The Hierarchical Rater Model for Longitudinal Ratings

**Abstract**

Rater effects in education testing and research have the potential to impact the quality of scores in constructed response and performance assessments. The hierarchical rater model (HRM) yields estimates of latent traits that have been corrected for individual rater bias and variability. We develop an extension of the HRM for ratings that come from longitudinal designs that includes an autoregressive time series component as well as a parameter for overall growth called the longitudinal HRM (L-HRM). We evaluate and demonstrate L-HRM feasibility and performance using a simulation study varying the type of trend and the number of raters, time points and sample size. Parameter recovery results reveal negligible bias for most parameters across conditions. We discuss limitations and future research underway to improve the L-HRM.

*Keywords*: item response theory, hierarchical rater model, ratings, longitudinal, trends, time series

The Hierarchical Rater Model for Longitudinal Ratings

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 [STAAR] 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 Project; Bill and Melinda Gates Foundation, 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). These scenarios in the education landscape demonstrate the need for advanced psychometric models that account for rater effects, or the error introduced into scores or trait estimates by the rating process. There are several types of rater effects (see Wolfe [2014] for an in-depth discussion), but the most common rater effects discussed in the literature include rater bias, a rater’s tendency to score higher (leniency) or lower (severity) on average, and rater centrality, a rater’s tendency to use the middle of the score scale versus the full extremes.

The need to account for rater effects extends to the longitudinal context where we seek to evaluate changes over time due to some intervention or naturally occurring growth. For example, in an IES-funded study, teachers and parents completed surveys to evaluate the behavior of children in order to determine if an intervention to improve character development was effective (IES, 2010). This is an instance where a longitudinal model is necessary to estimate traits or detect changes in traits over time points, but it is also necessary to account for rater effects introduced by the rating process.

There is no shortage of cross-sectional latent trait 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 hierarchical rater model (HRM; Patz, Junker, Johnson, & Mariano, 2002), Verhelst and Verstralen’s (2001) IRT model for multiple raters, and Wilson and Hoskins (2001) rater bundle models. Other models, such as the 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 the trait estimation thereby leading to trait estimates with falsely low standard errors (shown in Patz, et al, 2002; Mariano & Junker, 2007).

Longitudinal models for ratings are available. Hung and Wang (2012) proposed a generalized multilevel facets model applied to longitudinal ratings data. This model included three levels: level 1 models the item responses at specific time points using a facets model which treats rater severity as a random effect; level 2 models latent growth with an autoregressive residual structure to account for variation over time points in latent traits; and level 3 models variation in growth between examinees. This model assumes that item parameters are constant over administrations, but estimates time-specific latent traits and rater severity. Casabianca, Lockwood, and McCaffrey (in press) apply an augmented parameterization of a generalizability study model using B-splines to model changes in examinees and raters. Their model includes random and fixed effects for scoring trends in order to determine how individual raters change during the scoring period, how examinees change over time, and as well as how rater variation changes during the scoring period. Their work assumes that trends for the examinee and the rater can be decoupled. Hung and Wang (2012) do not address the conflation of examinee and rater trends and report parameter recovery for rater severity over time.

Our research focuses on the HRM, a multilevel IRT model that includes a level in between the latent trait and scores given by raters. This level of “ideal ratings” accounts for the nesting structure with item-level latent variables to represent the perfect rating associated with a response to a particular item. The HRM, however, does not include a mechanism by which we can model changes in longitudinal ratings data. In this paper 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. Longitudinal designs need not be associated with a specific intervention and there need not be any hypothesized changes in traits; the rating collection design simply must collect ratings from the same individuals over a collection of time points. This initial presentation of the L-HRM assumes that rater behavior does not change over time and keeps rater parameters, as well as item parameters, fixed. In the following sections, we first reintroduce the basic HRM and then introduce the L-HRM. Then, to evaluate this extension we present results from a simulation study and discuss further extensions and limitations.

**The Hierarchical Rater Model**

The HRM is a three-level hierarchical model for ratings data. 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 . The hierarchy is given by

 

Here, , the latent trait for examineeis normally distributed with meanand , is the ideal rating for examinee *i* on item *j* (*j*)and  is the observed rating given by rater *r* for examinee *i*'s response to item *j*. Note that specifying a normal distribution for the latent trait is a popular choice, but alternatives could be used.

This hierarchy connects a two-stage measurement process; the first stage is a “signal-detection-like'' model for measuring the ideal rating of an indicator based on multiple raters' observed ratings; and the second stage is an IRT model relating the ideal ratings to the latent trait variable, in this context. In other words, an examinee's work on *J* items may be hypothetically judged to have some true rating or quality, we call this the “ideal rating” and then a series of *R* raters evaluate the work, giving ratings conditional on the examinee's overall trait level. This notation is for the completely crossed design where all raters score each item. Within this framework incomplete designs are treated as missing completely at random (MCAR; Mislevy & Wu, 1996; Little & Rubin, 2002). Models for informative missingness (e.g. Glas & Pimentel, 2008; Holman & Glas, 2005) could also be incorporated directly into the HRM if needed.

 Suppose we have *N* persons to be rated and within each is a latent trait, ideal ratings per-item, and observed ratings per-item from each rater. In the second level, the ideal ratings represent the quality of person *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 , the item discrimination, , the item location, and the  threshold parameter for item *j*, or the locations on the scale of the latent trait distinguishing points between discrete score levels. 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  and *K* possible scores , the GPCM is given by:

 

Note, 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 ratingsand these ideal ratings are modeled using a discrete signal detection model which is specified to represent the quality of the response. A matrix of response probabilities defines the relationship between the observed and ideal rating probabilities such that . A simple signal detection model uses a discrete unimodal distribution for each row of the matrix to give the probability of observed rating  given ideal rating . The mode of this distribution is the rater bias, , and the spread of this distribution is the rater variability or unreliability, . The signal detection model can be specified such that probabilities in each row of the matrix are proportional to a Normal density with mean  and standard deviation :

 

The bias parameter 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 indicates a rater's variability; values near 0 indicate high consistency or reliability in rating (to the rubric or scoring guidelines) and high values indicate poorer consistency in rating.

--------------------------------------

Insert Figure 1 here

--------------------------------------

 To discuss the structural assumptions of the HRM, we provide itsdirected acyclic graph (DAG) in Figure 1 as originally given by Johnson, Sinharay, and Bradlow (2007). A DAG defines the conditional relationships between model quantities (“nodes”) by using directed edges (arrows). An arrow connecting two nodes indicates that the node from which the arrow originates (say “X”) influences the node to which the arrow is pointing (say “Y”). In this case we say that X is the “parent” of the “child” Y. In general, it is assumed that a quantity Y is conditionally independent of all nondescendant quantities given its parents (all nodes pointed to the Y).

Starting from the bottom of Figure 1, the observed rating  is determined by , which is a function of the latent trait , the characteristics of item *j* (), and characteristics of the rater *r* (). The observed rating is conditionally independent of the observed rating for other items and given by other raters  given . Further, the observed ratings assigned by rater *r* are conditionally independent across items and examinees given 

Estimation of the HRM as a Bayesian hierarchical IRT model is straightforward with Markov chain Monte Carlo (MCMC) estimation. This was documented in detail in Patz et al (2002). From the DAG we observe the parentless nodes that would be treated as random quantities and therefore need prior distributions: 

**The Longitudinal Hierarchical Rater Model (L-HRM)**

Time trends could be specified using a variety of models, however, we started with time series models (Hamilton, 1994; Hershberger, Molenaar, & Corneal, 1996). In the L-HRM, the latent traits are modeled within an autoregressive time series model of order 1 (AR[1]) and an overall growth trend that can take on any shape, including linear, logistic, cubic, etc. The AR(1) component accounts for the autocorrelation between time points and the trend accounts for an overall shift in the population of examinees.

Suppose a longitudinal rating design to be analyzed with the HRM has *M* time points with ideal and observed ratings at time *m* nested within each , where  is the trait for examinee *i* at time point . Now, instead of assuming some normal distribution for the traits, we model them using a longitudinal model, which is part trend and part autoregressive time series model of order 1, meaning that each time point is modeled using information only from the most previous time point. The other levels are the same but now we have a set of ideal ratings and observed ratings at each timepoint *m*. Item parameters and rater parameters remain the same over time.

 

In order to estimate, let the trait for subject *i* at time point *m* be a function of two quantities:

 

Here, is the trend in  at time *m* and  is the time series component of the model. We can decompose  into a function of the parameter *g* for overall growth, depending on the type of trend. A linear trend is a straightforward and reasonable approach to modeling linear relationships (whether positive or negative). If a linear trend, for example, then  decomposes into  where  the lagged value of  is weighted by  the autocorrelation parameter. In the second term the random noise for examinee *i* at time point *m* denoted  is distributed as  and is weighted by a function of the autocorrelation . Weighting  assures stationary variance of the noise (and therefore the resultant traits) across *M* time points. Together, the sum of the two weighted quantities  makes up the component of  that incorporates the individual’s trait information; the other component incorporates an average trend for all *N* examinees.

 To provide a better framework for discussing the longitudinal piece of the L-HRM below, we restate the model in using two steps which are implemented at each time point *m*:

 Step 1—The AR(1) process with no trend:  When *m* and there is no lagged value, we define . When m > 1, we place a normal prior with different parameters on this quantity, namely, . Restated in terms of precision instead of variances, these priors are  and  where precision 

 Step 2—Compute the trait from estimated parameters: 

 Performing these steps for each time point allows us to avoid a conflation of the trend and the lagged values. Structurally, the L-HRM is a simple extension of the HRM. Figure 2 shows the DAG for the L-HRM. Here the latent trait is now indexed for time point *m* and is a descendant of two new parameters,  and *g*. The variability of  is parameterized in terms of .

--------------------------------------

Insert Figure 2 here

--------------------------------------

**Method**

We tested the L-HRM parameterization with a simulation study that examined feasibility of parameter recovery with smaller samples (*N* = 250, 500) while varying the number of raters (*R* = 3, 6), time points (*M* = 3, 7) and type of trend (linear, logistic). Our original intentions were to also vary the number of items (*J* = 5, 20), however, the *J*=20 conditions required an exorbitant amount of computing resources and therefore we excluded that condition from the study. The final set of results includes 14 conditions: 8 conditions crossing *R*, *N* and trend type for *M* = 3, and an additional 6 conditions crossing the same factors for *M* = 7 with the exclusion of conditions where *R* = 6 and *N* = 500.

**Generation of Ratings**

To maximize our ability to generalize, we chose to consider traits, rater and item parameters as random effects and randomly sampled true parameter values from specified distributions. True rater bias (ϕr) and SD (τr) for *R* raters were drawn from *N*(0,1) and *lognormal*(1, 0.25), respectively. True GPCM item parameters for five items with five response categories 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 autocorrelation parameter was also fixed to 0.80 giving observations between times *m* and *m*-1 a strong correlation. The standard deviation of θ was fixed at  In this study, we tested a linear trend, , and a logistic trend,  , with the growth *g* fixed at 0.25. Note that we have set *g* = 0.25 but we are modeling raw growth as  over the length of observation period. In this way we estimate growth in standard deviation units, making the amount of change easier to conceptualize. Therefore, when *g* = 0.25 that is only a quarter of a ω increase over *M* time points. Since , then growth , and raw growth is equal to growth in standard deviation units. Estimates of growth will rely on the estimated values of ω.

We generated ratings data by drawing the baseline trait level  from a *N*(0,1) distribution and using the time series model in (1) to compute the additional  Using the true values for traits, raters and items, we generated ideal ratings first using the GPCM as shown in . The generated ideal ratings were then used in the SDM as shown in (2) to generate observed ratings. We generated 100 replications for each condition.

**Estimation of L-HRM Parameters**

We placed noninformative priors on all parameters (see Table 1). For the three parameters with non-negative values  we used a Gamma(1,1), Normal priors for *g*, , and a uniform(-1,1) for ρ.

--------------------------------------

Insert Table 1 here

--------------------------------------

We fitted the L-HRM as a Bayesian model using MCMC estimation in Jags (Plummer, 2003) via the R2jags package (Su & Yajima, 2012) with 8,000 iterations, 3 chains, burn-in of 2,000 and thinning every 6. We followed the two-step approach as detailed in the previous section using the R2Jags code provided in the Appendix. We relied on computational resources from the Texas Advanced Computing Center’s (TACC) supercomputer, Stampede[[1]](#footnote-1), where replications were parallelized across computing nodes to minimize computation time. We evaluated each replication for the convergence of chains according to the Gelman-Rubin convergence diagnostic (R-hat; Gelman & Rubin, 1996); a replication was excluded if there was one or more estimated parameter with R-hat > 1.1. Values higher than 1.1 indicate that the chains did not converge. Upon the rejection of a replication for this reason, we then estimated additional replications so that our final set of results included 100 converged replications for each condition. After a close assessment of convergence rates, we identified the largest issue with the convergence to be with the precision of the traits, 

**Results**

This purpose of this study was evaluate the recovery of parameters, particularly the parameters associated with the longitudinal model component and latent traits. To evaluate results, we report the median absolute bias for all parameters and conditions. Where appropriate, we also provide plots of bias as a function of true parameter value to highlight the variation in bias. We computed bias as the difference between the true parameter value and the posterior median, .

--------------------------------------

Insert Table 2 and Figure 3 here

--------------------------------------

**Recovery of Longitudinal Model Parameters and Latent Traits**

Table 2 provides the median absolute bias across replications for growth, autocorrelation, the standard deviation of latent traits, and estimates of latent traits at each time point. True growth was fixed at 0.25 standard deviation units such that there was a growth of 0.25 latent trait over *M* time points. The growth parameter, *g*, had bias that ranged from 0.044 to 0.077. Bias was consistently smaller with the N=250 conditions (versus the N=500 conditions). There were no striking differences between the two trend types, however, for the linear trend we observed that the M=7 cases yielded larger bias than their M=3 counterparts. This phenomenon did not occur in the logistic trend.

Estimation of the autocorrelation yielded very little bias, ranging from 0.011 to 0.029. Recall that the autocorrelation was fixed at 0.80, which represents the correlation between adjacent time points as specified in the time series model. No difference appears between the trend types, however, we do observe better recovery with the N=500 conditions and the M=7 conditions, the combination of those two yielding the least bias.

The standard deviation of θ, or ω, was fixed at 1.0; this value keeps the standard deviations of latent traits consistent over time. Median absolute bias for ω ranged from 0.272 to 0.319. Figure 3 provides an index plot of bias values for ω for different study conditions. Here, the N=250 conditions are in light grey and the N=500 conditions in black symbols. From this plot we observe that the bias is mostly positive and that the model mainly overestimated the standard deviation of θ, in a few outlying cases, by more than 1.5.

Table 2 provides median absolute bias for latent traits overall and by time point. Values ranged from 0.405 to 0.444. Overall, bias was lower when M=7. Further, in the M=3 conditions, having more raters (and therefore more ratings), appears to be associated with lower bias. While bias was large, there were very high correlations between the true and estimated latent traits (all > 0.86). The recovery of ω undoubtedly affects latent trait recovery; there is a strong correlation between the median absolute bias for parameters ω and θ, *r*(14)=0.787, p = 0.001. For both M=3 and 7 we observe that recovery of the latent traits is worse for the first and last time points.

**Recovery of Item and Rater Parameters**

Median absolute bias for rater bias (ϕ) and standard deviation (τ) was small, ranging from 0.027 to 0.087 and 0.015 to .029, respectively (see Table 3). There appears no pattern related to level of *R*, however, recovery of ϕ and τ was mostly better when N=500. Figures 4 and 5 reveal how the bias per replication was distributed as a function of the true parameter values for rater bias. The panel of plots on the top of Figure 4 show rater-level bias computed using the posterior median of the rater bias parameter (ϕ) for all replications by condition for the linear trend, and the same for the logistic trend in the bottom panel. While most bias was clustered around 0, there are scatters of replications from various conditions with bias reaching from -1.0 to 1.0. Recall that an absolute ϕ value of 0.5 indicates a half score point in either negative or positive rater bias. Therefore, median absolute bias for some conditions was substantial, but on average, minimal. 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 ϕs. It seems 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. This phenomenon occurred for both linear and logistic trends.

-----------------------------------------------------------

Insert Table 3, Figure 4 and Figure 5 here

-----------------------------------------------------------

The panel of plots on the top of Figure 5 show rater-level bias computed using the posterior median of the rater variability parameter (τ) for all replications by condition for the linear trend, and the same for the logistic trend in the bottom panel. Recovery of HRM rater standard deviation  is most difficult when the true parameter value is less than about 0.5. This difficulty is apparent in the bias plots for τ. Bias for true τ values larger than 0.5 were clustered around 0, however, bias for values below 0.5 were large. In fact, the bias plots show an elbow shape reflecting a dramatic decrease in bias as τ increases from just above 0 to .5.

The median absolute bias for item discriminations ranged from 0.219 to 0.295. The N=500 conditions generally had less bias than when N=250. While there was no apparent trend related to R for the logistic trend conditions, the linear trend conditions has smaller median absolute bias when R=3. There were no distinct differences between M levels. Item difficulties and item step parameters had median absolute bias that ranged from 0.075 to 0.157 and from 0.171 to 0.213, respectively. Interestingly, the conditions that yielded the best parameter estimates were M=7, N=250 and R=3 for the linear trend, and M=3, N=500, and R=6 for the logistic trend. Though inconsistent across trend type, these were the conditions that consistently yielded the least bias (for most or all parameters).

**Discussion**

In this research we introduce a new parameterization of the HRM that permits the modeling of changes in latent traits over time using a time series model. The parameterization presented here relies on an autoregressive time series model of order one. This component of the model includes an autocorrelation parameter that accounts for serial dependence between two adjacent time points, *m* and *m*-1. The L-HRM also includes an overall growth trend that can be flexibly specified. This parameterization was tested for various level of sample size, number of raters, and number of time points using a linear and logistic trend where growth was fixed to be 0.25 of a standard deviation of the trait, the standard deviation of the trait was fixed at 1.0, and autocorrelation was fixed at 0.80.

Recovery for most parameters was excellent—bias was minimal for the rater parameters, growth, and autocorrelation. Bias for item parameters was larger than expected. Median absolute bias in ρ was smaller for M=7, which is consistent with a general result found in the literature that states that estimation of ρ is acceptable if k ≤ M/4 (Box, Jenkins, & Reinsel, 2013) where k is the order of the time series model. This condition was not satisfied when M=3.

The phenomena that we observed in the plots for ϕ and τ come as no surprise. 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, thus leading to more variation in estimation. 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.

Perhaps the most informative result was yielded from the bias for ω and. The median absolute bias for these parameters was large considering the scale of the variables. We placed a noninformative prior on the precision of θ: . This parameter keeps variance stationary over time. We presented the results in terms of ω; plots of the bias per conditions and replications (Figure 3) revealed substantial positive bias. We believe a consequence of this positive bias is the extreme bias seen in latent trait estimates. That is, the standard deviation of the latent trait is estimated too large, leaving the individual latent trait estimates too spread out on the scale. However, it should be noted that the correlations between the true and estimated traits were high in every condition meaning that although there was large bias in estimating traits, they maintained a similar sorting to the true values. This result has already informed our decisions for future studies—we will change the prior on the precision to keep it closer to 1.0.

Results also showed larger variation at the first and last time points which actually presents as bias in summary measures. 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 *t*-1 and *t*+1 in the measurement at time *t*. The first and last time points do not benefit from this smoothing as we do not have both estimates of both θt-1 andθt+1. Therefore, we generally end up with more precise estimates at each time point 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. Other approaches will treat this differently. For example, a filtering procedure will only make use of the information at time *t*-1 in the measurement at time *t*. With filtering, we would expect to see larger variance only in the first time point. Figure 6 plots the posterior standard deviations by time point for all 100 replications of the linear, M=7, R=3, N=250 condition. The reader may observe the positive shift in the distributions for m=1 and m=7. The larger variability explains the larger biases observed in Table 2 for the latent trait at these time points. Computing solutions for smoothing are readily available, however, this is not true for filtering. Further work is needed to develop a Kalman filter (Kalman, 1960) as well as to investigate the differences between these approaches.

--------------------------------------

Insert Figure 6 here

--------------------------------

We experienced computing challenges that prevented us from completing the replications for the study. Depending on the size and complexity of the condition, any one replication of a L-HRM fitting in Jags required between 5 to 35 hours to complete (with 8,000 iterations, burn-in of 2000, thinning every 60th iteration). Fortunately, replications were run in parallel so that several were in process simultaneously. Some conditions that we attempted required even more time and computing resources (memory) than we were allocated, and therefore we had to exclude those conditions from the study. The largest computational burden occurred when *J*=20 and *M*=7—in fact, we had to remove all *J*=20 conditions from the study. While this appears as a limitation, it has prompted the creation of an R package to streamline the parallelization of chains in Jags which will be used in future HRM studies[[2]](#footnote-2). In testing runs for a new L-HRM study we found that a single L-HRM fitting with 4 time points, 400 examinees, 10 , and 10 raters, takes over 48 hours to complete using one computing node on a supercomputer. Using one computing node on the supercomputer with the R package that parallelizes chains over cores of the computing node, this same L-HRM fitting took only 18 hours. Access to supercomputers is rare, however, so computing remains a limitation. These computing issues do not pose a serious problem in the event that a researcher wants to fit the L-HRM to a specific dataset, but it will in the case of a simulation study requiring several replications of multiple student conditions. Future development of the HRM framework will include maximum likelihood estimation methods for the HRM which will offer results at a much greater speed than Bayesian methods.

This study did not investigate model performance under different amounts of autocorrelation and growth. This study also only examined one time series model specification. Research currently underway is investigating parameter recovery under AR(1) for various combinations of autocorrelation and growth (Casabianca & Junker, n.d.) and well as a general formulation that permits the fitting of the class of time series models (Bond, Casabianca, & Junker, n.d.).

Most importantly, note that here we are assuming that rater behavior does not change over time and we keep rater parameters fixed. In most studies of ratings taken over time, rater drift can be a serious issue (see for example: Casabianca, Lockwood, & McCaffrey, 2014; Casabianca, et al, 2013). However, the rating collection design must allow for the identification of the two separate rater and trait trends. This was possible in a study of teaching where instructional trends and rater trends were decoupled by collecting observations from classrooms on video and then having them scored out of sequence at a later date (Casabianca, Lockwood, & McCaffrey, 2014; Casabianca, et al, 2013). Future treatments of the L-HRM will accommodate changes in traits and rater behavior in datasets that permit this type of analysis.

References

Bill and Melinda Gates Foundation (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Seattle, WA: Kane, T. J., & Staiger, D. O. Retrieved from <http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf>

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

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2014). Trends in Classroom Observation Scores. *Educational and Psychological Measurement*, 0013164414539163.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*(5), 757-783.

Gelman, A., & Rubin, D. B. (1996). Markov chain Monte Carlo methods in biostatistics. *Statistical Methods in Medical Research*, *5*(4), 339-355.

Glas, C. A. W. & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*, 907-922.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton: Princeton University Press.

Hershberger, S. L., Molenaar, P. C., & Corneal, S. E. (1996). A hierarchy of univariate and multivariate structural time series models. *Advanced structural equation modeling: Issues and techniques*, 159-194.

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.

Holman, R., & Glas, C. W. (2005). Modelling non-ignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1-17.

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>

Johnson, M. S., Sinharay, S., & Bradlow, E. T. (2007). Hierarchical item response theory models. *Handbook of statistics: Psychometrics*, *26*, 587-606.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, *82*(1), 35-45.

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

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data. New York: Wiley.

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.

Mislevy, R. J. & Wu, P. K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. (Research Report 96-30). Princeton, NJ: Educational Testing Service

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–177.

Patz, R. J. (1996). Markov Chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh PA.

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.

Plummer, M. (2003, March). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March* (pp. 20-22).

Su, Y. S., & Yajima, M. (2012). R2jags: A Package for Running jags from R. *R package version 0.03-08, URL http://CRAN.R-project.org/package=R2jags*.

Verhelst, N. D., & Verstralen, H. H. (2001). An IRT model for multiple raters. In *Essays on item response theory* (pp. 89-108). Springer New York.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283-306.

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

Table 1

*Prior distributions used in fitting L-HRM. The parameters displayed are the mean and precision (1/variance) of the prior distribution.*



Table 2

*Median absolute bias for L-HRM parameters from simulation study. The top table provides results for the linear trend conditions and the bottom table for the logistic trend conditions.*





Table 3

*Median absolute bias for L-HRM item and rater parameters from simulation study. The top table provides results for the linear trend conditions and the bottom table for the logistic trend conditions.*







*Figure 1.* Direct acyclic graph of the HRM.



*Figure 2.* Direct acyclic graph of the L-HRM parameterization.





*Figure 3.* Bias plots for the standard deviation of θ (true value, ω=1). The *top* plot show bias values for each replication of the linear trend conditions and the *bottom* plot show the same for the logistic trend.

 



*Figure 4.* Bias plots for L-HRM rater bias (severity/leniency) parameter. The *top* plots show bias values for each replication of the linear trend conditions over the true values of rater bias φ (left) and rater standard deviation τ (right). The *bottom* plots show the same for the logistic trend.



**

*Figure 5.* Bias plots for L-HRM rater variability (SD) parameter. The *top* plots show bias values for each replication of the linear trend conditions over the true values of rater bias φ (left) and rater standard deviation τ (right). The *bottom* plots show the same for the logistic trend.



*Figure 6.* Distributions of posterior standard deviations for the latent trait, by time point for the linear, M=7, R=3, N=250 condition.

**Appendix: R2Jags Syntax to Estimate the L-HRM**

library(R2jags)

data <- test$data

N <- length(unique(data$student))

R <- length(unique(data$rater))

J <- length(unique(data$item))

M <- length(unique(data$time))

Km1 <- dim(test$g)[2]

K <- Km1 + 1 # could go for max(data$y)...

NNN <- dim(data)[1] # N\*R\*J\*M, extent of y

bugs.hrm.manytimes.lin.AR1 <- function()

{

 # likelihood - sdm part

 for (i in 1:NNN) { # NNN = N\*J\*R\*M

 y[i] ~ dcat(p[i,])

 for (k in 1:K) {

 d[i,k] <- k - xi[student[i],item[i],time[i]] - phi[rater[i]]

 z[i,k] <- exp(-d[i,k]\*d[i,k]\*tau2[rater[i]]/2)

 }

 for (k in 1:K) {

 p[i,k] <- z[i,k]/sum(z[i,])

 }

 }

 # likelihood - pcm part

 for (m in 1:M) {

 for (n in 1:N) {

 for(j in 1:J) {

 xi[n,j,m] ~ dcat(pcm[n,j,m,])

 psi[n,j,m,1] <- 0

 for (k in 1:(K-1)) {

 psi[n,j,m,k+1] <- a[j]\*(th[n,m] - (b[j] + g[j,k]))

 }

 for (k in 1:K) {

 term[n,j,m,k] <- exp(sum(psi[n,j,m,1:k]))

 }

 for (k in 1:K) {

 pcm[n,j,m,k] <- term[n,j,m,k]/sum(term[n,j,m,])

 }

 }

 }

 }

 # priors - sdm part

 for (r in 1:R) {

 phi[r] ~ dnorm(PHI.MEAN,1/(PHI.SD\*PHI.SD)) # dnorm(0,0.0001)

 tau2[r] ~ dgamma(TAU2.ALPHA, TAU2.BETA) # dgamma(1,1)

 tau[r] <- sqrt(tau2[r])

 }

 # priors - pcm part

 for (j in 1:J) {

 a[j] ~ dgamma(A.ALPHA,A.BETA)

 b[j] ~ dnorm(B.MEAN,1/(B.SD\*B.SD))

 for (k in 1:(K-1)) {

 g[j,k] ~ dnorm(G.MEAN,1/(G.SD\*G.SD))

 }

 }

 for (m in 1:M) {

 # Linear Growth

 trend[m] <- growth.raw\*(m-1)/(M-1)

 }

 for (n in 1:N) {

 Z[n,1] ~ dnorm(0,th.prec)

 th[n,1] <- trend[1] + Z[n,1]

 for (m in 2:M) {

 Z[n,m] ~ dnorm(rho\*Z[n,m-1],th.prec/(1-pow(rho,2)))

 # this is the time series part!

 th[n,m] <- trend[m] + Z[n,m]

 }

 }

 th.prec ~ dgamma(TH.PREC.ALPHA,TH.PREC.BETA)

 rho ~ dunif(-1,1)

 growth.raw ~ dnorm(0,0.01)

 # transformations -- for theta sd and for pcm item params...

 growth <- growth.raw/sd.theta

 sd.theta <- 1/sqrt(th.prec)

 for (j in 1:J) {

 beta[j] <- mean(bstar[j,])

 for (k in 1:(K-1)) {

 bstar[j,k] <- b[j] + g[j,k]

 gamma[j,k] <- bstar[j,k] - beta[j]

 }

 }

}

bugs.hrm.manytimes.lin.AR1.inits <- function()

 list(

 phi = rnorm(R,0,1),

 tau2 = rgamma(R,1,1),

 th.prec = rgamma(1,1,1),

 rho = runif(1,0,1),

 growth.raw = rnorm(1,0,1),

 a = runif(J,.5,1.5),

 b = rnorm(J,0,1),

 g = matrix(rnorm(J\*(K-1),0,0.25),nrow=J,ncol=K-1)

 )

bugs.hrm.manytimes.lin.AR1.data <- c(list(data,NNN=NNN,N=N,

 J=J,K=K,R=R,M=M),

 list(

 PHI.MEAN = 0,

 PHI.SD = 100,

 TAU2.ALPHA = 1,

 TAU2.BETA = 1,

 A.ALPHA=1,

 A.BETA =1,

 B.MEAN = 0,

 B.SD = 1,

 G.MEAN = 0,

 G.SD = 1,

 TH.PREC.ALPHA = 1,

 TH.PREC.BETA = 1

 )

 )

hrm.manytimes.lin.AR1 <- jags(

 data = bugs.hrm.manytimes.lin.AR1.data,

 inits= bugs.hrm.manytimes.lin.AR1.inits,

 parameters.to.save=c("a","beta","gamma","sd.theta","th","rho",

"growth","phi","tau"),

model.file = bugs.hrm.manytimes.lin.AR1, n.iter=8000,n.burn=2000,n.thin=6,n.chains=3

)

1. For more information on Stampede, visit: <https://www.tacc.utexas.edu/stampede/> [↑](#footnote-ref-1)
2. The R package was obtained through personal correspondence with Ruizhu Huang and Lei Huang at Texas Advanced Computing Center (TACC). [↑](#footnote-ref-2)