A Framework for the Analysis of Ratings:

The Expanded Hierarchical Rater Model

1 Overview

The proposed research will extend the hierarchical rater model, a statistical model for multiple ratings of responses, behavior and performance, (HRM; Casabianca, Junker, & Patz, 2013; Patz, 1996; Patz, Junker, Johnson, & Mariano, 2002), to a modeling framework for hierarchical and longitudinal designs with multidimensional assessments.

The HRM was introduced within the educational measurement literature for the analysis of student work on various rating scales, as well as analysis of rater behavior. However, the potential applications of the HRM are unlimited and the HRM is relevant in any field of research or practice that relies on ratings, for example, consumer product, psychological, and medical research.

Recently, another educational application appears most appropriate for analysis with the HRM; the surge of educational policy interest in teacher accountability has generated attention to the rating of teachers in classrooms. Often in these scenarios, we must account for the nesting of classrooms within teachers, teachers within schools, and schools within districts. Further, teachers are typically observed on multiple occasions during an academic year. Neither the current HRM nor other approaches such as generalizability theory are capable of handing these study designs, and we propose to perform the necessary theoretical advancements to make the HRM flexible for such measurement situations.

just don't need to pick a fight so early in the proposal

Ratings of rich response formats have been a part of the assessment landscape for as long as there have been assessments. Short answer and multiple choice question formats largely eliminate extraneous variability in scoring, but in many areas of assessment, rater bias, variability and other factors affect assessment scores. Unusual rating behavior (DeCarlo, 2008; Patz, Junker, Johnson, & Mariano, 2002; Wolfe & McVay, 2002), factors in raters' backgrounds (Winke, Gass, & Myford, 2011), the circumstances of rating (Mariano & Junker, 2007), and their effects on procedures for producing and reporting assessment scores (e.g. Yen, Ochieng, Michaels, & Friedman, 2005) continue to be of central interest.

Many studies of rater effects—including several of those cited above—employ an item response theory (IRT) model like the generalized partial credit model (Muraki, 1992; Patz & Junker, 1999b) or the Rasch Facets model (Linacre, 1989). As noted by Patz et al. (2002) and proven formally by Mariano (2002), however, these approaches have a fundamental flaw: as the number of raters increase—even for a single item!—the standard error of measurement for the examinee tends to zero. This cannot be: repeatedly rating the same item response or task behavior can tell us more about the quality of that particular response, but should not reduce measurement error for the underlying latent trait variable to zero.

ref: van der linden and hambleton

The HRM is an extension of polytomous IRT models for items with multiple ratings or scores, that corrects this flaw in IRT and Facets models, by composing two measurement stages: the first stage is a "signal-detection-like" model for measuring the ideal rating of an item based on multiple raters' observed ratings; and the second stage is an IRT model relating the ideal ratings to the underlying examinee proficiency or trait variable. Other

approaches to correcting this flaw have also been proposed (Bock, Brennan, & Muraki, 2002; Muckle & Karabatsos, 2009; Wilson & Hoskens, 2001).

In comparison to generalizability theory (Brennan, 2001), the HRM provides richer information on the rating process, including estimates of bias and reliability for individual raters, while also considering differences in the rubric indicators or items and overall level and variance of the responses or performance under study. Specifically, from the HRM we obtain estimates of item parameters, examinee trait level means and variance, and rater bias and variance. Since its introduction, the HRM has been extended to accommodate (i) the fitting of rater covariates permitting the use of the HRM as an explanatory model (Mariano & Junker, 2007), and (ii) the fitting of rater effects beyond simple rater bias (severity) and rater variability (consistency) (DeCarlo, Kim, & Johnson, 2011). Large-impact studies such as the Measures of Effective Teaching (MET; Bill and Melinda Gates Foundation, 2012) use generalizability theory to decompose the variance in scores from scoring rubrics used in the classroom. With the proposed HRM framework, we hypothesize the HRM will far outperform the generalizability theory approach.

We will build on the basic HRM framework by: expanding and extending it for known problems in the social sciences; developing practical data analysis and computing methodology so that other researchers can use the HRM; and illustrating our work through real and simulated data analyses relevant to education research and policy, particularly with regard to multidimensional ratings over a duration of time in complex measurement scenarios. In particular, we seek to:

- Develop approaches to incorporate a time-varying component for longitudinal ratings and create a mechanism to evaluate pre- and post-test differences;
- Use multidimensional IRT models to properly capture the factor structure of the traits and items being rated;
- Investigate how the HRM can be used to model complex hierarchical designs and data structures;
- Generate a maximum likelihood approach for estimation of the full HRM framework and assess the impact of sample size on the quality of parameter estimation;
- Combine the basic HRM framework and these extensions and expansions into a single HRM framework with unified notation and formulations;
- Develop and code model fitting algorithms, and assess our work using simulation studies, as well as analysis of real education data. Data will include teacher ratings from the MET project (BMGF, 2012), available through the *Inter-university Consortium for Political and Social Research*.[1]
- Write open source code so that our algorithms and models will be accessible to the research community.

[1]MET project data will be made available through restricted use agreements in fall 2013.

We next provide some background on the HRM and discuss the analysis of ratings in general and in the specific context of classrooms. In Section 3 we outline our approach to developing the longitudinal HRM, and in Section 4 we provide preliminary results from a simulation showing that our approach is feasible for the longitudinal analysis of ratings. In Sections 5 and 6 we outline our approach to extending the HRM for multidimensional assessments as well as complex hierarchical study designs. In Section 7 we discuss another goal, which is to develop an alternate estimation approach to make the HRM framework accessible to researchers. Section 8 details our research questions, plans, and broader impacts of our proposed activities.

2 The Hierarchical Rater Model

2.1 Model Formulation and Notation

The basic HRM is composed of a three-level hierarchy. We present the hierarchy in the context of classroom observation where the indicators on classroom observation scoring protocols or rubrics are considered the "item" and an "item response" is simply the sampling of teachers' instruction that relates to a particular indicator.

The first level of the hierarchy models the distribution of ratings given the quality of response (i.e. teaching/instruction), the second level models the distribution of a teacher's response (or a teacher's teaching) given their latent trait, and the third level models the distribution of the latent trait θ_p. The hierarchy is given by

$$\left.\begin{array}{lcl} X_{pir} & \sim & \text{a polytomous signal detection model, } r = 1, \ldots, R, \text{ for each } p, i. \\ \xi_{pi} & \sim & \text{a polytomous IRT model, } i = 1, \ldots, I, \text{ for each } p \\ \theta_p & \sim & \text{i.i.d. } N\left(\mu, \sigma^2\right), p = 1, \ldots, P \end{array}\right\} \quad (1)$$

Here, θ_p, the latent trait for teacher p $(p = 1, \ldots, P)$ is normally distributed with mean μ and σ^2, ξ_{pi} is the ideal rating for teacher p on indicator i $(i = 1, \ldots, I)$ and X_{pir} is the observed rating given by rater r for teacher p's response to indicator i. Note that specifying a normal distribution for the latent trait is a popular choice, but alternatives could be used instead.

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, the quality of teaching. In other words, the teacher's teaching is videotaped/observed so that a set of I indicators may be judged (with ideal ratings) and then a series of R raters evaluate the video/observation, giving ratings conditional on the teachers' instruction. This notation is for the completely crossed design where all raters score each indicator. Within this framework incomplete designs are treated as missing completely at random (MCAR; Mislevy & Wu, 1996; Rubin & Little, 2002). Models for informative missingness (e.g. Glas & Pimentel, 2008; Holman & Glas, 2005) could also be incorporated directly into the HRM if needed.

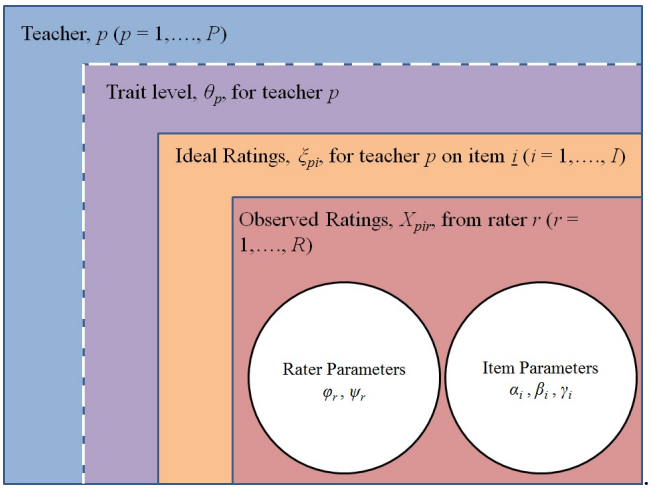

Figure 1: Hierarchy in the basic HRM

Figure 1 pictorially shows the HRM as it was introduced originally; here we have p teachers (or more generally, examinees), and within each is a latent trait, ideal ratings per-item, and observed ratings per-item from each rater. The white dashed line in this figure represents the distinction between the structure of the data, here, just teachers with no nesting, and the estimated and observed parameters, θ_p, ξ_{pi}, and X_{pir}. Two additional sets of parameters are estimated for raters, (ϕ_r, ψ_r), and items/indicators, (α_i, β_i, γ_i).

In the second level, the ideal ratings ξ_{pi} represent the quality of teacher p's response to indicator i, and are latent variables modeled using a polytomous IRT model, such as the A-category generalized partial credit model (GPCM; Muraki, 1992). From the GPCM component of the HRM we estimate α_i, the indicator discrimination, β_i, the indicator location, and γ_{ia}, the a^{th} threshold parameter for indicator i, 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 A, the number of response categories per indicator, need not be constant across items. With ideal rating ξ_{pi} and A possible scores ($a = 1, \ldots, A$), the GPCM is given by:

$$P\left[\xi_{pi} = \xi | \theta_p, \alpha_i, \beta_i, \gamma_{i\xi}\right] = \frac{\exp\left\{\sum_{a=1}^{\xi} \alpha_i \left(\theta_p - \beta_i - \gamma_{ia}\right)\right\}}{\sum_{h=0}^{A-1} \exp\left\{\sum_{a=1}^{h} \alpha_i \left(\theta_p - \beta_i - \gamma_{ia}\right)\right\}}. \tag{2}$$

Note, the ideal rating is the rating that teacher p would receive on indicator i, by a rater exhibiting *no* rater bias and perfect rating consistency. In the HRM the deviations between actually observed ratings X_{pir} and these ideal ratings ξ_{pi} 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 $p_{\xi ar}$ = (Rater r rates a|ideal rating ξ). A simple signal detection model uses a discrete unimodal distribution for each row of the matrix to give the probability of observed rating X_{pi} given ideal rating ξ_{pi}. The mode of this distribution is the rater bias or

severity, ϕ_r, and the spread of this distribution is the rater variability or unreliability, ψ_r. The signal detection model can be specified such that probabilities in each row of the matrix are proportional to a Normal density with mean $\xi + \phi_r$ and standard deviation ψ_r:

$$p_{\xi ar} = P\left[X_{pir} = a | \xi_{pi} = \xi\right] \propto \exp\left\{-\frac{1}{2\psi_r^2}[a - (\xi + \phi_r)]^2\right\}. \quad (3)$$

The severity parameter ϕ_r indicates a rater's deviation from the ideal rating; values near 0 indicate no deviation, negative values indicate severity (negative bias), and positive values indicate leniency (positive bias). The spread parameter ψ_r 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.

2.1.1 Covariates

Characteristics of the observation or rating process, examinees, raters, and/or of raters' ratings, have the potential to influence rater bias and variability. We incorporate covariates into the HRM within the signal detection model component to represent any of these characteristics (see Mariano & Junker, 2007, for more details). Covariates may be fixed for all ratings from the same rater (e.g. gender, race, hours of rater training) or they may differ over ratings from the same rater (e.g. time to complete rating or scoring mode). Non-zero rater covariate effects reveal areas in which attention is needed; analysis of bias and variability effects of these covariates may be useful for an audit of the rating process to adapt features of the rating design while the study is in progress (e.g. rater training/calibration). It could also be useful to compare the effects of changing scoring procedures.

2.1.2 Estimation

The HRM begins with the hierarchy in Equation 1, an IRT model such as the GPCM in Equation 2 for ideal ratings, and the signal detection model in Equation 3 connecting observed ratings to ideal ratings. Estimating the Bayesian HRM with MCMC is a straightforward extension of the MCMC approach to estimating a GPCM (Patz & Junker, 1999a,b); the extension must include the additional parameters for rater bias and variability and the ideal ratings (see Patz et al., 2002, for an in-depth discussion of MCMC for HRM). Additional modifications for rater covariates are discussed by Mariano and Junker (2007). To use the Bayesian framework, we must also specify priors for seven sets of parameters: GPCM item parameters α_i, β_i, and γ_{ia}, for $i = 1, \ldots, I$ and $a = 0, \ldots, A-1$; rater parameters ϕ_r , ψ_r, $r = 1, \ldots, R$; and examinee proficiency distribution parameters μ, σ^2. Typically we use some noninformative parameters for the HRM parameters to reflect little prior knowledge. The prior distributions for the GPCM specified in Equation 2 should account for location indeterminacy by constraining either the latent proficiency mean μ, or the item difficulty parameters β_i. A similar scale indeterminacy problem can be addressed by constraining either the item discrimination parameters α_i or the latent proficiency variance σ^2. These constraints may be hard linear constraints, or soft constraints imposed through prior distributions.

The HRM has also been fitted with marginal maximum likelihood (MML; Hombo & Donoghue, 2001). DeCarlo et al. (2011) fitted the a variation of the HRM using posterior

modal estimation (PME) implemented with an Expectation-Maximization algorithm (EM, Dempster, Laird, & Rubin, 1977; Wu, 1983). Although the HRM has been estimated with maximum likelihood methods (Hombo & Donoghue, 2001), the literature has predominantly treated HRM as the Bayesian model (Patz et al., 2002) described here. Faster, more scalable ML methods for the HRM have not yet been developed.

2.2 The HRM in a Specific Context: Ratings from Classroom Observation

While the HRM can be applied in any field, much of our work is motivated by applying the HRM in the analysis of teacher ratings. States and districts are increasingly introducing classroom observation scoring protocols as measures of teaching quality into their teacher accountability systems. Ratings from multiple lessons are aggregated to derive measures of an individual teacher's teaching. There are specific issues related to using ratings from classroom observation including the complex hierarchy of studying multiple instances of teaching within classrooms within schools, within districts, where there are multiple raters with differing levels of education, training and experience. There are very high stakes attached to the proper measurement of teaching practice.

A major factor that may impact teacher ratings is the trend in rating over time; in other words, how raters change in their use of the score scale (Casabianca, Lockwood, & McCaffrey, 2013; Casabianca, et al., 2012; Leckie & Baird, 2011; Myford & Wolfe, 2009). Other influences varying from lesson-to-lesson include idiosyncrasies in the observation and/or rating process that may affect the dynamics of the classroom observation or the rater. For example, lessons for which the video or audio equipment was poor quality. Day-to-day variation in classroom activities, and variation in curricula and contexts across different classrooms (e.g. middle school teachers who teach multiple sections or courses each year), mean that a small sample of lessons could vary widely on scores even when scored by a common rater.

Historical and recent research has used generalizability theory models to analyze different facets of teacher ratings (see, for example, BMGF, 2012; Hill, Charalambous, & Kraft, 2012); that approach falls short in accommodating some of the complexities in classroom observation. That is, in addition to not providing estimates of latent traits, item parameters, or rater parameters, frequently, univariate models are run for rubric indicators, thereby ignoring multidimensional structure of the rubric. Traditional G study models also fail to incorporate a time-varying component for lessons, thereby ignoring information about time trends in scores. (Casabianca, Lockwood, & McCaffrey, 2013, are currently using more complicated "G study" models to investigate time trends, but this approach will not be pursued as part of the propsed work here.) Additionally, generalizability theory analysis provides an analysis of the sources of error variance, but covariates are not incorporated into the model to explain the variance decomposition and score reliabilities.

3 Developing the HRM for Multiple Timepoints: the Longitudinal HRM

To explicitly model changes in the latent trait or changes in rating behavior (rater bias and reliability) we add a level for time to the hierarchy.

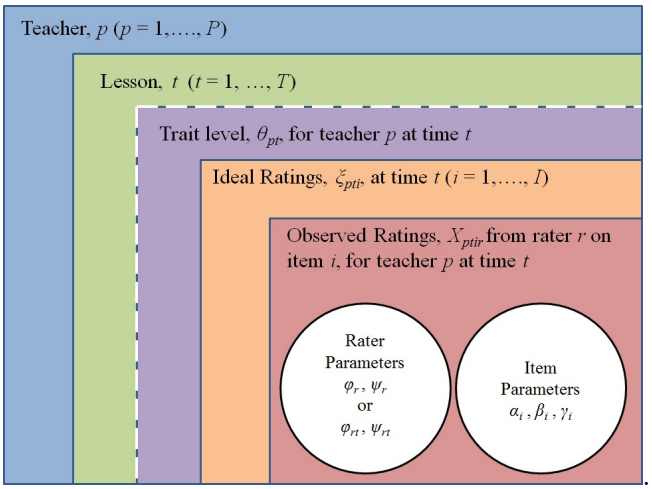

Figure 2: The longitudinal HRM: the addition of multiple lessons per teacher suggests another layer to the hierarchy that could result in multiple parameters for the trait, as well as multiple parameters to examine how raters' ratings change over time (in terms of bias and reliability).

Figure 2 is an augmented version of Figure 1 for the basic HRM; here we have a layer of lessons observed within each teacher. In this figure, the trait level is indexed with t for time, and so are the rater parameters, ϕ_{rt}, ψ_{rt}. This is so because the additional level could be specified in such a way that: (i) multiple traits are estimated (one at each timepoint) and the ideal ratings differ because it is assumed that the trait level (or teaching quality) is improving (or declining); (ii) rater parameters are estimated at each timepoint in order to study the changes in bias and reliability (e.g. due to rater learning and experience) over the study duration; or (iii) we can estimate a trait level for each timepoint and a set of rater parameters for each rater at each timepoint. We acknowledge that both the trait level and rater behavior can change over time, and there is a conflation between these trends, therefore scenario (iii) is the most challenging because of the need to parse the concurrent trends.

The additional level for time could be specified using a variety of models for time trends; we will investigate time series models (Hamilton, 1994; Hershberger, Molenaar, & Corneal, 1996) and growth curve models (Raudenbush & Bryk, 2002).

3.1 Time Series Model Approaches

In an initial time series model approach, we treat time ordinally with a simple autoregressive time series model of order 1. The trait for subject p at time point t (where $t = 1, \ldots, T$) is

$$\theta_{pt} = \eta_0 + \eta_1 \theta_{p(t-1)} + \varepsilon_{pt}. \tag{4}$$

Here, η_0 is the baseline increase or decrease in θ_{pt}, η_1 is the autocorrelation between θ_{pt} and $\theta_{p(t-1)}$, and ε_{pt} represents the noise or variation for subject p at timepoint t which follow a normal density, $\varepsilon_{pt} \sim N(0, \sigma^2)$, with constant variance. We estimate the θ_{pt} as well as the η_0 and η_1 parameters.

Note that under this formulation, the longitudinal HRM has T timepoints with ideal and observed ratings at time t nested within each θ_{pt}. A similar model will be used to evaluate change in the rater parameters. Additionally, different time series models will also be explored including autoregressive models of different orders and moving average models.

A preliminary feasibility study (see Section 4) demonstrates that we can successfully estimate trends with this approach (Casabianca & Junker, 2013).

3.2 Growth Curve Model Approaches

Growth curve models use a hierarchical linear model for repeated observations. Unlike time series models, however, growth curve models use an interval scale so that the actual duration in between observations could be incorporated in the model for estimating time trends. A simple model for quadratic growth is

$$\theta_{pt} = \gamma_{0p} + \gamma_{1p}t + \gamma_{2p}t^2 + \varepsilon_{pt}, \tag{5}$$

where, as before, θ_{pt} is the trait for person p at time t, γ_{0p} is the initial status at time $t = 0$, γ_{1p} is the growth trajectory, t is a temporal dimension that here is assumed to be the same for all individuals, γ_{2p} is the curvilinearity of the growth trajectory, and ε_{pt} is the disturbance term which is i.i.d.

We provide a model for quadratic growth as recent research shows quadratric, cubic, and even quartic polynomial trends in teacher ratings (Casabianca et al, 2012; Casabianca, Lockwood, & McCaffrey, 2013). However, we will explore a number of different growth curve modeling approaches for changes in the trait (as shown above), changes in rater parameters, and changes in both types of parameters.

4 Proof of Concept: Pilot Study for the Longitudinal HRM

We are currently performing a pilot simulation study to evaluate and understand some preliminary longitudinal HRMs for changes in traits (Casabianca & Junker, 2013). To determine how the model performs under various conditions we vary the following: (i) Sample size, $P = 80, 250, 500, 1500$; (ii) Number of time points, $T = 3, 5, 7$; (iii) Number of items/indicators, $I = 5, 13$; and (iv) Number of raters, $R = 2, 4, 6$.

We generate data by drawing the baseline trait level θ_{p0} from a $N(0, 4)$ distribution and using the time series model with $\eta_0 = \eta_1 = 1$ to compute the additional θ_{pt}. In other words, we modeled a 1-unit increase in trait location at each t so that the average true trait values were roughly $\theta_{p0} = 0$, $\theta_{p1} = 1$, and $\theta_{p2} = 2$. Then, the ideal ratings are generated using the θ_{pt} and GPCM item parameters derived from classical test theory-based item indices from real teacher ratings on a widely-used classroom observation scoring rubric under study in Casabianca, et al (2012). Lastly, the observed ratings are generated using the generated ideal ratings and true rater parameters reflecting raters with high and low levels of rater bias and reliability. The design is fully crossed so that each combination of factors is evaluated.

Thus far we have tested the new model formulation under one condition; we fit the longitudinal HRM as specified in Equation 4 with three timepoints $T = 3$, and $P = 500$, $R = 3$, $I = 5$, using MCMC estimation in Winbugs with 2,000 iterations and burn-in of 1,000.

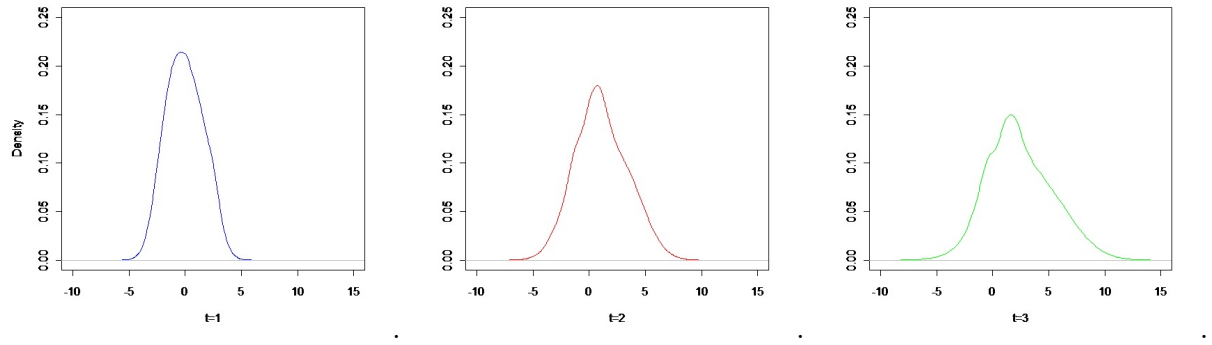

Figure 3: Estimated trait distributions at each time t. Mean($t = 1$)=-0.04, SD($t = 1$)=1.69; Mean($t = 2$)=1.10, SD($t = 2$)=2.43; Mean($t = 3$)=2.44, SD($t = 3$)=2.98.

The MCMC estimation converged for all parameters and rater and item parameters were recovered fully and stably. The purpose of this test was to determine if the HRM specified with the autoregressive time series model would recover latent traits following the trend used to generate the simulation data.

Figure 3 gives posterior densities for θ_{pt} at each time point such that: $\theta_{p1} = \eta_0 + \eta_1\theta_{p0} + \varepsilon_{p1}$; $\theta_{p2} = \eta_0 + \eta_1\theta_{p1} + \varepsilon_{p2}$; and, $\theta_{p3} = \eta_0 + \eta_1\theta_{p2} + \varepsilon_{p3}$. We observe a shift of over one unit on the θ_{pt} scale between $t = 1$ to 2 and $t = 2$ to 3. The variability of the estimates increased with each timepoint and there is positive skew at $t = 2$ and $t = 3$. The expected growth trends were indeed captured by this parameterization of the longitudinal HRM. Since the longitudinal HRM yielded estimated traits that increased with time we can conclude that the HRM model formulation as we introduced it here is feasible and warrants further study. Full results of this pilot research will be presented at a national conference in April 2013 (Casabianca & Junker, 2013) and will inform the proposed extension of the HRM for multiple timepoints.

5 Extending the HRM to Multidimensional Measures

Complex IRT models that incorporate the multidimensional structure of assessments are becoming more accessible and popular in practice. Currently, the HRM utilizes a unidimensional IRT model for polytomous responses. To analyze assessments with groups of items related to multiple factors or traits, we must estimate multiple HRMs; this approach is not desirable. We will extend the current formulation of the HRM to include multidimensional IRT (MIRT; Reckase, 2009) models for polytomous items. This extension will increase the generality of the HRM so that the framework can handle assessments and rubrics with complex factor structure.

An example of a strightforward MIRT model is the noncompensatory multidimensional IRT model is given by:

$$P(X_{pi} = 1|\theta_p; \alpha_i, \beta_i) = \prod_{d=1}^{D} \frac{\exp(\alpha_{di}\theta_{dp} + \beta_{di})]}{1 + \exp(\alpha_{di}\theta_{dp} + \beta_{di})]}, \tag{6}$$

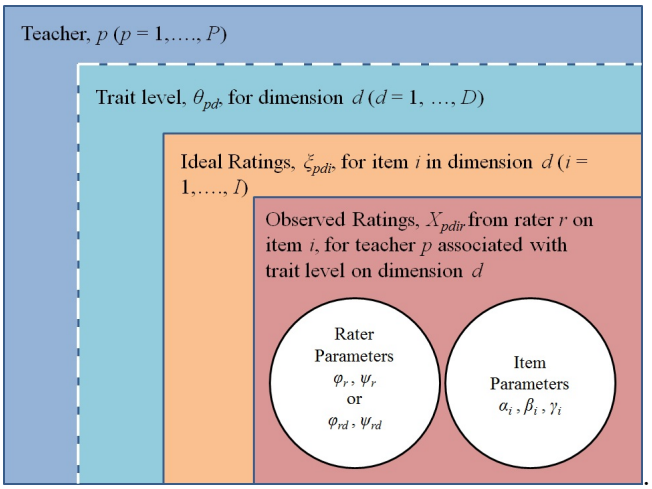

Figure 4: The HRM for multidimensional assessments. In this scenario, there are multiple traits; each trait is estimated based on a group of items. It is also possible to allow rater parameters to vary by trait dimension if it is believed that rater behavior (bias and reliability) change by the type of trait being measured.

where p, i, and d subscript represents examinees, indicators, and traits/dimensions, respectively, θ_p represents a vector of latent trait variables for examinee p, α_i represents multiple discrimination parameters associated with indicator i, and β_i represents an indicator's location on an item response surface.

Two examples of multidimensional assessments are the Advanced Placement (AP) foreign language exams (College Board, n.d.) and the Classroom Assessment Scoring System-Secondary (CLASS-S; Pianta, Hamre, Haynes, Mintz, & Paro, 2007) for teacher/classroom evaluation. The AP language exams have listening, reading, oral, and speaking sections; portions of these sections are based on observed performance or essays and consequently rated by trained raters. To incorporate these four distinct constructs for language ability into the analysis, we would estimate four HRMs. The multidimensional HRM extension will be flexible such that only one model would be needed to estimate the parameters assuming four latent traits.

The CLASS-S is a popular classroom observation scoring protocol, which is frequently used longitudinally; it is very widely-used, even for evaluating teachers in Federal programs such as Head Start (Head Start Program Final Rule, 2011). Lessons are evaluated on 10 dimensions of teaching, and each dimension is related to one of three domains: Emotional Support, Instructional Support, or Classroom Organization. Currently, three HRMs would be estimated (one for each domain score). Of course, an alternative to estimating multiple HRMs is to ignore the multidimensionality completely and use a unidimensional model with all indicators included.

Figure 4 displays the HRM hierarchy with multidimensional assessment structure for an assessment such as the CLASS-S. Here we display trait levels for multiple dimensions denoted by θ_{pd}. Ideal and observed ratings are indexed within these traits to indicate the correspondence between ratings and the trait categories. We will explore how to vary rater

parameters by trait dimension as well by estimating ϕ_{rd} and ψ_{rd}.

Inherently, we know that adding a MIRT model to the HRM is a very challenging task; MIRT and other complex IRT models are plagued with issues surrounding identifiablity (Azevedo, 2009; Haberman, 2005; Maris & Bechger, 2009), parameterization (Martín, González, Tuerlinckx, 2009), and estimation (Partchev, 2009). Using a MIRT model within a hierarchical framework, and with a timing component, could prove troublesome, especially with the potential for parameterizations under which the parameters vary by the trait dimension. However, other, very complex, latent variable model frameworks, for example, the general diagnostic model (von Davier, 2005), also include MIRT models. Part of our expansion will include rigorous testing using simulations to determine model performance under various situations. We will provide guidelines for best practice in terms of how the HRM framework should be used (e.g. How many traits are too many?; What is the required sample size to obtain valid and precise estimates? etc.).

6 Extending the HRM to More Complex Hierarchies

Given the potential for vast application in many fields, the HRM framework needs to accommodate hierarchical data structures that may appear in these fields to maximize the flexibility and therefore accessibility of the model. Thus far in our narrative, we have discussed a sample of teachers with no hierarchy or nesting. Additional hierarchical structure that may be found in educational settings includes classrooms, schools and districts. In addition, often the unit of observation on rubrics for classroom observation are in segments; multiple subsets, or segments, of lessons are observed so that there are multiple ratings for each lesson. Figure 6 shows that addition of classrooms and segment of observation as new levels to the HRM hierarchy.

Another example where the HRM framework would be useful is in a study by Khan et al (2011); their research compared ratings from radiologists and surgeons in their judgments on computer tomography (CT) imaging features, and found low to moderate agreement using a simple Cohen kappa coefficient. In this type of scenario there may be groups of medical personnel in different departments and/or hospitals participating in the study. This type of structure is not accommodated and therefore not accounted for by the current formulation of the HRM.

instead of saying the old HRM can't do it, say that your new framework will do it.

"Research tells us" sounds like you don't think you belong to the club of researchers (you do!!!).

~~Research tells us that~~ Ignoring hierarchical structure in data results in underestimation of the variance of the estimated coefficients. Other problems could manifest as well. Therefore, while we may not at all be interested in estimating parameters for different levels of the hierarchy (e.g. overall traits for a specific school), we still want to incorporate it into analysis. This will entail developing formulae for the HRM that extends the data into a 3-level hierarchy (in addition to the structure already in the HRM). [NOT SURE WHAT ELSE TO PUT HERE]

I don't think you have to say anything more

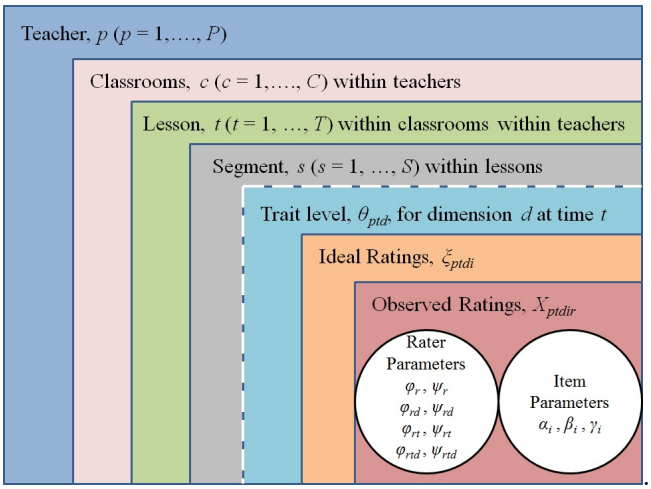

Figure 5: The complete HRM framework with complex data hierarchy, time-varying components, and multidimensional assessment.

7 Making the HRM More Accessible with Maximum Likelihood Estimation Techniques

Estimating the HRM as a Bayesian model with MCMC is not straighforward; knowledge of Bayesian statistics and MCMC is required, and the parameter estimates are sensitive to priors and therefore, are not unique. talk abotu scalability to large data sets instead

While the HRM is traditionally estimated in this way, ML equations have been used in research (Hombo & Donoghue, 2000). There are advantages and disadvantages to each estimation approach, however, we feel that accessibility to researchers is limited because syntax and software for applying ML estimation equations are unavailable. To rectify this, we will write ML equations within an EM (Demspter, Laird, & Rubin, 1977) algorithm for the most complex case where there are multiple timepoints, multidimensional traits, and complex hierarchical structure. The ML equations will be coded to estimate the HRM within an R package, and made publicly available.

As in a traditional IRT model, typically a practitioner will perform multiple estimation routines to obtain estimates of different types of parameters; for example, item parameter calibration will provide estimates of IRT item parameters, and IRT scoring will provide estimates of latent traits. We expect the ML approach for the HRM framework to require the user to run a specific set of routines or algorithms (as series of EM algorithms) to obtain the desired estimates.

Much like the other components in this proposal, developing ML methods for the full HRM framework will be a challenge due to the parameterization and the number of estimated parameters. However, ML methods have been used for other complex frameworks as well (GDM, von Davier, 2005), and therefore, development of these techniques is feasible. One possible route to ensure success is to consider the use of parsimonious nonparametric methods for ML estimation (Casabianca, 2011; Casabianca & Junker, 2013; Casabianca & Lewis, 2012; Casabianca, Xu, Jia, & Lewis, 2010).

8 Research Plan

We propose to develop and disseminate the methodology, application and software for fitting the framework of Hierarchical Rater Models over a three-year period.

8.1 Year 1

In year 1, we plan to conduct research contributing answers to the following three questions.

Question 1.1 *Which longitudinal models should be used to incorporate an additional level to the HRM to properly analyze longitudinal assessments in education research and research in other fields?*

Preliminary results provided in Section 4 suggest that an additional level in the HRM is feasible, at least with the time series model in Equation 4. A full feasibility study will be conducted for the HRM for both time series and growth curve models, using simulation conditions that mimic large-scale testing situations, teacher evaluation, and longitudinal research studies in psychology and medicine. In addition to data simulations for feasibility, we will use the MET project data (BMGF, 2012) to evaluate our longitudinal model formulations with actual teacher ratings. We may also use longitudinal functioning magnetic resonance imaging (fMRI) data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) for real data evaluation.

Question 1.2 *How can we best model simultaneous changes in trait and rater behavior over time?*

We hypothesize that the addition of a time-varying component will be straightforward for examining the change in the trait *or* rater parameters (assuming the other remains constant). However, the longitudinal HRM, as part of the full HRM framework, should be capable of modeling concurrent changes in traits *and* in rater behavior. We will develop a model formulation that allows both to vary over time. We will evaluate the formulation with simulations and using the MET project dataset, which is appropriate as there are observed trends in teaching quality and rater behavior.

Question 1.3 *By what mechanism can we evaluate differences between timepoints using the longitudinal HRM?*

There are several possible research designs where ratings are used to evaluate an intervention involving the examinees (educational program) or the raters (rater calibration and training). We will develop hypothesis tests and effect sizes for evaluating the impact of an intervention on traits and/or rater behavior (in terms of rater bias and reliability). The developed methods for hypothesis tests and effect sizes will also be evaluated with simulations; real intervention data will be acquired to determine feasibility and performance with actual data.

8.2 Year 2

In year 2, we plan to conduct research contributing answers to the following three questions.

Question 2.1 *How can we incorporate MIRT models into the HRM framework?*

Currently, a univariate HRM is fitted for each latent trait in a multidimensional assesment, or all indicators are combined and the multidimensional structure is ignored. We will analytically explore model formulations for a variety of MIRT models to be included in the HRM framework. In doing so we will also explore the impact on the rater parameters and determine parameterizations for estimating fixed rater parameters or a set of rater parameters for each trait. Simulations and real data applications using the MET project data, which uses five different multidimensional classroom observation tools, will be used to evaluate our formulations.

Question 2.2 *What is the impact of adding levels to the hierarchy to represent complex structure in the data?*

As it stands, the 'hierarchy" in the HRM refers to the multiple observed ratings associated with a single, latent, ideal rating for an item. This hierarchy ignores any hierarchical structure or nesting in the data design. We will design an expanded HRM up to three three levels representing data structure, for example, schools, teachers, and classrooms. With simulations we will analyze the differences in estimated parameters with and without accounting for hierarchical data structure.

Question 2.3 *How can we fully integrate the basic HRM, with covariates, time-varying components, multidimensional assessments and up to three levels of hierarchical stucture, into a unifed model formulation for the full HRM framework?*

Creating the full HRM framework will be possible after we address the aforementioned research questions. Our goal is to derive a model equation that is generalized enough such that most measurement scenarios for ratings data can be captured and fitted as a special case of the full framework. Creating the full model framework will involve extensive mathematical and analytical work.

The full HRM framework will accommodate a measurement scenario that, for example, has three timepoints, a four dimensional rating scale, and a three-level hierarchy. We will evaluate the performance of the model using simulations and determine the constraints in terms of estimation as a Bayesian model in regard to the number of items, dimensions/traits, examinees, raters, ratings, etc. We will demonstrate the applicability of the model with real data.

In addition we will begin considering ML estimation in anticipation of Year 3 work.

8.3 Year 3

In year 3, we plan to conduct the following two research activities which are related to providing an alternative to estimating the HRM as a Bayesian model.

Activity 3.1 *Developing and coding ML estimation algorithms, and assessing our work using simulation studies as well as analyses of real ratings data.* There are extant applications of maximum likelihood for estimation of the HRM (DeCarlo, Johnson, & Kim, 2011; Hombo & Donoghue, 2000), however, there is no real documentation of the ML equations and they are unavailable in software. We will have started to consider ML algorithms in Year 2; in Year 3, we will develop ML algorithms for the unified, full, HRM framework. We anticipate there to be some numerical problems to work out, and possibly some trouble with identifiability (Haberman, 2005). However, with this in mind, we will explore different routes of ML estimation including more parsimonious estimation models (Casabianca, 2011; Casabianca & Junker, 2013; Casabianca & Lewis, 2012; Casabianca, Xu, Jia, & Lewis, 2010). Analysis of real and simulated data, to understand the operating characteristics of our models and fitting algorithms, will also be ongoing throughout the project.

Activity 3.2 *Writing open source code so that our algorithms and models will be accessible to the research community* To made the full HRM framework available to researchers, we will develop and make available an R library, comparable in functionality to an IRT library such as ltm, that will make it possible for practitioners to fit and make inferences from ratings data, using some of our algorithms and models. We expect that this library will consist of a mixture of R functions that serve as a direct interface for users who wish to use our methods. We also expect that this library will be revised and extended during and beyond the course of this project, if it is funded, eventually becoming similar to, or perhaps a component of, a package such as statnet (Handcock et al., 2008).

8.4 Data Application: Measures of Effective Teaching (MET) Project

The Bill and Melinda Gates Foundation's *Measures of Effective Teaching* (MET) project is one of the largest and most extensive studies of classroom teaching ever undertaken in the United States. MET researchers collected a variety of indicators of teaching quality over a two-year period (2009-2011) in the classrooms of more than 2,500 fourth- through ninth-grade teachers working in 317 schools located in 6 large school districts in the United States. The data collected on teachers and their teaching include video-recorded lessons taught by a teacher and scored by independent observers using multiple classroom observation protocols. These data will become available in fall 2013. Casabianca has been working with simiar data in recent studies (Casabianca, et al, 2012; Casabianca, Lockwood, & McCaffrey, 2013).

8.5 Broader Impacts of the Proposed Activities

The HRM framework will be developed so that it can be applied in any context where there are raters observing and judging performance. In truth, the potential for the application of the HRM is limitless. In education, applications may include: ratings of student work (Casabianca, Junker, & Patz, 2013; Mariano & Junker, 2007), judgments made in educational standard settings (Kalinski et al, 2012), or evaluations of teachers' teaching (Casabianca et al, 2012; Hill et al, 2012). In other disciplines, some measurement scenarios that would benefit from the HRM framework include: consumer product ratings (Horn & Salvendy, 2006), ratings of psychological traits (Clare, Gudjonsson, Rutter, & Cross, 2011;

Edens, Boccaccini, & Johnson, 2010) and and ratings of medical personnel performance and reasoning (Berger et al, 2012; Yeates, O'Neill, Mann, & Eva, 2012).

Extant approaches used to analyze ratings in the aforementioned contexts typically involve descriptive numerical summaries, traditional estimates of rater reliability, and generalizability studies. The HRM framework would be useful in this scenario, to estimate latent traits, parameters describing rubric indicators or items, and rater parameters to investigate and possibly reveal information that would contribute to subsequent alignment between raters.

8.6 Work Flow and Dissemination

Casabianca will take the lead on directing the project, under consultation from Junker, and will be responsible for organizing and implementing the study.

In all three years of the project, planning and coordination of work will take place in weekly or biweekly meetings among Casabianca, Junker, and two graduate research assistants (RA).

In Years 1 and 2 we expect to publish 2–4 peer-reviewed articles deriving from the project per year, in methodological journals such as the *Journal of the American Statistical Association* and *Annals of Applied Statistics*, to domain-specific research journals like *Psychometrika*, *Journal of Educational and Behavioral Statistics*, *Psychological Assessment*, as well as journals related to the disciplines of teacher evaluation and Alzheimer's disease. In Year 3, publication will also include an R package for doing analyses with the HRM, in addition to peer-reviewed journal articles.

We will also make the work known through conference presentations, graduate coursework and seminars, and (under)graduate research projects.

9 Prior NSF Support

Junker received the NSF Award #SES-1229271, "Hierarchical Models for the Formation and Evolution of Ensembles of Social Networks", for the period September 2012–August 2014. The total award amount is $169,999.

9.1 Intellectual Merit

We are assembling candidate data sets for analysis for the goal of assessing how hierarchical network models can best be employed to model multiple social networks in social science and education settings. Candidates include the AddHealth and Spillane data mentioned in the proposal, as well as additional data sets on adolescent bullying in the US and on friendship ties in Afghanistan. In addition, we are beginning theoretical work on (a) assessing power to detect interventions on networks; and (b) understanding the relationship between network size and effect size in multiple network settings.

9.2 Broader Impact

We plan to publish widely and disseminate free software, in order to maximize the impact of our work. We will also train junior personnel, but because the award was made so recently we have not yet identified a suitable trainee.

9.3 Publications and other accomplishments

Because the award is so new, there are no publications to report at this time.