observed ratings are nested within these unobservable item-level parameters, and their expected value is the quality of the response. In the G study model framework, this ideal response is a true score for the observed ratings. Under the HRM framework it is termed the *ideal rating*. The primary

difference is that G study models are parameterized with the Normal-theory true score model framework, and HRMs are parameterized with IRT and discrete signal detection distributions.

Some models for longitudinal data

Rater models

Instances of longitudinal parameterizations of rater models appearing in the literature are generally few and far between, with the exception of a few noteworthy articles. Recently, Casabianca et al. (2015) applied an augmented parameterization of a G study model using B-splines to model time trends in teaching quality and rater behavior. Their model included random and fixed effects for scoring trends in order to determine how individual raters change during the scoring period, how teachers change over time, as well as how rater variation changes during the scoring period. It was only due to the data collection design, and the decoupling of the rater scoring timeline and the timeline for teachers' observation, that they were able to estimate separate trends for the teachers and for the raters.

Hung and Wang (2012) applied a three-level generalized multilevel facets model to longitudinal ratings data. Level 1 modeled the item responses at specific timepoints using a facets model which treats rater severity as a random effect; level 2 modeled latent growth with an autoregressive residual structure to account for variation over timepoints in latent traits; and level 3 modeled variation in growth between examinees. This model assumed that item parameters are constant over administrations, but estimated time-specific rater severity. Since this model relies on the facets model parameterization to model rater effects, it does not account for the dependencies from multiple ratings of the same examinee's work. Thus, Hung and Wang's model is theoretically very different from the L-HRM that we will propose in this article, at least in terms of the approach to treating these dependencies.

In much the same vein as Hung and Wang (2012), Guo (2014) introduced a 3-level cross-classified random-effects model to correct for rater effects in data designs that do not include complete nesting of ratings but instead multiple ratings at different timepoints are cross-classified by raters and examinees. Guo (2014) and Hung and Wang (2012) report parameter recovery for rater severity over time but do not address the conflation of examinee and rater trends. The aforementioned longitudinal models each proposed to analyze ratings with different foci, goals, and assumptions.

Longitudinal modeling under the SEM framework

IRT modelers may not intuitively consider IRT for analyzing longitudinal data as IRT must make assumptions about the stability of the latent trait (McArdle, Petway, & Hishinuma, 2015). However, researchers have indeed explored the use of IRT to model repeated measures data but with the necessary assumptions that the construct definition remains constant, and that changes we observe in individuals are due to real changes or growth, and not due to changes in the construct definition over time (Andersen, 1985; Embretson, 1991, 1997; Millsap, 2010; Roberts & Ma, 2006). For example, items on a depression scale may perform differently for pre-adolescents versus adolescents as the construct of "depression" may have different definitions at different stages of life. The structural equation modeling (SEM) literature explicitly addresses this by discussing different levels of longitudinal measurement invariance. Measurement invariance in general is often discussed in terms of invariance for subgroups, but we may apply the same principles to establish invariance over time (for general information on measurement invariance, see: Horn & McArdle, 1992; Little, 2013; McArdle et al., 2015; Meredith, 1993). Recently, Liu et al. (2017) proved that the invariance criteria for attributing changes over time to actual changes in latent common factors over time is different for continuous and ordered-categorical latent responses. That is, in order to make this claim, *unique factor invariance* must be established, which requires invariance of factor loadings, thresholds, and unique factor variances over time. This is the level of invariance that would need to be established to make such claims with the L-HRM.

Longitudinal models under the SEM framework take on different forms and respond to different questions—for example, there are latent growth curve models, latent change score models, and longitudinal panel models, just to name a few (see Little, 2013 for a comprehensive discussion on longitudinal SEMs). In general, however, for our purposes we classify the structure that defines the changes for the latent trait into one of four models (following McArdle et al., 2015). A simple linear change score model permits growth/changes at the individual level (by modeling a mean slope, variance of growth, [thereby allowing individuals to vary], and a correlation between the initial level scores and change score). An autoregressive change score model uses an examinee's trait at the previous timepoint to predict the trait at the current timepoint. A dual change score (DCS) model permits additive and proportional influences to influence the changes in the latent trait (by combining the linear change score and autoregressive change score model components). Lastly, the triple change score model builds on

the DCS model by allowing the shape of the changes to be estimated.

The L-HRM as it will be described may be considered a variant of the DCS model in which the latent accumulation of changes is part additive (by way of a growth component) and part proportional (by way of an autoregressive component). However, we show in the following sections that the parameterization of the L-HRM presented here is further restricted in that it estimates overall growth (and not individual growth/trends). It is flexible in that it permits the explicit modeling of different types of growth (in this paper, we study linear growth). Furthermore, we make an assumption that the items hold the same properties and relationship with the trait over time and much like in Hung and Wang (2012), we estimate one common set of item parameters for all timepoints.

The hierarchical rater model

The HRM is a three-level Bayesian hierarchical model for ratings. The first level of the hierarchy models the distribution of ratings given the quality of response, the second level models the distribution of an examinee's response given their latent trait, and the third level models the distribution of the latent trait θ. The hierarchical representation is given by

$$
\begin{aligned}
X_{ijr}|\xi_{ij}, \tau_r^2, \ \phi_r &\sim \text{polytomous signal detection model}, \\
&\quad r = 1, \ldots, R, \text{ for each } i, \ j. \\
\xi_{ij}|\theta_i, \ \alpha_j, \ \beta_j, \ \gamma_{jk} &\sim \text{polytomous IRT model}, \\
&\quad j = 1, \ldots, J, \text{ for each } i \\
\theta_i &\sim Normal(\mu_\theta, \sigma_\theta^2), \\
&\quad i = 1, \ldots, N \text{ where } \sigma_\theta^2 = {}^1\!/_\omega \\
\omega &\sim Gamma\,(a_\omega, b_\omega) \\
\alpha_j &\sim Gamma\,(a_\alpha, b_\alpha) \\
\beta_j &\sim Normal(\mu_\beta, \sigma_\beta^2) \\
\gamma_{jk} &\sim Normal(\mu_\gamma, \sigma_\gamma^2) \\
1/\tau_r^2 &\sim Gamma\left(a_{1/\tau^2}, b_{1/\tau^2}\right) \\
\phi_r &\sim Normal(\mu_\phi, \sigma_\phi^2)
\end{aligned} \tag{1}
$$

Here, θ_i, the latent trait for examinee i ($i = 1, \ldots, N$) is normally distributed with mean μ_θ and σ_θ^2, ξ_{ij} is the ideal rating for examinee i on item j ($j = 1, \ldots, J$), and X_{ijr} is the observed rating given by rater r for examinee i's response to item j. In the above hierarchical Bayesian representation, we have assumed prior distributions for all unknowns in the model.

The HRM hierarchy connects a two-stage measurement process; in the first stage, an examinee's work on J items may be hypothetically judged to have some true rating or quality of response, we call this the "ideal rating"; in the second stage, a series of R raters evaluate the work, giving ratings based on their observations in accordance with a scoring rubric, with some added noise resulting from their rating tendencies. An IRT model relates the ideal ratings to the latent trait variable in the first stage, and a "signal-detection-like" model measures the ideal rating of an indicator based on multiple raters' observed ratings in the second stage.

Suppose we have N examinees to be rated, and for each examinee there is a latent trait, ideal ratings per-item, and observed ratings per-item from each rater. In the second level, the ideal ratings ξ_{ij} represent the quality of examinee i's response to item j, and are latent variables modeled using a polytomous IRT model, such as the K-category generalized partial credit model (GPCM; Muraki, 1992). From the GPCM component of the HRM we estimate α_j, the item discrimination, β_j, the item difficulty, and γ_{jk} the kth item step parameter for item j. Note that other polytomous IRT models can be used in this level, and that K, the number of response categories per indicator, need not be constant across items. With ideal rating ξ_{ij} and K possible scores ($k = 1, \ldots, K$), the GPCM is given by:

$$
\begin{aligned}
&P(\xi_{ij} = \xi\,|\theta_i, \alpha_j, \beta_j, \gamma_{j\xi}) \\
&\quad = \frac{\exp\left\{\sum_{k=1}^{\xi} \alpha_j(\theta_i - \beta_j) - \gamma_{jk}\right\}}{\sum_{h=0}^{K-1} \exp\left\{\sum_{k=1}^{h} \alpha_j(\theta_i - \beta_j) - \gamma_{jk}\right\}}.
\end{aligned} \tag{2}
$$

The ideal rating is the rating that examinee i would receive on item j, by a rater exhibiting *no* rater bias and perfect rating consistency. In the HRM, the deviations between actually observed ratings X_{ijr} and these ideal ratings ξ_{ij} are modeled using a discrete signal detection model. Specifically, we use the signal detection model to define the relationship between the observed and ideal rating *probabilities*, which is easily conceptualized in matrix form. Consider a matrix of response probabilities with ideal ratings in the rows and observed ratings in the columns, and each entry in the matrix is defined by $p_{\xi kr} =$ (Rater r rates k|ideal rating ξ). The signal detection model is specified such that probabilities in each row of this matrix are proportional to a Normal density with mean $\xi + \phi_r$, and standard deviation τ_r:

$$
\begin{aligned}
p_{\xi kr} &= P(X_{ijr} = k|\xi_{ij} = \xi) \\
&\propto \exp\left\{-\frac{1}{2\tau_r^2}[k - (\xi + \phi_r)]^2\right\}.
\end{aligned} \tag{3}
$$

The bias parameter ϕ_r indicates a rater's predictable deviation from the ideal rating, and represents a consistent bias in the rater's ratings; values near 0 indicate no deviation, negative values indicate negative bias (sometimes referred to as "severity"), and positive values

indicate positive bias (or "leniency"). The spread parameter τ_r indicates a rater's variability; values near 0 indicate high consistency or reliability in rating and high values indicate poorer consistency in rating.

We estimate the HRM with Markov chain Monte Carlo (MCMC) estimation (see Patz et al. [2002] for details on estimation including complete conditional distributions). To specify the model we take note of the conditional independence assumptions. The observed rating X_{ijr} is conditionally independent of the observed rating for other items and given by other raters $X_{ij'r'}$ given θ_i. Further, the observed ratings assigned by rater r are conditionally independent across items and examinees given $\boldsymbol{\lambda}_\mathbf{r} = [\phi_r, \tau_r]$. The quantities ω_θ, $\boldsymbol{\lambda}_\mathbf{r} = [\phi_r, \tau_r]$ and $\boldsymbol{\eta}_\mathbf{j} = [\alpha_j, \beta_j, \gamma_{jk}]$ are treated as random and therefore they need prior distributions. Equation (1) provides the prior distributions that we use to estimate these parameters. Note that typically we use Gamma prior distributions to restrict ω, α, and the rater precision $1/\tau^2$ to be positive, but other priors may be used (e.g. log-normal).

Model identification

There are multiple methods to identify the GPCM in the HRM when using a Bayesian approach to estimation. Identification of the GPCM can be implemented using "hard" (fixing certain parameter values) or "soft" constraints (using very strong priors), as long as they adequately solve both location and scale indeterminacies. The location indeterminacy may be solved by constraining either μ_θ (such that $\mu_\theta = 0$) or constraining the βs. There is an additional location indeterminacy in the γ_{jk}s. Thus, one approach is to constrain the model by centering the mean of θs at 0 ($\mu_\theta = 0$) and applying a sum-to-zero constraint on γ_{jk} such that $\gamma_{j(K-1)} = -\sum_{k=1}^{K-2} \gamma_{jk}$. One could instead apply sum-to-zero constraints on both the βs and the γ_{jk}s and estimate the mean of the θs. In addition, in the GPCM, a scale indeterminacy must be addressed by either setting the latent trait precision to 1 ($1/\sigma_\theta^2 = 1.0$) or applying a sum-to-zero constraint on the log of the discrimination parameters, such that $\log \alpha_J = -\sum_{j=1}^{J-1} \log \alpha_j$ or $\alpha_J = 1/\prod_{j=1}^{J-1} \alpha_j$. Alternatively, soft constraints may be imposed by specifying very strong priors on some parameters. For example, instead of setting the precision to 1.0, one could place a very strong hyperprior on the precision hyperparameter so that the estimate would be essentially 1 (or very close). We apply different methods of setting priors in our example and simulations.

The longitudinal hierarchical rater model (L-HRM)

Autoregressive time series models provide a mechanism to incorporate information from previously observed data in the estimation of the current state (Box, Jenkins, & Reinsel, 2013; Hamilton, 1994; Hershberger, Molenaar, & Corneal, 1996). In the case of trait estimation over multiple timepoints, we posit that previous trait estimates may be informative. The L-HRM presented here models latent traits using an autoregressive time series model and an overall growth trend that can take on any shape, including linear, logistic, and cubic. As we were motivated by evaluating change due to interventions, we present a parameter for *overall* growth. Individual growth terms may also be included at the expense of parsimony. Further, note that this parameterization does account for changes in rater and item effects, but assumes these parameters are constant over the study period. The model as we will show below could be adapted to instead estimate these parameters at each timepoint.

Suppose ratings from a longitudinal design to be analyzed with the HRM are from M timepoints ($m = 1, \ldots, M$). In the L-HRM, ideal and observed ratings at time m are nested within each θ_{im}, where θ_{im} is the trait for examinee i at timepoint m. Now, instead of assuming a Normal distribution for the traits, we specify them using a longitudinal model. All other levels of the HRM are the same with the exception of the ideal ratings, now indexed as ξ_{ijm}. The corresponding hierarchical representation is given in (4)—the quantities in this representation will be described in detail as the focus of the remainder of this section.

$$
\begin{aligned}
X_{ijrm}|\xi_{ijm}, \tau_r^2,\ \phi_r &\sim \text{polytomous signal detection model,} \\
&\quad r = 1, \ldots, R, \text{ for each } i,\ j,\ m \\
\xi_{ijm}|\theta_{im},\ \alpha_j,\ \beta_j,\ \gamma_{jk} &\sim \text{polytomous IRT model ,} \\
&\quad j = 1, \ldots, J, \text{ for each } i,\ m \\
\theta_{im}|\rho, \delta &\sim \text{longitudinal model,} \\
&\quad m = 1, \ldots, M, \text{ for each } i \\
\rho &\sim \mathit{Uniform}\left(a_\rho, b_\rho\right) \\
\delta &\sim \mathit{Normal}\left(\mu_\delta, \sigma_\delta^2\right) \\
\alpha_j &\sim \mathit{Gamma}\left(a_\alpha, b_\alpha\right) \\
\beta_j &\sim \mathit{Normal}\left(\mu_\beta, \sigma_\beta^2\right) \\
\gamma_{jk} &\sim \mathit{Normal}\left(\mu_\gamma, \sigma_\gamma^2\right) \\
1/\tau_r^2 &\sim \mathit{Gamma}\left(a_{1/\tau^2}, b_{1/\tau^2}\right) \\
\phi_r &\sim \mathit{Normal}\left(\mu_\phi, \sigma_\phi^2\right)
\end{aligned} \tag{4}
$$

The longitudinal model we will use for θ_{im} lets the trait for examinee i at timepoint m be a function of two quantities:

$$\theta_{im} = G_m + Z_{im}. \tag{5}$$

Here, G_m is the trend in θ_{im} at time m and Z_{im} is the time series component of the model. We model Z_{im} using a stationary autoregressive model of order 1, or

AR(1). That is, $Z_{im} = \rho Z_{i(m-1)} + (\sqrt{1-\rho^2})\varepsilon_{im}$, where $Z_{i(m-1)}$ the lagged value of Z_{im}, is weighted by ρ, the autocorrelation parameter which quantifies the correlation between adjacent timepoints. The random noise for examinee i at timepoint m denoted ε_{im} is distributed as $Normal(0, \sigma_\theta^2)$ and is weighted by a function of the autocorrelation $\sqrt{1-\rho^2}$. Weighting ε_{im} assures stationary variance of the noise (and therefore the resultant traits) across M timepoints. We place a $Uniform(-1,1)$ prior distribution on ρ. The trend G_m can be any deterministic function representing overall (positive or negative) growth. For example, a linear trend could be modeled as $G_m = \delta \times (m-1)/(M-1)$. The overall trend or growth over the M timepoints is δ, which we express in terms of standard deviation (SD) units using the estimated value of σ_θ such that growth in SD units $= \delta/\sigma_\theta$. To reflect little prior knowledge about growth we assign δ an uninformative $Normal(0, \sigma_\delta = 10)$ prior.

The model in (5) can be restated using two steps implemented at each timepoint m:

Step 1: AR(1) process (detrended), Z_{im}:
  a. When $m = 1$ and there is no lagged value, we define $Z_{im} \sim Normal(0, \sigma_\theta^2)$.
  b. When $m > 1$, we place a Normal prior with different hyperparameters, namely, $Z_{im} \sim Normal(\rho \times Z_{i(m-1)}, \sigma_\theta^2(1-\rho^2))$.

Step 2: The latent trait at time m is computed as an additive function of the estimated parameters: $\theta_{im} = G_m + Z_{im}$ where G_m is a function of the estimated growth δ.

The autoregressive time series model of order 1 uses only information from the previous timepoint and assumes that the correlations between the traits decrease as the distances between the timepoints increase. Certainly, other orders could be modeled to incorporate information from more than one previous timepoint. In addition, more general parameterizations of time series models, such as autoregressive moving average models (ARMA; for a general reference, see Box at al., 2013), could be useful in estimating traits over time. The autoregressive structure and order are consistent with the modeling approaches found in several articles in the psychometric literature related to longitudinal latent variable models (Andrade & Tavares, 2005; Azevedo, Fox, & Andrade, 2015; Cagnone, Moustaki & Vasdekis, 2009; Dunson, 2003; Eid & Hoffmann, 1998; Fu, Tao, Shi, Zhang, & Lin, 2011). While different parameterizations follow from the structure we present here, we introduce evaluations of this specific AR(1) parameterization with an eye toward educational and psychological applications, which often have fewer timepoints. For example, an AR(2) may not be sensible when $M = 4$ or 5. To motivate the benefit of modeling rater effects in longitudinal rating-based studies with the L-HRM, we present its application using data collected for a large-scale study before discussing results from a simulation study testing parameter recovery.

Motivating example: Changes in character strength in CDAP

Sample and measures

The CDAP (Stanford Center on Adolescence, 2014) is a longitudinal study that examined growth in adolescents' character strength. Beginning in 2014, the two-year study followed a cohort of 1,777 eighth-grade students over four school semesters (spring 2014, fall 2014, spring 2015, and fall 2015), and measured different attributes of character strength using a variety of behavioral measures. Data were collected through student self-report measures and interviews, as well as teacher ratings of students' character strengths.

The CDAP focused on a core set of character strengths, relating to the six items under study here: (i) actively open-minded thinking (AOT), (ii) gratitude, (iii) grit, (iv) prosocial purpose, (v) interpersonal self-control, and (vi) work-related self-control. For student-based responses, these six character strengths were measured via short (five to six item) adapted measures (e.g. Grit-S Scale [Duckworth and Quinn, 2009]; the Gratitude Questionnaire-6 [Froh et al., 2011]), as well as newly developed measures for prosocial purpose, AOT and the two self-control constructs. At each timepoint, teacher ratings, analyzed in this example, measured a common construct of *character strength* by asking teachers to rate each child on each of the six strengths as observed within the past month. For each of the six items, a description of behaviors related to the associated character strength was included that used the items comprising the associated student self-report survey. Each item was measured on the same five-point Likert scale (1 = Never True, 2 = Rarely True, 3 = Sometimes True, 4 = Often True, 5 = Always True), where higher scores denoted higher levels of character strength. To assess the dimensionality of the rating instrument, we performed an exploratory factor analysis using a varimax rotation on the original data by treating each teacher-student pair as a unique respondent. Results yielded a one-factor solution based on the Kaiser–Guttman rule, and visual inspection of a scree plot. The scale was found to be highly reliable ($\alpha = .95$).

Eight middle schools across Pennsylvania, California, Texas, and Idaho were selected to participate in the study. Because the CDAP followed eighth graders into ninth

grade, the sample of schools also included seven high schools. For each semester of data collection, students' English, Math, Science, and Social Studies teachers were invited to complete the rating instrument. Thus, at each measurement occasion, students received ratings from up to four different teachers. In total, ratings were collected for 1,777 students from 161 teachers over four timepoints. Our application of the L-HRM focuses on a subset of 81,552 ratings collected for 1,065 students evaluated at each timepoint by 145 teachers. Across the four timepoints, students received ratings from an average of 6.16 raters ($SD = 2.25$), and teachers rated an average of 64.70 students ($SD = 37.14$).

Analysis

A plot of the average ratings for each item and overall over the four timepoints in the CDAP shows very little to no change over time (see Figure 1). All score means are in the approximate score range of 3–4, with the average score hovering around 3.3–3.5, indicating a prevalence for ratings of character strength near "Sometimes True" or "Often True." There was a slight dip from timepoint 1 to timepoint 2 (0.10) and then an increase from timepoint 2 to timepoint 3 (0.14), but this may have occurred because there was a new set of teachers evaluating the children starting at timepoint 3, or we could attribute these minor changes to noise. Further, this observed trend does not take into account item or rater effects. The CDAP data did not allow a rater drift analysis because the observation of the child and the evaluation of the behavior occurred simultaneously and also because the set of teachers participating in the study changed from grade to grade. Thus, the goal of our analysis is to estimate latent traits representing character strength and determine if there was an overall change.

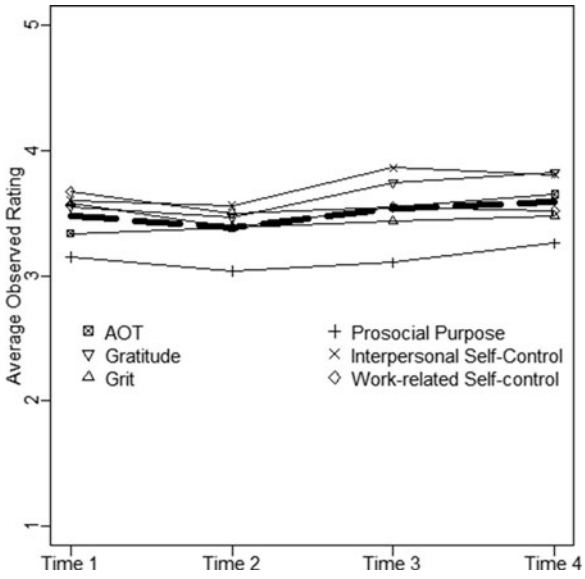

Figure 1. Average observed trends on the CDAP Character Strength scale. The thick dashed line is the average trend across items. The thin lines show the observed average item scores over time. (Note: AOT = Actively Open-minded Thinking; CDAP = Character Development in Adolescence Project).

For demonstration purposes, we fit the L-HRM parameterization as outlined in (4) with a linear trend. We used a partial credit model (PCM) instead of the GPCM because in preliminary runs we found all six of the discrimination estimates to be close to one. Thus, to reduce the estimation problem, we opted to constrain all discriminations to 1.0. Note that the character strength scale in the CDAP is indeed an ordinal/Likert scale, with responses ranging from 1 to 5. There are several polytomous IRT models available for use, but not all would be appropriate for this type of scale. For example, the nominal categories item response model would be inappropriate because it does not incorporate the ordinal nature of the scale. Perhaps the polytomous IRT model that most people would consider for this problem is the rating scale model (Andrich, 1978), which is simply a constrained version of the GPCM that holds thresholds equal across items.[1] To avoid making such strong assumptions about the thresholds, however, we continue with the partial credit family of models.

Model specification

We experimented with priors and settled on the following for the CDAP data.[2] For both models, we assigned a $Normal(0,1)$ prior for the θ_{i1} (the first timepoint), a $Normal(0,1)$ prior for the β_j, a $Normal(0,0.5)$ prior for the γ_{jk}, a $Normal(0,1)$ for the ϕ_r, and a $Gamma(1,1)$ prior for the $1/\tau_r^2$. We placed a $Normal(0,10)$ prior on δ and a $Uniform(-1,1)$ prior on ρ. Sum-to-zero constraints were used to identify the IRT model. We estimated the model using MCMC estimation in JAGS (Plummer, 2003) with 3 chains, 35,000 iterations each, with the first 5,000 iterations discarded as burn-in, and a thinning interval of 10. This yielded a final sample of 9,000 (with the chains combined). We evaluated the convergence of chains according to the Gelman–Rubin convergence diagnostic ($\hat{R}$; Gelman & Rubin, 1996); all $\hat{R}$ values were below 1.1. Values higher than 1.1 indicate that the chains did not converge.

To evaluate the model's absolute fit, we performed posterior predictive model checks. Posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996; Rubin, 1984; Sinharay, 2005; Sinharay, Johnson, & Stern,

[1] It should be noted that the GPCM is appropriate for use with rating scale data despite traditional viewpoints on how the GPCM should be used (Hambleton, van der Linden, & Wells, 2010, p. 31).

[2] Equation (4) presents the Normal hyperparameters in terms of variances, but when discussing the actual values for the hyperparameters here, we provide SDs to assist the readers' understanding.

2006) is a Bayesian method used to compare observed data to data replicated or generated from the model under evaluation via the predictive distribution. If the statistical summaries generated from the observed and replicated data indicate systematic discrepancies, this indicates misfit as the model is inadequately capturing that aspect or structure of the observed data. To perform a general check of overall model fit we compared the observed total score distribution (where a "score" is the sum of the item scores, and the item scores are the average of all ratings for that item) to the posterior predictive distribution (based on replicated ratings from the data-generating process associated with the model using the same sample size as in the analyses). The posterior predictive distribution is similar to the observed distribution along the entire total test score scale in terms of central tendency, spread, and shape, and the observed distribution is within the 95% posterior predictive intervals in areas of large mass (see Supplementary Material, Figure A1). This result allows us to conclude that the L-HRM with AR(1) is an adequate fit to the CDAP data. Note that there have been no studies evaluating the use of PPMC for the HRM. Thus, the usefulness of PPMC for the HRM/L-HRM is not fully clear.

Results

Statistical summaries of the posterior distributions show very small Monte Carlo errors compared to the posterior SDs, signifying the MCMC simulation yielded good approximations of the theoretical posterior expectations. Item 4 ("prosocial purpose") was the most difficult, while all other items were easier in comparison, especially Items 2 ("gratitude") and 5 ("interpersonal self-control"). The step parameter estimates were inconsistently spaced across items, supporting our use of the partial credit model (PCM) (over the RSM). (See Table A1 in the Supplementary Material for summaries of the posterior distributions for these parameters.)

The posterior median for the autocorrelation ρ was 0.89 (95% Credible Interval: [0.88, 0.91]), indicating a very strong correlation between adjacent timepoints. There was a loss in character strength over the four timepoints (which equates to two years of school)—the posterior median for δ was −0.49 [−0.63, −0.35]). This trend was not visible on the plot of observed rating means over time (Figure 1), however, after adjustments for item characteristics and rater effects, the trend of latent traits may not directly correspond to the trend (or lack of trend) in the raw ratings. We see this in Figure 2 which provides latent trait profiles for a random sample of 100 children from the CDAP. Most children had trait estimates in the −2 to 2 range, and experienced decline over the four timepoints, though there was variability in trajectories.

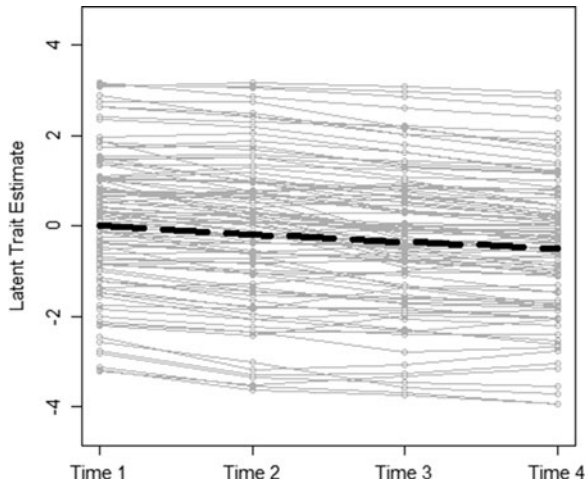

Figure 2. Estimated latent trait trends in the CDAP Character Strength scale. The gray lines represent profiles of latent trait estimates from a random sampling of the 1,065 CDAP adolescent participants. The black dashed line represents the overall fitted linear trend depicting an average decline of 0.49 over the four timepoints. (Note: CDAP = Character Development in Adolescence Project).

The bolded, black, dashed line shows the average trend, which indicates an average loss in character strength of half a SD. Importantly, this decline in character strength is not attributable to any intervention or policy change.

Diagnostic information about the raters (teachers) can be gleaned from Figure 3, which plots rater SD (y-axis) and rater bias (x-axis). Most raters were indeed in the "safe" zone, between −0.50 and 0.50. However, many other raters were in the positive space above 0.5, all the

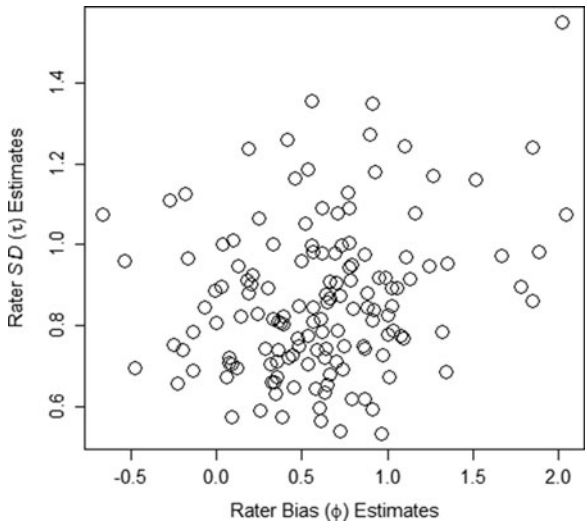

Figure 3. Rater parameter estimates from the L-HRM fit to ratings from the character strength scale administered in the CDAP. The scatterplot depicts the distribution of rater bias (x-axis) and rater SD (y-axis) for N = 145 raters. (Note: SD = standard deviation; τ = rater standard deviation; ϕ = rater bias; CDAP = Character Development in Adolescence Project; L-HRM = longitudinal hierarchical rater model).

way up to 1.6. This indicates that many raters were lenient in the scoring, at least relative to ideal ratings. There were a few raters who were very lenient and also had high rater SDs, which equates to large inconsistency in their rating tendencies. Interpretation of these rater diagnostics could have been performed at early points in the study to determine which teachers are too severe or lenient and/or inconsistent. They could then be remediated before the next timepoint to reduce rater errors.

Connecting the prevalence of rater leniency and the estimated traits, we see in the fitted latent trait trends a slight decline in character strength which is converse to what was in the observed plot. This could be a function of the model correcting for the positive rater bias and instead capturing the true trend in character strength.

Simulation studies to evaluate the L-HRM

Study 1: Evaluate L-HRM recovery with varying levels of N, R and M

We first tested the L-HRM parameterization with a simulation study that examined feasibility and parameter recovery with relatively small samples ($N = 250, 500$) while varying the number of raters ($R = 3, 6$), and the number of timepoints ($M = 3, 7$). We decided to test this model with conditions comparable to real data sets, including our CDAP empirical illustration as well as other longitudinal studies using ratings (e.g. the Measures of Effective Teaching project; Kane & Staiger, 2012). The two M conditions selected were comparable to what has been evaluated in terms of simulations and empirical data in other studies (e.g. Andrade & Tavares, 2005; Azevedo et al., 2015; Cagnone et al., 2009; Dunson, 2003). In all cases we used $J = 5$ rated test items; exploration showed that a greater number of rated test items reduced standard errors for estimating θ_{im}, however, we kept J small because typically, constructed responses compose shorter tests due to the burden and expense involved in rating process. We modeled a linear trend in all conditions. The final set of results includes eight conditions, completely crossing R, N, and M.

Generation of ratings

To maximize our ability to generalize, we chose to consider traits, rater parameters, and item parameters as random effects and randomly sampled true parameter values from specified distributions. True rater bias (ϕ_r) and SD (τ_r) for R raters were drawn from *Normal*(0,1) and *Lognormal*(0.5, 0.25) distributions, respectively. True GPCM item parameters for five items with five response categories ($K = 5$) were generated by drawing from *lognormal*(1, 0.25), $N(0, 0.125)$, and $N(0, 1)$ for the discrimination, difficulty, and step parameters, respectively.

The SD of θ was fixed at $\sigma_\theta = 1.0$ and therefore the precision was also $\omega = 1.0$. In this study, we tested a linear trend, $\delta \times (m-1)/(M-1)$. To reflect realistic growth over an academic year, we set the total growth at ¼ of a population SD in θ. Thus, "raw" growth was $\delta = 0.25$ and growth in SD units was also 0.25 since $\sigma_\theta = 1.0$. The autocorrelation parameter was fixed to $\rho = 0.80$, a relatively strong correlation between traits at times m and $m-1$. Note that when we examine the results, we will assess growth as a fraction of the estimated SD of θ. We will also discuss the recovery of the variability of the θs in terms of σ_θ instead of ω.

We generated ratings by drawing the baseline trait level θ_{i0} from a *Normal*(0,1) distribution and using the time series model in (5) to compute the additional θ_{im} values. Using the true values for traits, raters, and items, we generated ideal ratings first using the GPCM as shown in (2). The generated ideal ratings were then used in the signal detection model as shown in (3) to generate observed ratings. We generated 100 replications for each condition ($L = 100$).

Model estimation and outcome measures

We fit the L-HRM in JAGS (Plummer, 2003) via the R2jags package (Su & Yajima, 2012) with 8,000 iterations, 3 chains, and a burn-in of 2,000. To reduce autocorrelation, we thinned the chains such that we kept every 6th iteration. We relied on computational resources from the Texas Advanced Computing Center's (TACC) supercomputer, Stampede,[3] where replications were parallelized across computing nodes to reduce the time required to complete the simulation study (see Xu, Huang, Zhang, El-Khamra, & Walling, [2016] for more information on parallelization). We evaluated each replication for the convergence of chains according to the Gelman–Rubin convergence diagnostic ($\hat{R}$; Gelman & Rubin, 1996); a replication was excluded if there was one or more estimated parameter with $\hat{R} > 1.1$.[4] Values higher than 1.1 indicate that the chains did not converge. Upon the rejection of a replication for this reason, we then estimated an additional replication so that our final set of results included 100 converged replications for each condition.

In terms of prior distributions, we assigned non-informative priors to all estimated parameters. For $1/\tau_r^2$, α_j, and ω we used a *Gamma*(1,1) prior density,

[3] For more information on Stampede, visit: https://www.tacc.utexas.edu/stampede/

[4] In many practical situations, the analyst may request more iterations on the same set of chains to arrive at convergence, however, the remote manner in which we ran our simulation study did not support this process. For this reason, we simply started a new set of chains with a new set of randomly generated initial values.

reflecting a noninformative prior for the nonnegative quantities. We assigned *Normal*(0,10) priors for δ and ϕ_r, and *Normal*(0,1) priors for β_j and γ_{jk}. We assigned a *Uniform*(−1,1) prior for ρ. Sum-to-zero constraints were applied to the IRT item parameters to identify the model.

The purpose of this simulation study was to evaluate the recovery of parameters, particularly the parameters associated with the longitudinal model component and latent traits, under various data structures and conditions. To evaluate results, we report the mean absolute bias for all parameters and conditions—bias was computed as the difference between the true parameter value and the posterior median, $(\hat{p} - p)$. The mean of all absolute bias values across the parameter type within a condition is the mean absolute bias. We also report another measure of accuracy, the root mean square error (RMSE), computed as the square root of the average squared difference between the true parameter value and the posterior median, $\text{RMSE}(\hat{p}) = \sqrt{\sum_{l=1}^{L=100} (\hat{p}_l - p)^2 / L}$ (averaged over L replications). Where appropriate, we also provide plots of bias as a function of true parameter value to highlight the variation in bias. In order to put all parameter estimates on the same scale as their generating parameter values before evaluating recovery, we applied a linear transformation to the estimates of the traits and item parameters using the estimated SD of the latent traits.

Recovery of longitudinal model parameters and latent traits

Table 1 provides the mean absolute bias and RMSEs across replications for all estimated parameters. The growth parameter, δ, had mean absolute bias that ranged from 0.044 to 0.080 and RMSEs that ranged from 0.052 to 0.095. Mean absolute bias and RMSEs were consistently larger in the $N = 250$ conditions (versus the $N = 500$ conditions) and the $M = 7$ conditions (versus the $M = 3$ conditions). There does not appear to be an effect related to the number of raters.

Estimation of the autocorrelation yielded very little bias and error, with mean absolute bias ranging from 0.01 to 0.026 and RMSEs ranging from 0.013 to 0.031. There was better recovery with the $N = 500$ conditions and the $M = 7$ conditions, with the combination of those two yielding the least error. The number of raters did not have an effect on the autocorrelation.

Mean absolute bias for σ_θ ranged from 0.089 to 0.106 and RMSEs ranged from 0.110 to 0.134. The bias was mostly positive. There were no consistent patterns differentiating the number of timepoints or number of raters. Both mean absolute bias and RMSEs were smaller in the $N = 500$ conditions when $M = 3$. For the latent traits, the mean absolute bias ranged from 0.336 to 0.362 and RMSEs ranged between 0.424 and 0.459. Recovery was consistent across conditions. While bias/RMSEs were large, there were very high correlations between the true and estimated latent traits (all $r > 0.88$). Upon an analysis of the bias and RMSEs by timepoint (not reported here for brevity), we observed that recovery of the latent traits is worse for the first and last timepoints, a phenomenon we will return to in the Discussion.

Recovery of item and rater parameters

The mean absolute bias for item discriminations ranged from 0.045 to 0.088 and the RMSEs ranged from 0.062 to 0.114. Item difficulties and item step parameters had mean absolute bias that ranged from 0.061 to 0.185 and from 0.064 to 0.143, respectively. For some conditions, RMSEs for these parameters were very large in comparison to the mean biases.

Mean absolute biases for rater bias (ϕ) and standard deviation (τ) were small, ranging from 0.019 to 0.069 and 0.014 to 0.026, respectively. However, when compared to mean absolute biases, the RMSEs were fairly large for rater bias (0.039–0.230) because RMSEs as a measure of accuracy are more sensitive to outlying estimates. Mean absolute bias was smaller for ϕ when $N = 250$ and smaller for τ when $M = 7$ and/or $R = 6$. Figure 4 reveals how the

Table 1. Parameter recovery results from simulation study 1 where $\rho = 0.8$ and $\delta = 0.25$.

Condition			Growth (δ)		Autocorrelation (ρ)		SD of Traits (σ_θ)		Latent Trait (θ)		Item Discrimination (α)		Item Difficulty (β)		Item Step Parameters (γ)		Rater Bias (ϕ)		Rater SD (τ)	
M	*N*	*R*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*
3	250	3	0.066	0.079	0.025	0.031	0.098	0.125	0.362	0.459	0.088	0.114	0.096	0.193	0.120	0.178	0.044	0.113	0.026	0.046
3	250	6	0.062	0.071	0.026	0.031	0.094	0.120	0.350	0.442	0.083	0.110	0.103	0.285	0.115	0.236	0.045	0.147	0.022	0.057
3	500	3	0.044	0.052	0.019	0.023	0.092	0.111	0.355	0.449	0.064	0.084	0.095	0.327	0.118	0.299	0.053	0.182	0.022	0.044
3	500	6	0.044	0.055	0.020	0.023	0.089	0.110	0.348	0.441	0.058	0.077	0.103	0.365	0.111	0.309	0.051	0.178	0.019	0.058
7	250	3	0.080	0.095	0.017	0.022	0.104	0.134	0.345	0.438	0.063	0.083	0.095	0.276	0.098	0.217	0.042	0.134	0.021	0.040
7	250	6	0.074	0.088	0.015	0.020	0.095	0.118	0.336	0.424	0.051	0.066	0.061	0.076	0.064	0.083	0.019	0.039	0.014	0.027
7	500	3	0.058	0.065	0.011	0.013	0.101	0.128	0.344	0.436	0.046	0.062	0.137	0.474	0.120	0.383	0.060	0.190	0.018	0.036
7	500	6	0.058	0.070	0.010	0.013	0.106	0.125	0.338	0.427	0.045	0.063	0.185	0.588	0.143	0.469	0.069	0.230	0.015	0.049

Note. *Bias* is the mean of the absolute biases where bias is computed as the difference between the posterior median and the true value. *SD* = standard deviation; ρ = autocorrelation; δ = growth; *M* = number of timepoints; *N* = number of examinees; *R* = number of raters; *RMSE* = root mean square error.

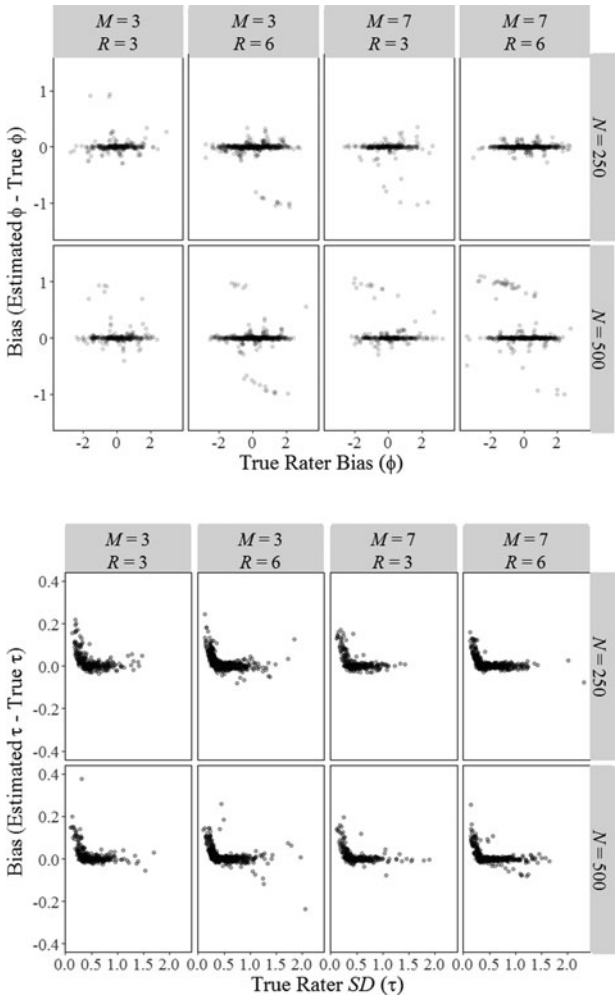

Figure 4. Bias plots for L-HRM rater bias (severity/leniency) and rater SD parameters from Simulation Study 1. The *top* plots show bias values for each replication over the true values of rater bias ϕ. The *bottom* plots show bias values for each replication over the true rater standard deviation τ. A transparency filter was applied to the plotted points to aid the reader. The shading of the points indicates the relative concentrations of plotted values such that there are more points in areas of darker shading. (Note: L-HRM = longitudinal hierarchical rater model; *SD* = standard deviation; M = number of timepoints; N = number of examinees; R = number of raters; τ = rater standard deviation; ϕ = rater bias).

bias per replication was distributed as a function of the true rater parameter values. The panel of plots on the top of Figure 4 show (rater-level) bias in rater bias ($\hat{\phi}_r - \phi_r$) computed using the posterior median of the rater bias parameter for all replications, by condition. While most bias was clustered around 0, there are scatters of replicates from various conditions with bias reaching from −1.0 to 1.0. An absolute ϕ value of 0.5 indicates a half score point in either negative or positive rater bias. Therefore, mean absolute bias for *some* conditions was substantial, but on average, minimal. The parameter recovery statistics we see in Table 1 are a function of these outlying values. In addition, interesting patterns appear along the scale of the *x*-axis—the L-HRM had a tendency to overestimate and underestimate at different values of the true ϕ's.

The panel of plots on the bottom of Figure 4 show (rater-level) bias in rater SD ($\hat{\tau}_r - \tau_r$) computed using the posterior median of the rater variability parameter for all replications by condition. Recovery of HRM rater standard deviation τ_r was most difficult when the true parameter value was less than about 0.25.

Study 2: Evaluate L-HRM recovery with varying levels of ρ and g

Study 1 fixed the longitudinal model parameters of the L-HRM so that growth was small ($\delta = 0.25$) and the autocorrelation was high ($\rho = 0.80$). To assess the parameter recovery of the L-HRM under a variety of longitudinal model conditions, we performed an additional simulation study that varied growth ($\delta = 0.25$, 0.50, and 0.75) and autocorrelation ($\rho = 0.00$, 0.30, 0.60, and 0.90) with a linear trend. There were a total of 12 conditions formed by fully crossing these growth and autocorrelation terms. We kept constant the number of timepoints ($M = 4$), the sample size ($N = 400$), the number of raters ($R = 10$), the number of items ($J = 10$), and the number of rating categories ($K = 5$).

Generation of ratings

Rater parameters, including rater bias (ϕ_r) and the rater SD (τ_r), were kept fixed across the 12 conditions. For all ten raters, the true ϕ_r values were, respectively, −1.00, −0.50, −0.50, −0.25, 0.00, 0.00, 0.25, 0.50, 0.50, and 1.00. For this respective list, the rater SDs (τ_r) began with τ_r set to 1.50, and then alternated between 1.50 and 1.00, respectively. These values relate to low levels of rater reliability, especially considering $K = 5$. Crossing the true rater bias and SD parameters, we yield a set of raters with high levels of variability with small to moderate levels of bias. Similarly, we simulated parameters for all ten items as fixed across conditions. Difficulty parameters varied from −1.25 to 1.25 in increments of 0.25. Discrimination parameters had a mean of 1.00 and a *SD* of 0.25. Step parameters varied from −1.27 to 1.76, with a mean of zero and a *SD* of 0.80. All growth terms were simulated as linear.

Latent traits were generated separately for each condition. Identical to the first study, we first drew an initial latent trait, θ_{i0}, from a *Normal*(0,1), then generated the remaining latent traits following (5). With these latent traits, rater, and item parameters, we first generated ideal ratings and then observed ratings for 100 replications per condition.

Model estimation and outcome measures

In Study 2 we estimated the L-HRM in exactly the same fashion (i.e. same prior distributions, convergence criteria, etc.) as in Study 1 except that instead of applying a sum-to-zero constraint on the discrimination parameters to address the scale indeterminacy, we applied a soft constraint by way of tightening the prior for ω from a *Gamma*(1,1) to a *Gamma*(10,10). Also identical to Study 1, we computed mean absolute bias and RMSEs for all parameters in each condition to determine under which settings parameter recovery is optimized but first rescaled the final estimates of the latent traits and item parameters using a linear transformation based on the estimated trait SD (within replication).

Recovery of longitudinal model parameters and latent traits

Table 2 reports the mean absolute bias and RMSEs for all parameters; the first four columns contain these quantities for the longitudinal model parameters and latent traits. Recovery of ρ, δ, and θs was better when the true ρ was larger. Comparatively, varying δ seemed to have little to no impact on any parameter's recovery except for σ_θ; the mean absolute bias and RMSEs of σ_θ were reduced as δ increased. Specifically, when the true growth δ was 0.25, the mean absolute bias of σ_θ was between 0.080 and 0.083; when δ was 0.50, the mean absolute bias was between 0.066 and 0.069; and finally, when δ was 0.75, the mean absolute bias was between 0.046 and 0051. RMSEs for δ were not very different from these mean absolute bias values.

Mean absolute bias and RMSEs for ρ were small, with the mean absolute bias ranging from 0.007 to 0.031. When $\rho = 0.9$, the bias ranged only from 0.007 to 0.009. Within each value of δ, as ρ increased, the estimation of ρ was less accurate (as measured by the mean absolute bias and RMSEs). The top plot in Figure 5 supports this observation with boxplots of the bias in ρ, by δ. Figure 5 also suggests that there are about the same number of positively and negatively biased estimates.

High ρ was related to smaller bias in δ, but different levels of growth did not impact the estimation of δ. The bottom plot in Figure 5 shows that bias in δ is nondirectional, and it also depicts the narrower distribution of biases with larger ρ. Manipulating δ had little to no effect on the quality of the θ estimates, which has median absolute biases ranging from 0.246–0.301.

Recovery of item and rater parameters

Overall the recovery of item and rater parameters in the present study was adequate and the bias/RMSEs of these parameters did not vary much by levels of growth and autocorrelation (see Table 2). The mean absolute biases were approximately 0.015 and 0.009 for ϕs and τs, respectively. The mean absolute biases were approximately 0.06 for αs, 0.06 for βs, and 0.11–0.12 for γs.

Discussion

In this article we introduced and tested a new parameterization of the HRM that permits the modeling of changes in latent traits over time using an overall trend and an autoregressive structure. The autoregressive component of the model includes an autocorrelation parameter that accounts for serial dependence between adjacent timepoints, m and $m - 1$. The overall growth trend can be flexibly specified.

We demonstrated a real-world situation in which the L-HRM might be useful. In the CDAP illustration, we

Table 2. Parameter recovery results from simulation study 2 where $M = 4$, $N = 400$, $R = 10$, $J = 10$, $K = 5$.

Condition		Growth (δ)		Auto-correlation (ρ)		SD of Traits (σ_θ)		Latent Trait (θ)		Item Discrimination (α)		Item Difficulty (β)		Item Step Parameters (γ)		Rater Bias (ϕ)		Rater SD (τ)	
δ	ρ	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*
0.25	0.00	0.058	0.074	0.028	0.033	0.083	0.086	0.297	0.376	0.056	0.072	0.056	0.070	0.110	0.143	0.016	0.020	0.008	0.011
0.25	0.30	0.056	0.070	0.025	0.032	0.082	0.085	0.296	0.375	0.055	0.071	0.054	0.069	0.117	0.152	0.015	0.019	0.009	0.011
0.25	0.60	0.055	0.068	0.020	0.025	0.082	0.086	0.285	0.360	0.056	0.073	0.055	0.071	0.114	0.148	0.015	0.019	0.008	0.011
0.25	0.90	0.033	0.041	0.009	0.011	0.080	0.085	0.246	0.310	0.055	0.071	0.062	0.079	0.114	0.148	0.015	0.019	0.008	0.011
0.50	0.00	0.057	0.071	0.031	0.037	0.069	0.072	0.300	0.379	0.053	0.069	0.052	0.066	0.115	0.150	0.015	0.018	0.009	0.011
0.50	0.30	0.055	0.070	0.023	0.030	0.068	0.072	0.298	0.376	0.055	0.071	0.056	0.069	0.119	0.156	0.015	0.020	0.009	0.011
0.50	0.60	0.059	0.074	0.018	0.022	0.066	0.072	0.288	0.364	0.058	0.073	0.062	0.079	0.116	0.152	0.015	0.019	0.009	0.011
0.50	0.90	0.039	0.048	0.007	0.010	0.066	0.071	0.247	0.312	0.060	0.077	0.061	0.077	0.116	0.151	0.015	0.019	0.008	0.011
0.75	0.00	0.050	0.064	0.026	0.033	0.051	0.057	0.301	0.383	0.056	0.071	0.053	0.066	0.118	0.155	0.015	0.020	0.008	0.011
0.75	0.30	0.058	0.073	0.022	0.027	0.046	0.053	0.299	0.380	0.054	0.070	0.052	0.067	0.115	0.151	0.015	0.019	0.008	0.011
0.75	0.60	0.059	0.075	0.020	0.023	0.047	0.053	0.289	0.366	0.055	0.071	0.059	0.075	0.117	0.155	0.015	0.019	0.009	0.011
0.75	0.90	0.040	0.050	0.008	0.011	0.046	0.052	0.247	0.313	0.054	0.070	0.057	0.073	0.115	0.150	0.016	0.021	0.008	0.011

Note. *Bias* is the mean of the absolute biases where bias is computed as the difference between the posterior median and the true value. *SD* = standard deviation; M = number of timepoints; N = number of examinees; R = number of raters; J = number of items; K = number of score levels; δ = growth; ρ = autocorrelation; *RMSE* = root mean square error.

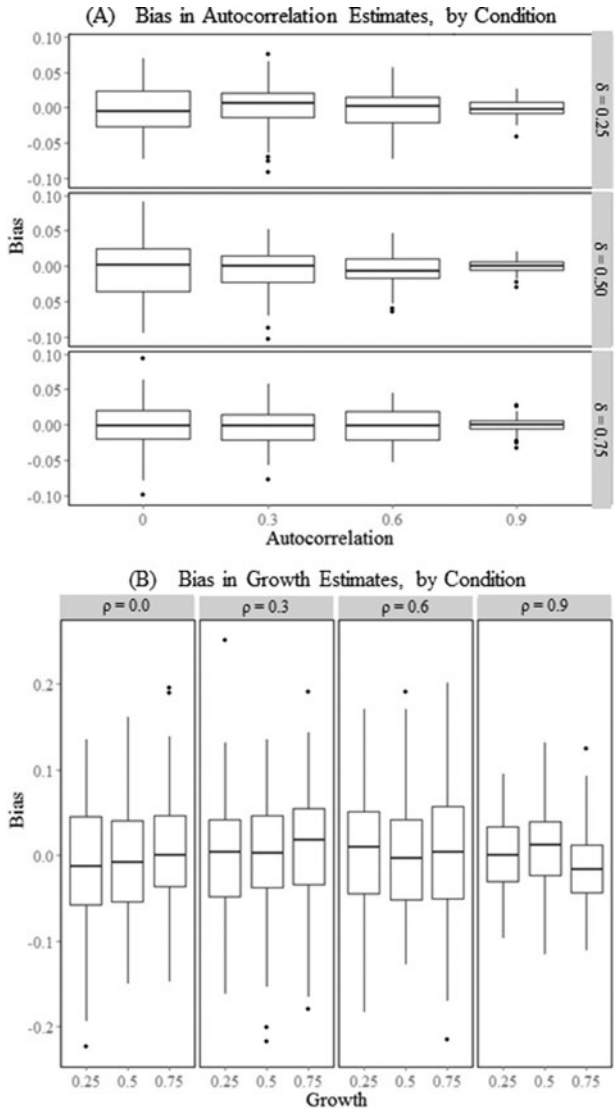

Figure 5. Bias in L-HRM parameter estimates from Simulation Study 2. Panel (A) shows bias in autocorrelation estimates by level of true growth and true autocorrelation. Panel (B) shows bias in growth estimates by level of true autocorrelation and true growth. (Note. δ = true growth level; ρ = true autocorrelation value; L-HRM = longitudinal hierarchical rater model).

were interested in estimating traits representing adolescents' character over four timepoints spanning the eighth and ninth grades. This study was different from other educational research studies in that there was no intervention and therefore no expectation of any strong trend for changes—the goals were simply to estimate overall changes and study each student's character over time while incorporating information about their traits from timepoint to timepoint. We found that several of the teacher raters in CDAP were lenient in their evaluation of students' character strength, relative to the ideal ratings. Examining pictorial displays of rater parameter estimates provides a mechanism by which to recognize raters who are evaluating children with large bias and/or unreliability. In terms of growth, we also found a slight decline in character strength of about half a SD. The very strong estimated autocorrelation shows the importance of using prior levels of character strength to predict the future; the value of modeling an autoregressive structure was also demonstrated in Study 2. These trait estimates can be used to compare adolescents or as outcomes in subsequent analyses for the CDAP.

Two simulation studies evaluated the L-HRM. Study 1 tested this parameterization under various levels of sample size, number of raters, and number of timepoints using a linear trend where growth was fixed to be 0.25 of a SD of the trait, the SD of the trait was fixed at 1.0, and autocorrelation was fixed at 0.80. Recovery for some parameters was excellent—mean absolute bias and RMSEs were minimal for the rater parameters, growth, and autocorrelation, (however, RMSEs for the rater bias parameters were large). Mean absolute bias and RMSEs for item and trait parameter estimates were also larger than desired. However, it is not surprising given that there are only five items on our simulated assessment, which may be considered insufficient to adequately support the estimation of item and trait parameters (Mariano & Junker, 2007). Mean absolute bias and RMSEs for ρ estimates were smaller for $M = 7$, which is consistent with a general result found in the literature that states that estimation of ρ is acceptable if $q \leq M/4$ (Box et al., 2013) where q is the order of the time series model. This condition was *not* satisfied when $M = 3$. Generally, time series models are popular in scenarios where there are numerous timepoints, however, the modeling of autoregressive processes is gaining popularity in the social sciences where fewer timepoints are observed (see examples in: Andrade & Tavares, 2005; Azevedo et al., 2015; Cagnone et al., 2009; Dunson, 2003).

The phenomena that we observed in the plots for ϕ and τ (Figure 4) come as no surprise. Any value for τ less than 0.25 or so is consistent with observing $X = \xi + \phi$, and so τ is not well identified (Patz et al., 2002). This difficulty is apparent in the bias plots for τ where bias for true τ values larger than 0.25 were clustered around 0, however, bias for values below 0.25 were larger. Additionally, Patz et al. (2002) observed that more reliable raters, with lower values of τ, tend to have the "least-well" estimated ϕ parameters, and vice versa. They noted the cause might be the use of a continuous rating bias parameterization to model discrete shifts in the observed rating away from the ideal rating. That is, when a rater is very consistent in their scoring, their rating bias will be within some one-point range on the score scale. However, because they are so consistent, it is difficult to pin-point exactly where they lie in this narrow range. This continuous treatment of a discrete quantity also explains the patterns of bias for ϕ which clearly shows optimal estimation of ϕ at the exact

discrete score points. Figure 4 shows that overestimation occurred at true bias levels that were just under a discrete score point and underestimation occurred at true bias levels that were just over a discrete score point—the best estimation occurred when bias was a whole number. In other words, the model provided an estimate of bias aligned to the nearest whole number on the bias scale. In terms of bias, recovery of the rater parameters was exceptional, however, we did observe large RMSEs which is partly a manifestation of this issue. Improvements could certainly be made to the model to mitigate this problem.

Study 2 examined recovery of L-HRM parameters under different values for δ and ρ for a 10-item test given to 400 examinees at 4 timepoints, and evaluated by 10 raters. There were no patterns for different levels of growth δ and autocorrelation ρ revealed in the mean absolute bias or RMSEs of item or rater parameters. Recovery of ρ was insensitive to growth. Recovery of ρ improved with larger values of ρ—the variability in bias in ρ across replications decreased as ρ increased. Recovery of both traits and the growth parameter also improved as ρ increased. Given this result, we conclude that if there is indeed truly serial dependence between latent traits over time, inclusion of that in the model will provide information to the estimation process and thereby yield better trait estimates and detection of accurate growth trends.

RMSEs for rater bias and item step parameters were substantially smaller in Study 2 than in Study 1. This is in part because in Study 1 we used a random effects approach for generating true parameter values. In each replication, rater biases and item step parameters were drawn from a $N(0,1)$, leading to some simulated raters having large rater bias values ($\phi_r > 2$) and some simulated items having very large item step parameters, which are both more difficult to estimate, thereby yielding larger biases. In Study 2, true rater bias values were fixed across conditions and replications to values less than or equal to $|1.0|$, leading to less sampling variance. The latent trait estimates were also better recovered under Study 2 which may be a function of the way we generated data (i.e. based on the same fixed set of item parameters for all replications and conditions). Another obvious explanation is the difference in test length ($J = 5$ in Study 1 vs. $J = 10$ in Study 2).

In both studies we observed larger variation at the first and last timepoints which actually initially presented as bias until we further investigated. While not well documented in the educational statistics and psychometrics literature, this is a phenomenon that occurs when using time series models with certain estimation procedures. Specifically, smoothing procedures, such as Gibbs sampling, will incorporate information from time $m - 1$ and $m + 1$ in the measurement at time m. The first and last timepoints do not benefit from this smoothing as we do not have both estimates of both θ_{m-1} and θ_{m+1}. Therefore, we generally end up with more precise estimates at each timepoint in between the first and last, as the variance is reduced based on our knowledge of what occurred in the past and what occurs in the future.[5] Other approaches will treat this differently. For example, a filtering procedure will only make use of the information at time $m - 1$ in the measurement at time m. With filtering, we would expect to see an inflated variance only in the first timepoint (although variances at all timepoints may be larger, because we would not be conditioning on information at time $m + 1$). Fully Bayesian computing solutions for smoothing are readily available, however, this is not true for filtering. Future work will incorporate a Kalman filter (Kalman, 1960), for example, into an MCMC-like estimation algorithm, and investigate the differences between smoothing and filtering approaches.

Future directions

The simulation studies only investigated a linear trend, however, our preliminary analyses comparing linear and logistic trends showed very few differences in terms of parameter recovery. The logistic trend had better recovery with larger sample sizes and more timepoints. This may be related to the nature of the logistic curve reaching an asymptote as it approaches positive and negative ∞, thereby making it easier to detect change (or lack thereof) over a longer duration. While the inclusion of a linear, logistic, or some other deterministic trend is a straightforward process, it is an oversimplification of the possible trajectory of traits over time. The ability to include a flexible specification of trends, using B-splines, for example, would better capture the true changes in traits. Further, this research was focused on the problem similar to the one that the CDAP researchers were facing—examining overall growth. However, individual growth may be of interest. If this is the case, then individual, possibly varying types of, growth terms could be modeled. We chose to keep our investigation to overall growth only for parsimony.

This parameterization of the L-HRM accounts for changes in traits only. The benefit of this model is that it permits us to estimate serially dependent traits, while accounting for rater bias and SD, assuming these two characteristics of raters remains the same over time. However, the L-HRM is indeed a "rater" model, and thus it is intuitive to conceptualize a version of this model that accounts for rater drift in bias and changes in rater SD. As mentioned throughout this article, decoupling

[5] Figure A2 in the Supplementary Material provides a pictorial depiction of this phenomenon.

changes in examinees and raters is possible with very specific rating collection designs (see Casabianca et al., 2015). If the design facilitated such an analysis, the time series and/or growth components could be added to the signal-detection portion of the model in order to estimate rater parameters at each timepoint and model overall and/or individual growth.

Article Information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant SES 1324587 from the National Science Foundation.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: This work was supported by Grant SES 1324587 from the National Science Foundation. We are grateful to the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high performance computing resources that have contributed to the research results reported within this paper (URL: http://www.tacc.utexas.edu). We are especially grateful to David Walling, Ruizhu Huang, and Lei Huang for their personal assistance with implementing parallelization on the supercomputer, Stampede. This work also used resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Special thanks to Yisi Wang (former student at UT Austin) for her data analysis work early in the project. We also would like to gratefully acknowledge Angela Duckworth for granting permission to use data from the Character Development in Adolescence Project (CDAP; https://coa.stanford.edu/content/character-development-adolescence). The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or the National Science Foundation is not intended and should not be inferred.

References

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*(1), 3–16. https://doi.org/10.1007/BF02294143

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: population parameter estimation. *Journal of Multivariate Analysis, 95*(1), 1–22. https://doi.org/10.1016/j.jmva.2004.07.005

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573. https://doi.org/10.1007/BF02293814

Azevedo, C. L., Fox, J. P., & Andrade, D. F. (2015). Longitudinal multiple-group IRT modelling: covariance pattern selection using MCMC and RJMCMC. *International Journal of Quantitative Research in Education, 2*, 213–243. https://doi.org/10.1504/IJQRE.2015.071737

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2013). *Time series analysis: Forecasting and control*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781118619193

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag. doi: 10.1007/978-1-4757-3456-0

Cagnone, S., Moustaki, I., & Vasdekis, V. (2009). Latent variable models for multivariate longitudinal ordinal responses. *British Journal of Mathematical and Statistical Psychology, 62*(2), 401–415. https://doi.org/10.1348/000711008 × 320134

Casabianca, J. M., Junker, B. W., & Patz, R. (2016). The hierarchical rater model. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (pp. 449–465). Boca Raton, FL: Chapman & Hall/CRC.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*, 311–337. https://doi.org/10.1177/0013164414539163

DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48*(3), 333–356. https://doi.org/10.1111/j.1745-3984.2011.00143.x

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment, 91*, 166–174. https://doi.org/10.1080/00223890802634290

Dunson, D. B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association, 98*(463), 555–563. https://doi.org/10.1198/016214503000000387

Eid, M., & Hoffmann, L. (1998). Measuring variability and change with an item response model for polytomous variables. *Journal of Educational and Behavioral Statistics, 23*(3), 193–215. https://doi.org/10.1198/016214503000000387

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*(3), 495–515. https://doi.org/10.1007/BF02294487

Embretson, S. E. (1997). Structured ability models in tests designed from cognitive theory. In M. Wilson, G. Engelhard, Jr., & K. Draney (Eds.), *Objective measurement: theory into practice* (Vol. 4, pp. 223–236). Greenwich, CT: Ablex.

Froh, J. J., Fan, J., Emmons, R. A., Bono, G., Huebner, E. S., & Watkins, P. (2011). Measuring gratitude in youth: Assessing the psychometric properties of adult gratitude scales in children and adolescents. *Psychological Assessment, 23*(2), 311–324. https://doi.org/10.1037/a0021590

Fu, Z. H., Tao, J., Shi, N. Z., Zhang, M., & Lin, N. (2011). Analyzing longitudinal item response data via the pairwise fitting

method. *Multivariate Behavioral Research, 46*(4), 669–690. https://doi.org/10.1080/00273171.2011.589279

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 733–760. Retrieved from http://www.jstor.org/stable/24306036

Gelman, A., & Rubin, D. B. (1996). Markov chain Monte Carlo methods in biostatistics, *Statistical Methods in Medical Research, 5*(4), 339–355. https://doi.org/10.1177/096228029600500402

Guo, S. (2014). Correction of rater effects in longitudinal research with a cross-classified random effects model, *Applied Psychological Measurement, 38*(1), 37–60. https://doi.org/10.1177/0146621613488821

Hambleton, R. K., van der Linden, W. J., & Wells, C. S. (2010). IRT models for the analysis of polytomously-scored data: Brief and selected history of model building advances. In R. Ostini & M. Nehring (Eds.), *Handbook of polytomous item response theory models* (pp. 21–42). London: Routledge Academic.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43–58. https://doi.org/10.1111/j.1745-3984.2009.01068.x

Hershberger, S. L., Molenaar, P. C. M., & Corneal, S. (1996). A hierarchy of univariate and multivariate structural time series models. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 159–194). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*(2), 371–384. https://doi.org/10.17763/haer.83.2.d11511403715u376

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

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*, 64–86. https://doi.org/10.1037/1082-989X.5.1.64

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*, 403–424. https://doi.org/10.1037/1082-989X.4.4.403

Hung, L. F., & Wang, W. C. (2012). The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics, 37*(2), 231–255. https://doi.org/10.3102/1076998611402503

IES (Institute of Education Sciences). (2010). Efficacy of schoolwide programs to promote social and character development and reduce problem behavior in elementary school children. Retrieved from http://ies.ed.gov/ncer/pubs/20112001/index.asp

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering, 82*(1), 35–45. https://doi.org/10.1115/1.3662552

Kane, T.J., & Staiger, D.O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from ERIC database. (ED540962).

Kingsbury, F. A. (1922). Analyzing ratings and training raters. *Journal of Personnel Research, 1*, 377–383.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399–418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506. https://doi.org/10.1037/met0000075

Mariano, L. T. (2002). *Information accumulation, model selection and rater behavior in constructed response student assessments* (Doctoral dissertation). Carnegie Mellon University.

Mariano, L. T., & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics, 32*, 287–314. https://doi.org/10.3102/1076998606298033

McArdle, J. J., Petway, K. T., & Hishinuma, E. S. (2015). IRT for growth and change. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 435–456). New York: Routledge.

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

Millsap, R. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4*, 5–9. https://doi.org/10.1111/j.1750-8606.2009.00109.x

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–177. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

Patz, R. J. (1996). *Markov chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress* (Doctoral dissertation). Carnegie Mellon University.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*, 341–384. https://doi.org/10.3102/10769986027004341

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Paper presented at the 3rd International Workshop on

Distributed Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Roberts, J. S., & Ma, Q. (2006). IRT models for the assessment of change across repeated measurements. In R. W. Lissitz (Ed.), *Longitudinal and value added modeling of student performance* (pp. 100–127). Maple Grove, MN: JAM Press.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*(4), 1151–1172. https://doi.org/10.1214/aos/1176346785

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*(4), 375–394. https://doi.org/10.1111/j.1745-3984.2005.00021.x

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*(4), 298–321. https://doi.org/10.1177/0146621605285517

Stanford Center on Adolescence. (2014). *Character development in adolescence*. Retrieved from https://coa.stanford.edu/content/character-development-adolescence

Su, Y. S., & Yajima, M. (2012). R2jags: A package for running jags from R. R Package Version 0.03-08.

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In A. Boomsma, M. van Duijn, & T. Snijders (Eds.), *Essays on item response theory* (pp. 89–108). New York, NY: Springer-Verlag.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283–306. https://doi.org/10.3102/10769986026003283

Wolfe, E. W. (2014). *Methods for monitoring rating quality: Current practices and suggested changes*. (White Paper). Iowa City, IA: Pearson Education.

Xu, W., Huang, R., Zhang, H., El-Khamra, Y., & Walling, D. (2016). Empowering R with high performance computing resources for big data analytics. In R. Arora (Ed.), *Conquering big data using high performance computing* (pp. 191–217). New York, NY: Springer-Verlag.