Journal of Educational and Behavioral Statistics
September 2007, Vol. 32, No. 3, pp. 287–314
DOI: 10.3102/1076998606298033
© AERA and ASA. http://jebs.aera.net

Covariates of the Rating Process in Hierarchical Models for Multiple Ratings of Test Items

Louis T. Mariano
RAND Corporation

Brian W. Junker
Carnegie Mellon University

When constructed response test items are scored by more than one rater, the repeated ratings allow for the consideration of individual rater bias and variability in estimating student proficiency. Several hierarchical models based on item response theory have been introduced to model such effects. In this article, the authors demonstrate how these models may be extended to include covariates of the rating process. For example, how do features of an essay grader's training affect his or her performance? The authors show how to include covariates by embedding a linear model at appropriate levels of the model hierarchy. Depending on the level, such covariates may be thought of as determining fixed effects or random effects on the rating process. The authors also discuss the appropriate design matrix for such covariates, discuss how to incorporate needed identifiability constraints, and illustrate the methods using data from a rating study of a student assessment.

Keywords: *repeated measures; item response theory; rater bias; Bayesian hierarchical models*

Section 1: Introduction

Repeated ratings are often collected in the scoring of constructed response test items. For example, to understand a student's writing proficiency, a single student's essay on a standardized test may be scored by more than one rater (also referred to as a reader or a grader). The subjectivity of the rating process causes an extra level of variability, that attributable to the raters, to be considered in the examination of examinee proficiency. A variance components analysis, such as that provided by generalizability theory, facilitates examination of the performance

This research was supported in part by an Educational Testing Service Harold T. Gulliksen Psychometric Fellowship awarded to Mariano. The opinions expressed are solely the authors' and do not represent those of their institutions or sponsors. The authors thank the California Department of Education and CTB/McGraw-Hill for making available data from the Golden State Exam Image-Scoring Pilot Study and Matthew Johnson for sharing his original software for the Hierarchical Rater Model (HRM).

efficiency of the raters as a group. The availability of repeated ratings, however, allows for the examination of individual rater performance as well. Research in this area has been dominated by applications to educational testing. Current literature provides several modeling options for within-rater performance, based wholly or in part on item response theory (IRT). Engelhard (1994, 1996) demonstrates the use of the Facets model (Linacre, 1989) with rated responses in analyzing the quality of rater judgments in an assessment of written composition and in the evaluation of rater performance against a set of benchmark expert ratings. Patz and Junker (1999b) elaborate on this model. Patz, Junker, Johnson, and Mariano (2002) include an analysis of rater bias and variability in a rating study of the Florida Comprehensive Assessment Test (FCAT) using the Hierarchical Rater Model (HRM). Verhelst and Verstralen (2001) present a different version of a hierarchical IRT model for multiple raters (MMR). Wilson and Hoskens (2001) use the same FCAT data set to demonstrate their Rater Bundle Model, which generalizes the Facets model to account for correlated ratings.

In considering the performance of the raters, it is natural to question how covariates of the rating process may affect a rater's performance. Covariates informative on the performance of the raters may prove valuable both in the ability to adjust for the covariate effects and in providing information to be used in the future training of raters. For example, if a particular group of essay graders seated at a common grading table exhibits bias in the form of harsher grading, proficiency estimates may be adjusted for that table's bias; if essay graders exhibit increased variability on the 2nd day of grading, in the future, a brief training review session could be implemented at the beginning of the 2nd day.

In this article, we present a methodology for including covariates of rater behavior within the structure of three different hierarchical IRT-based models: Patz et al.'s (2002) HRM, Verhelst and Verstralen's (2001) IRT MMR and a hierarchical version of Linacre's (1989) Facets model. These three models are reviewed in Section 2. In Section 3, we demonstrate a generalized format for characterizing the covariates and two different strategies for incorporating the covariates into our target models. In Section 4, we use data from an image-scoring pilot study of California's Golden State Examination (GSE) to demonstrate and contrast both of these methods in the HRM. The article concludes with a discussion in Section 5.

Section 2: Rater Models

In this section, we review three Bayesian hierarchical models for repeated ratings, which we will later expand to accommodate the rating covariates. All three may be used for both dichotomous and polytomous rating categories; throughout, we will demonstrate the general polytomous case. We will consider the scenario of N independent subjects $i \in \{1, \ldots, N\}$, each with underlying latent trait θ_i, responding to an administration of J independent items $j \in \{1, \ldots, J\}$, each having features—item location, discrimination—characterized by a vector of parameters.

Each item j has K_j response categories, indexed by $k \in \{0, \ldots, K_j - 1\}$. The observed data $X_{ijr} = k$ represent the rating k given to subject i's response to item j by rater r. We let R_{ij} denote the set of raters who rate subject i's response to item j, where there are Q raters total. Note that the design does not need to be fully crossed (i.e., not all subjects need to respond to all items, and not all raters need to rate all responses).

The three rater models we consider here all contain IRT components, which may be seen as an extension of an IRT model for polytomous items. In this article, we focus on Muraki's (1992) Generalized Partial Credit Model (GPCM; see also Muraki, 1997), but in other contexts, we would use a different IRT model as appropriate. The item category response functions (ICRFs), $P(X_{ijr} = k|\theta_i)$, for the GPCM are generated from

$$\ln \frac{P(Y_{ij} = k|\theta_i)}{P(Y_{ij} = k-1|\theta_i)} = \alpha_j(\theta_i - \beta_j - \gamma_{jk}), \tag{1}$$

where Y_{ij} is an objective item score, α_j is the item discrimination parameter, and β_j is the overall item location. Category-specific deviations from the overall location are represented by the item step parameters γ_{jk}. The sum of the item location and item step parameters, $\beta_j + \gamma_{jk}$, determine where the ICRFs for adjacent categories intersect.

For identifiability in all the examples we discuss, we constrain the latent parameter θ_i, setting its expectation to zero, $E(\theta_i) = 0$, and its variance to one, $\mathrm{Var}(\theta_i) = 1$. The overall item location parameter β_j is left free, and the item step parameters γ_{jk} are set to sum to zero.

The Hierarchical Rater Model

Patz (1996) and Patz et al. (2002) introduced the HRM, a hierarchical Bayesian model for discrete rated response data that incorporates an IRT model into a generalizability theory (Brennan, 1992) structure, allowing for the identification of the contribution of multiple sources of measurement error to the variability of observations. The HRM takes advantage of the natural hierarchical structure of the sources of variability, modeling the distribution of the latent trait, the distribution of a subject's response given their latent trait, and the distribution of the ratings given the quality of response. Thus, the HRM treats the administration and scoring of items as a two-stage process: First, the subject responds to the item, and then the rater evaluates that response so that the rating is a direct consequence of the response.

The HRM incorporates the following hierarchy, which is discussed in detail below:

$$\left.\begin{array}{rcl} \theta_i & \sim & i.i.d.\, N(\mu, \sigma^2),\ i = 1, \ldots, N, \\ \xi_{ij} & \sim & \text{a polytomous IRT model (e.g., GPCM)},\ \forall i,\ j = 1, \ldots, J, \\ X_{ijr} & \sim & \text{a polytomous signal detection model},\ \forall (i,j), r \in R_{ij}. \end{array}\right\}, \tag{2}$$

Within each examinee, let ξ_{ij} represent the latent quality of response reflected in subject i's response to item j (i.e., ξ_{ij} is the value that the ratings attempt to quantify). Patz et al. (2002) use the terminology *ideal ratings* or *ideal scores* for the ξ_{ij}, which is adopted herein. Consider the latent ideal scores ξ_{ij} in the context of data augmentation (Tanner & Wong, 1987; see also Maris, 1995). If these scores were known and discrete, they could be modeled using a traditional IRT model, such as the GPCM of Equation 1 (i.e., they could be treated as traditional objective scores). The HRM models the subject responses using such a traditional IRT model for the ideal scores.

The actual ratings X_{ijr} are treated as noisy versions of the ideal ratings ξ_{ij}, modeled with a simple signal detection model. As an example, consider five scoring categories:

$$\begin{array}{ll|ccccc}
 & & \multicolumn{5}{c}{\text{Observed Rating } (k)} \\
 & & 0 & 1 & 2 & 3 & 4 \\
\hline
 & 0 & p_{00r} & p_{01r} & p_{02r} & p_{03r} & p_{04r} \\
\text{Ideal} & 1 & p_{10r} & p_{11r} & p_{12r} & p_{13r} & p_{14r} \\
\text{Rating} & 2 & p_{20r} & p_{21r} & p_{22r} & p_{23r} & p_{24r} \\
(\xi) & 3 & p_{30r} & p_{31r} & p_{32r} & p_{33r} & p_{34r} \\
 & 4 & p_{40r} & p_{41r} & p_{42r} & p_{43r} & p_{44r}
\end{array} \tag{3}$$

Here, $p_{\xi kr} \equiv P(\text{rater } r \text{ rates } k|\text{ideal score } \xi)$, with the sum over the row of observed ratings, k, set equal to one. The signal detection parameters may be extended to capture potential interaction between examinees, items, and raters as well.

As a Bayesian model, appropriate prior distributions are assigned to the population, item, item step, and signal detection parameters.

Note that we have just described a class of HRMs. Whereas normality is a popular choice for the latent trait distribution, alternative distributions may be used when normality is not a reasonable assumption. Similarly, any parametric or nonparametric partial credit model (e.g., Hemker, Sijtsma, Molenaar, & Junker, 1996), such as Samejima's (1969) Graded Response Model (GRM), may be used to model the ideal scores.

To incorporate the concepts of rater severity and consistency, Patz et al. (2002) parameterize each row of the signal detection model by setting the probabilities $p_{\xi kr}$ proportional to a normal density in the categories k, with location $\xi + \phi_r$ and scale ψ_r:

$$p_{\xi kr} = P(X_{ijr} = k|\xi_{ij} = \xi) \propto \exp\left\{-\frac{1}{2(\psi_r)^2}[k - (\xi + \phi_r)]^2\right\}. \tag{4}$$

Here, the location parameter ϕ_r indicates the rater's severity. Following Patz et al. (2002), we refer to this severity parameter as rater bias. Negative and positive bias values indicate that the rater is more severe or less severe, respectively, than the rating guidelines indicate. A bias greater in magnitude than 0.5 indicates that the rater favors an adjacent category over the ideal. The scale parameter ψ_r indicates the rater's variability. Variability values near zero imply that the rater has a high level of consistency in applying the guidelines, with higher values indicating a diminished consistency.

The HRM is fit using a Markov Chain Monte Carlo (MCMC; e.g., Chib & Greenberg, 1995; Gelman, Carlin, Stern, & Rubin, 1995) algorithm to sample from the posterior distribution of the model parameters. Patz et al. (2002) describe the complete conditional distributions necessary to implement a Metropolis-Hastings within Gibbs MCMC procedure.

A Hierarchical Facets Model

Linacre (1989) generalized the Rasch model (Rasch, 1960) to include additional sources of variability as additive effects on the logit scale. This extension, commonly known as the Facets model, provides a framework to include rater bias in the logistic context and is similar in form to the Linear Logistic Test Model (Fischer, 1973). Let ϕ_r represent the fluctuation in item difficulty caused by the bias of rater r. Then, the model

$$\ln \frac{P(X_{ijr} = k|\theta_i)}{P(X_{ijr} = k-1|\theta_i)} = \theta_i - \beta_j - \gamma_{jk} - \phi_r \tag{5}$$

is a special case of Linacre's Facets model that accounts for rater bias. The ratings are treated as independent direct evaluations of the latent trait, accounting for the rater bias by shifting the difficulty of the item (i.e., the difficulty of the item is now $\beta_j - \phi_r$ when rater r rates item j). Notice also that there is now a unique set of IRTs for each rater. Interaction terms may also be included additively on the logit scale, and the rater bias term, ϕ_r, could be expanded to account for rater interactions with the item and examinee. For consistency with the traditional form of the Facets model and ease of exposition below, we present the model below without an item discrimination parameter, α_j. Of course, a more general model, as in Equation 1, may also be considered.

Following Patz and Junker (1999b), we may formulate a Bayesian hierarchical version of the Facets model by treating each latent trait θ_i of the subject population as an observation from a normal distribution centered at μ with variance σ^2 and assigning appropriate prior distributions to the item, item step, rater, and population parameters. The Bayesian Facets hierarchy is then:

$$\left.\begin{array}{lll} \theta_i & \sim & i.i.d.\, N(\mu, \sigma^2), i = 1, \ldots, N, \\ X_{ijr} & \sim & \text{an IRT Facets model (Equation 5)}, \forall i,\ j = 1, \ldots, J,\ r \in R_{ij}. \end{array}\right\} \tag{6}$$

A discrimination parameter, α_j, may be included in Equation 5, taking the generalized form present in Equation 1. Below, we illustrate the $\alpha_j \equiv 1$ case. Notice that the Facets formulation also requires an additional identifiability constraint (otherwise, adding any arbitrary constant to all β_j and subtracting the same amount from all ϕ_r yields the same ICRFs). A common choice is

$$\sum_{r=1}^{Q} \phi_r = 0.$$

A crucial difference between the HRM and Facets models is the independence structure of the responses assumed by each model. IRT Facets treats the ratings, including repeated ratings, as independent evaluations of the subject's latent trait, so that the repeated ratings are independent given the latent trait. The HRM treats the ratings as independent evaluations of the quality of the response, so that the repeated ratings are independent conditional on the ideal scores but are dependent when conditioning only on the latent trait. These independence structures help guide the appropriate choice of model; the structure among repeated ratings of a single response to a single item is better reflected in the HRM, whereas ratings of multiple responses to a single item are better reflected in the Facets model.

HRM and Facets also differ in how they characterize rater performance—a shift in item difficulty representing bias under Facets versus a two-dimensional representation affecting the center (representing bias) and scaling of the probability distribution describing the rating under the HRM. Note that this means that the bias parameter ϕ_r has different interpretations of bias in the two models.

Patz and Junker (1999a, 1999b) explain the general implementation of an MCMC algorithm for sampling from the posterior distribution of IRT model parameters. MCMC for the hierarchical Facets model is included in Patz and Junker (1999b).

Verhelst and Verstralen's IRT Model for Multiple Raters

Verhelst and Verstralen's (2001) IRT MMR uses a hierarchical structure similar to the HRM, treating the administration as a two-stage process where the subject first responds to the item and then the response is evaluated. However, although ideal scores in HRM are discrete, the MMR allows a continuous quality of response variable ξ_{ij} (note that the *ideal score* terminology does not apply, because the quality of work is not on the same metric as the ratings). Here, the ξ_{ij} are treated as independent observations from a normal distribution centered to reflect subject–item interaction and with a common variance across all subjects and responses,

$$\xi_{ij} \sim N(\theta_i - \beta_j^*, \sigma_\xi^2).$$

Under the MMR, the ratings are then described using an IRT Facets model similar to Equation 5 with the latent trait θ_i being replaced by the quality of response ξ_{ij}.

$$\ln \frac{P(X_{ijr} = k|\xi_{ij})}{P(X_{ijr} = k-1|\xi_{ij})} = \xi_{ij} - \beta_j - \gamma_{jk} - \phi_r, \tag{7}$$

where ϕ_r again represents rater bias and is constrained as in the hierarchical Facets model to sum to zero. An item discrimination parameter α_j may also be included as in Equation 1; for ease of exposition, we discuss the $\alpha_j \equiv 1$ case below.

The full hierarchy of the MMR is then

$$\left.\begin{array}{rcl} \theta_i & \sim & i.i.d.\,N(\mu, \sigma^2),\ i = 1, \ldots, N, \\ \xi_{ij} & \sim & ind.\,N(\theta_i - \beta_j^*, \sigma_\xi^2),\ \forall i, j = 1,\ \ldots, J, \\ X_{ijr} & \sim & \text{the IRT model of Equation (7)}\ \forall (i,j),\ r \in R_{ij}. \end{array}\right\} \tag{8}$$

In the Bayesian context, we place appropriate prior distributions on the population, item, item step, and rater parameters. Similar to the other models discussed, we may sample from the posterior distribution of the parameters of the MMR using an MCMC algorithm.

In the quality of response level of the model, the notation β_j^* indicates that the item contribution is not necessarily the traditionally defined item location. In the usual IRT context, location is relative to the scoring categories, which, in the constructed response case, are defined by the rating rubric (i.e., "How difficult is it to receive a score of k, given the respondent's latent trait, θ_i, and the definitions of the rating categories as determined by the rubric?"). This notion of item location is captured by β_j in the IRT level of the model. However, when allowing for an item effect in the quality of response level, that effect will not be dependent on the scoring rubric and the rating categories it defines. Note that β_j^* and β_j are not separately identifiable. Verhelst and Verstralen (2001) transform the current parameterization so that the sum $\delta_j = \beta_j^* + \beta_j$ is estimated as a single parameter. For a complete treatment of the estimation of the item parameters, see Verhelst and Verstralen (2001).

A dichotomous response version of this model is presented by Verhelst and Verstralen (2001); we have extended this to the polytomous response case for our purposes below. We have also reparameterized their original rater term, multiplying by -1, so that the rater bias is in terms of harshness instead of leniency, providing consistency with the other two models above. Finally, Equation 7 may also be expanded to include a discrimination parameter.

The independence structure of the MMR is similar to the HRM. However, the characterization of rater performance under the MMR is similar to that of the

Facets model. These differences, along with the representation of the quality of work, govern the appropriate model choice among these three models.

Section 3: Including Measurement Covariates

In this section, we provide a general method for describing the covariate effects and including them into the models described in Section 2. Although we will comment generally about the effects of rating covariates on both the bias and variability of the ratings, the reader should keep in mind that of the models discussed, in describing rater performance, the HRM incorporates both rater bias and variability but facets and the MMR only provide a characterization of bias.

Let $Z=\{Z_1,Z_2,\ldots,Z_S\}$ be values of a set of S covariates (or covariate factors) under which a rater may produce a rating. Here, a quantitative covariate would occupy a single Z_s, whereas in the C-categorical covariate case, each of the C factors might occupy its own indicator covariate $Z_c\in\{0,1\}$, with $Z_s+\cdots+Z_{s+C}=1$. In considering the rating covariates, the bias and variability of the ratings are influenced not only by the individual rater but also by the covariates under which the rater is performing. A rater may rate under different conditions, coded by the covariates, at different times (i.e., the values of Z may change within raters). Some covariate values may differ with each response. For example, the rater covariate "time, in seconds, to complete the rating" may be unique to each response. Other covariates may be constant over a set of individual responses, producing covariate values that apply over a range of rated responses. For example, the rating table at which an exam rater is seated may be constant.

To capture differences in a rater's bias and variability across various combinations of rater covariates, it is necessary to consider each unique Rater × Covariate combination present in the data separately, examining the effect of each Rater × Covariate combination on the probability of scoring in a particular category. Where we originally considered the effect of only the rater, we will now consider the effect of the individual Rater × Covariate combinations, referring to these Rater × Covariate combinations as pseudoraters.

Let V represent the total number of unique pseudoraters present in the ratings, with arbitrary ordering $v=1,\ldots,V$. Let ρ_v and ω_v represent the contributions to bias and variability, respectively, by pseudorater v. Note that V is, at minimum, equal to Q, the total number of raters (if each rater always rates under the same set of covariate values). Expand the notation of the data as $X_{ij\omega}=k$, to indicate that pseudorater ω scored examinee i's response to item j in category k, and consider the data in this more detailed form, treating each pseudorater separately in the models of Section 2. In the hierarchical facets model, the facets level may be reexpressed as

$$\ln\frac{P(X_{ijv}=k|\theta_i)}{P(X_{ijv}=k-1|\theta_i)}=\theta_I-\beta_j-\gamma_{jk}-\rho_v, \tag{9}$$

so that each unique pseudorater would now have its own set of item response functions. An analogous reexpression may also be made in the MMR:

$$\ln\frac{P(X_{ijv}=k|\xi_{ij})}{P(X_{ijv}=k-1|\xi_{ij})}=\xi_{ij}-\delta_j-\gamma_{jk}-\rho_v. \tag{10}$$

In the HRM, each pseudorater will now have its own signal detection model, with cell entries

$$p_{\xi kv}=P(X_{ijv}=k|\xi_{ij}=\xi)\propto \exp\left\{-\frac{1}{2\omega_v^2}[k-(\xi+\rho_v)]^2\right\}, \tag{11}$$

which corresponds to extending the original signal detection model of the raters to include interaction with the covariates.

To characterize the various combinations of raters and covariates present in the rating process, let Y be a design matrix containing V rows, one for each pseudorater and a total of $Q+S$ columns. The first Q columns of Y are indicators for each rater, and the remaining S columns hold the rating covariate values. For example, if pseudorater v corresponds to the second of four raters and there are three covariates, Z_{1v}, Z_{2v}, Z_{3v}, then row $Y_v=(0,1,0,0,\ Z_{1v},\ Z_{2v},\ Z_{3v})$. Of course, if controlling for the individual rater effects is not of interest, the corresponding first Q columns of Y may be eliminated.

To parameterize bias in the rating process, let $\eta=(\phi_1,\ldots,\phi_Q,\ \eta_1,\ldots,\ \eta_S)^T$, where η_S represents the bias effect of covariate (or covariate factor) Z_s. A linear model may be developed for the rating bias as

$$\rho_v=Y_v\eta. \tag{12}$$

Similarly, we may build a model for variability in the rating process (on the log scale) with rater and covariate variability effects $\ln\tau^2=(\ln\psi_1^2,\ldots,\ \ln\psi_Q^2,\ \ln\tau_1^2,\ldots,\ \ln\tau_S^2)^T$ as

$$\ln\omega_v^2=Y_v(\ln\tau^2). \tag{13}$$

In evaluating the effects of rater covariates, we wish to identify those individual covariates for which differences in the bias or variability effects associated with the available values of the covariate are nonzero. Such nonzero covariate effects may then be addressed in rater training, assignment of item responses to raters, and so on.

We next explore two possibilities for incorporating these rater behavioral models into the hierarchical structure of the models discussed in Section 2. The first is to include these linear structures at the same level as the rating data (the fixed rating effects option); the second is to include them one level removed from the rating data (the random rating effects option). Both options are then illustrated in Section 4 with data from an image-scoring pilot study of California's GSE, which contains multiple rater covariates.

Incorporating the Rating Covariates: Fixed Rating Effects

To include the covariates of rating bias at the same level as the rating data in the hierarchical Facets model, we replace the pseudorater bias term ρ_v in Equation 9 with the linear structure for the rating bias in Equation 12:

$$\ln \frac{P(X_{ijr}=k|\theta_i)}{P(X_{ijr}=k-1|\theta_i)} = \theta_i - \beta_j - \gamma_{jk} - Y_v\eta. \tag{14}$$

Appropriate prior distributions then need to be assigned to the covariate bias effects in η. An identical replacement in the facets level of the MMR will include the covariates of rating bias in that model as well.

To include the linear structures of Equations 12 and 13 for rating bias and variability at the same level as the rating data in the HRM, Equation 4 is augmented to

$$p_{\xi k v} = P(X_{ijv}=k|\xi_{ij}=\xi) \propto \exp\left\{-\frac{[k-(\xi+Y_v\eta)]^2}{2(\exp\{Y_v(\ln\tau^2)\})}\right\}, \tag{15}$$

with appropriate prior distributions assigned to the covariate effects η and τ. Patz et al. (2002) extended the HRM specifically to demonstrate the effect of the distribution of items among raters in a rating study of the FCAT; their example may be viewed as a special case of the fixed rating effects option for including rating covariates in the HRM.

For any of these three models, the rater and covariate effects may need to be constrained for identifiability (i.e., the design matrix Y may need to be constrained if it is not of full rank). The examples in the next section highlight this necessity, which is further discussed in Section 5.

In Equation 15, notice the dependence of the covariate effects for bias η and variability τ on the ideal scores ξ. This complicates the MCMC algorithm to draw from the posterior distribution of the rating parameters. Traditionally (e.g., Patz & Junker, 1999a), MCMC is implemented on IRT models via the Metropolis-Hastings within Gibbs method (e.g., Gelman et al., 1995), which passes iteratively through the complete conditional distributions of the model parameters. Even though Equation 12 expresses a linear relationship between the pseudorater bias and the rater and covariate effects, the complete conditional distributions for the covariate effects are not in the usual form for a linear model, and care must be taken to properly express that portion of the likelihood proportional to each model parameter. Mariano (2002) provides the complete conditional distributions for the HRM when implementing the fixed rating effects option.

In the current formulation of Y above, each pseudorater occupies a single line in the design matrix. Unless there is a different Rater × Covariate combination for each observed rating, Y will be much smaller, and hence a simpler expression of the design, than a matrix that includes a row for each observed rating, with multiple rows being replicates. Note that in the fixed rating effects option outlined above, for any of the three rater models considered, the MCMC algorithm will still consider the effect of the pseudorater associated with each response through the model likelihood, so that each row Y_v is considered not once but the actual number of times that responses are associated with pseudorater v.

Incorporating the Rating Covariates: Random Rating Effects

The second option for incorporating rating covariates into the hierarchical facets model is to preserve the pseudorater bias term in the facets level of the model (as in Equation 9) and, one level removed from the data in the hierarchy, use the linear model implied by Equation 12 to model the pseudorater bias (i.e., for $\rho = \{\rho_1, \ldots, \rho_V\}$):

$$\rho_v \sim \mathrm{N}(Y_v\eta, \sigma_\rho^2). \tag{16}$$

The population level of the model remains unchanged, and prior distributions are assigned to the rater and covariate effects η and the variance term σ_ρ^2. Again, the extension for the MMR is similar.

To include rating covariates into the structure of the HRM, retain the rating probabilities of Equation 11 as the signal detection portion of the model, which includes the pseudorater contributions to bias ρ_v and variability ω_v. Then, one level removed from the data and ideal scores, include the linear structure of Equation 16 to describe the rating bias and model the rating variability as

$$\omega_v \sim \mathrm{N}(Y_v\tau, \sigma_\omega^2), \tag{17}$$

where $\omega = \{\omega_1, \ldots, \omega_V\}$. The rater and covariate effects, η and $\tau^2 = (\psi_1^2, \ldots, \psi_Q^2, \tau_1^2, \ldots, \tau_S^2)^T$, and the variance terms, σ_ρ^2 and σ_ω^2, are assigned appropriate prior distributions. Note that the pseudorater variability could also be modeled on the log scale ($E[\ln \omega_v^2] = Y_v(\ln \tau^2)$, analogous to the treatment in Equation 13. By choosing the formulation of Equation 17, the component parameters of τ are easily interpretable as additive effects on rating variability under a Rater × Covariate combination (as opposed to multiplicative effects on the square of rating variability).

As was the case for the fixed rating effects option, using any of the three models discussed with this random rating effects option may require that constraints be placed on the rater and covariate effects for identifiability. Naturally, these constraints will be reflected in the prior distributions of the affected parameters.

Note that in the linear models defined in Equations 16 and 17, a result of using the design matrix Y is that each pseudorater is only considered once instead of the number of times that pseudorater occurred in the data. An alternative here is a weighted regression scenario where each pseudorater is weighted by the number of ratings that pseudorater produced.

The extension implied by the random rating effects option to include rating covariates is larger than the models produced using the fixed rating effects option in the sense that there are $V+1$ additional parameters in the hierarchical Facets and MMR and $2V+2$ additional in the HRM. The pseudorater bias and variability parameters are not explicitly included in the fixed rating effects option, but they are included in the random rating effects option, along with parameters for their respective variances. However, the MCMC algorithm to sample from the model of the random rating effects option is actually simpler. To see this for the HRM, first factor the full posterior distribution of the model parameters as

$$f(\mu, \sigma, \theta, \beta, \gamma, \xi, \rho, \omega, \eta, \tau, \sigma_\rho^2, \sigma_\omega^2 | X, Y)$$

$$= f(\mu, \sigma, \theta, \beta, \gamma, \xi, \rho, \omega | X) \tag{18}$$

$$\times\, f(\sigma_\rho^2 | \rho)\, f(\eta | Y, \rho, \sigma_\rho^2) \tag{19}$$

$$\times\, f(\sigma_\omega^2 | \omega)\, f(\tau | Y, \omega, \sigma_\omega^2). \tag{20}$$

Samples may be drawn from the posterior distribution of the population, item, and pseudorater parameters of Equation 18 in the same way that samples are drawn from the full posterior distribution of the original HRM defined by Equations 2 and 4 above, with the pseudoraters taking the place of the actual raters. With η and τ one level removed from the data and ideal scores, they may be treated as coefficients in a Bayesian linear regression. A factorization of the posterior distributions of the hierarchical Facets and MMR will also yield an isolation of the bias parameters identical to Equation 19. The technique for sampling from these distributions is well known. Following Gelman et al. (1995), using the standard noninformative prior $f(\eta, \sigma_\rho^2) \propto \sigma_\rho^{-2}$,

$$\Sigma = (Y^T C^{-1} Y)^{-1}, \tag{21}$$

$$\hat{\eta} = (Y^T C^{-1} Y)^{-1} Y^T C^{-1} \rho, \tag{22}$$

$$s_\rho^2 = \frac{1}{df} (\rho - Y\hat{\eta})^T C^{-1} (\rho - Y\hat{\eta}), \tag{23}$$

$$\sigma_\rho^2|\rho \sim \text{Scaled Inverse } \chi^2(df, s_\rho^2), \tag{24}$$

$$\eta|\sigma_\rho^2, \rho \sim \mathrm{N}(\hat{\eta}, \Sigma\sigma_\rho^2). \tag{25}$$

Here, C^{-1} is a $V \times V$ diagonal weighting matrix, where the diagonal entry in row v is the number of responses rated by pseudorater v if such weighting is desired, alternatively, C is set to the identity matrix; df is the number of rows minus the number of columns in a properly constrained full rank version of Y. Of course, the design matrix used must also have more rows than columns to fit the model. To draw from the distributions specified in Equation 19, first draw σ_ρ^2 as specified in Equation 24, then draw η as in Equation 25. The variability parameters for the HRM follow a similar form.

Proper prior distributions may also be considered for the bias (and variability) parameters. However, if V is small, shrinkage toward the prior mean may be of concern if the unweighted version ($C = I$) is used. With the prior information about each coefficient having the influence of an additional data point (Gelman et al., 1995), if the number of pseudoraters is not large relative to the number of coefficients, the posterior distribution will be sensitive to the prior.

Notice in Equation 16 that if the variance term, σ_ρ^2, associated with pseudorater bias in the random rating effects option is driven toward zero, the resulting deterministic relationship between the pseudorater bias parameters and the corresponding rater and covariate effects, $\rho = Y\eta$, is exactly the relationship featured in the fixed rating effects option. This suggests that, subject to design constraints, the more complex fixed rating effects model might be fit by using the random rating effects option and setting $\sigma_\rho^2 = 0$. One possibility for revising the MCMC algorithm to accomplish this would be to draw the pseudorater bias ρ as in the random rating effects option and then set $\eta = (Y^T C^{-1} Y)^{-1} Y^T C^{-1} \rho$, as suggested by Equations 25 and 22. The pseudorater variability ω and variability effects τ in the HRM would follow similarly.

Section 4: Golden State Exam Image-Scoring Pilot Study

The California Department of Education and CTB/McGraw-Hill have provided data from their image-scoring pilot study of the GSE. Patz, Awamleh, and Kelly (1999) provide detail on the design of this two-part study implemented to help assess the potential use of a computerized, image-based scoring (image scoring) system instead of the traditional paper-and-pencil (paper scoring) method. Image scoring proceeds with the rater viewing an image of the response over a computer network and completing the rating online. In principal, this may be done at any location convenient to the rater. For the GSE study, the image scoring was completed in a centralized location in the traditional configuration of raters assembled at scoring tables with a table leader assigned to each. We consider data from part 1 of the study, which is made up of ratings of responses by 9,356

TABLE 1
Number of Ratings per Response (R) for the 9,356 Examinee Responses Found in Part 1 of the Golden State Exam Image-Scoring Pilot Study

R	4	3	2	1
Frequency	28	767	6,678	1,883

examinees to a single four-category constructed response item. These responses were randomly drawn from the complete set of 1998 GSE test papers to be rerated for the purposes of the study. The number of ratings per response *R* range from 1 to 4. Table 1 gives the distribution of *R* in the data.

To assist in establishing the difficulty of the item and the proficiency of the examinees, responses to 30 GSE objectively scored multiple-choice items were also provided.

$Q = 28$ raters in two study groups alternated between image scoring and paper scoring in two half-day sessions. Each group is made up of two tables of seven raters per table (including the table leader). Raters of varying experience were provided by two separate sources, labeled herein as source A and source B. The sources represent two groups that, although both are likely to be employed as raters, have demonstrably different backgrounds.

Below, we consider the effect of the rating environment (image scoring vs. paper scoring) to illustrate modeling a rating covariate at the same hierarchical level as the rating data (fixed rating effects option). Then, we consider the effects of rating environment, as well as rating table and source, to demonstrate modeling the rating covariates one level removed from the rating data in the model hierarchy (the random rating effects option). As the GSE rating data correspond multiple ratings of individual item responses as applied from a categorical scoring rubric, the appropriate dependence structure and expression of the quality of item response indicate use of the HRM for these examples.

Expanding the HRM

The single constructed response item in the GSE data does not support the estimation of the difficulty of the item and provides little information about the examinees' proficiencies. For this reason, we include in this analysis each examinee's responses to the 30 objectively scored multiple-choice questions on the 1998 GSE. To evaluate this additional data concurrently with the constructed response item, the HRM must be expanded to also include objectively scored data.

In a general setting, suppose a version of the HRM is desired to simultaneously consider J_c rated items and $J_o = J - J_c$ objectively scored (e.g., multiple-choice) items. This may be accomplished by modeling the objectively scored responses using an IRT model (which may or may not be the same IRT model used for the ideal scores within the hierarchical structure of Equation 2). This approach combines the original HRM with the Bayesian approach to IRT for objectively scored items illustrated by Patz and Junker (1999a). Let $X_{ij} = k$ denote that examinee i responded to objectively scored item j in category k. Then this new version of the HRM for both objectively scored and constructed response items (herein referred to as the HRMoc) has the following form:

$$\left.\begin{aligned} \theta_i &\sim i.i.d.\, N(\mu, \sigma^2), i = 1, \ldots, N \\ \xi_{ij} &\sim \text{a polytomous IRT model (e.g., GPCM)}, \ \forall i, j = 1, \ldots, J_c \\ X_{ij} &\sim \text{a polytomous IRT model (e.g., GPCM)}, \ \forall i, j = J_c + 1, \ldots, J \\ X_{ijr} &\sim \text{a polytomous signal detection model}, \ \forall i, j = 1, \ \ldots, J_c, \ r \in R_{ij} \end{aligned}\right\}, \tag{26}$$

and the revised HRMoc has the following expanded likelihood:

$$\underbrace{\prod_{i=1}^{N} p(\theta_i|\mu, \sigma^2) \left[\left(\prod_{j=1}^{J_c} p(\xi_{ij}|\theta_i, \beta_j, \gamma_j) \prod_{r \in R_{ij}} p(X_{ijr}|\xi_{ij}, \phi_r, \psi_r) \right)}_{\text{original HRM likelihood}} \left(\prod_{j=J_c+1}^{J_c+J_o} p(X_{ij}|\theta_i, \beta_j, \gamma_j) \right) \right]. \tag{27}$$

Similar extensions to the MMR and hierarchical facets are easily envisioned, with the likelihood factoring into independent parts representing the objectively and subjectively scored items.

For the GSE example, $J_c = 1$ and $J_o = 30$. In the analysis below, the GPCM (Equation 1) was used for the ideal scores of the item $j = 1$ responses, with the two-parameter logistic model (Birnbaum, 1968) used for the multiple-choice items. Rating covariates may be included in the HRMoc for the rated items using either option described in Section 3.

Fitting the model via MCMC is straightforward. The original Metropolis-Hastings within Gibbs sampling algorithm of the HRM (Patz et al., 2002), as expanded in Section 3 for rating covariates, may be used; each objectively scored response is treated as an ideal score that remains constant at the actual scored value instead of being updated at every iteration.

When considering the GSE data, in particular the feature of a single constructed response item, it is necessary to place an additional constraint on the standard model. Table 2 details the distribution of ratings in each of the four rating categories for this item.

More than 50% of the ratings are in category 0, and slightly less than 90% are in category 0 or 1. Either this is a very difficult item or the raters as a group are misapplying the scoring rubric such that they are causing a negative bias. Without an additional item of less observed difficulty, we cannot distinguish between these two cases. For this illustration, we assume that the raters as a group understand the rubric well enough to apply it properly. We implement this

TABLE 2
Observed Distribution of Ratings in Each Scoring Category for the Golden State Exam Image-Scoring Pilot Study

Rating	0	1	2	3
Frequency	9,417	6,337	1,312	586

assumption by constraining the average rater bias to be less than 0.5 in magnitude. Note in Equation 4 that if an individual rater bias term is less than -0.5, then that rater will have higher probability of scoring in the adjacent lower category than of scoring in the ideal. In the fixed effects option, the constraint is implemented on the average of the covariate bias; for random effects, the constraint is placed on the average pseudorater bias. Finally, note that one could proceed with this particular data set without this additional constraint; however, the MCMC algorithm chosen would have to be modified to account for the resulting bimodal distributions.

Example: Fixed Rating Effects

Having extended the HRM to accommodate the multiple-choice GSE data, the effects of the GSE rating study covariates may now be analyzed. Recall that the primary goal of the GSE rating study was to assist in addressing the viability of an image-based scoring system. A natural question to consider is whether differences in rating environment resonate in the rating process in the presence of individual rater effects (i.e., are rating bias and variability unique to individual raters, or can a portion of these effects be attributed to the rating environment?). In this subsection, we investigate this question using the HRMoc model (Equations 26 and 27) with $J_c = 1$ constructed response item, $J - J_c = 30$ objective/multiple-choice items, and fixed rater covariate effects. Recall (Equation 15) that in this fixed effects model, we decompose the pseudorater location parameters linearly as $\rho_v = Y_v \eta$ and the pseudorater scale parameters linearly as $\ln \omega_v^2 = Y_v(\ln \tau^2)$. For this example, the rating design matrix Y has 56 rows, one for each of 28 raters in two environments, and 30 columns; the first 28 columns are indicators of the 28 individual raters, and the final 2 columns are indicators of the rating environment. For example, the row for rater 5 when image scoring is

$$Y_5 = (\underbrace{0000100000000000000000000000}_{\text{rater}}\underbrace{10}_{\text{env.}}).$$

Y is only of rank 29, indicating that the model is not fully identified. To see this, notice that the pseudorater bias term $Y_v \eta$ in Equation 15 reduces to $\phi_r + \eta_s$, where η_1 is the effect of image scoring and η_2 the effect of paper scoring. Thus, adding any constant value to all the η_ss and subtracting that same value for all the ϕ_rs will yield the exact same pseudorater probabilities in the signal detection model portion of the HRMoc. A similar problem exists for the pseudorater variability term $Y_v(\ln \tau^2)$ on the log scale. Ultimately, we are interested in understanding the differences, if any, between rating environment effects. These differences are well defined in Y (i.e., adding any constant value to all the η_ss does not change the value of $\eta_1 - \eta_2$). Below, we demonstrate constraining the design matrix to directly expose the differences. See Section 5 for a discussion of alternatives.

Because the effect of the rating environment is our question of interest and the individual rater effects are nuisance parameters, we impose sum-to-zero constraints on the bias and variability of the individual raters:

$$\sum_{r=1}^{28} \phi_r = 0 \text{ and } \sum_{r=1}^{28} \ln \psi_r = 0. \tag{28}$$

Constraining the individual rater effects in this way centers the rating environment effects at the overall mean effects of bias and variability present in the ratings so that the environment effects may be interpreted as deviations from overall rating behavior induced by each particular environment. Alternate constraints, such as sum-to-zero constraints on the rating environment effects or fixing the bias and variability of the first rater to 0 and 1, respectively, would impart different meaning to the individual rating environment parameters while still producing the same estimates for their differences.

To complete the model specifications, we followed Patz et al. (2002), using vague but proper prior distributions for the item and rating parameters. The γ_{1k} received Normal$(0, 3)$ priors, and the β_j, $\ln \alpha_j$, ϕ_r, $\ln \psi_r$, η_s, and $\ln \tau_s$ all received Normal$(0, 10)$ priors, conditional on the model constraints. We programmed the MCMC algorithm in C++. For the results we discuss below, parameter estimates are based on 10,000 iterations taken from five MCMC chains, after a burn-in of 3,000 iterations. The method of Gelman and Rubin (e.g., Gelman et al., 1995) was used to assess convergence. Posterior median and 95% credible interval estimates of the item parameters are displayed in Table 3.

Because the focus of this work is the influences that may affect performance of the raters, we forego additional discussion of the estimated item parameters in the interest of space and focus on the rating covariate estimates below.

Table 4 displays 95% credible interval estimates of the difference in the two environment bias parameters $\eta_1 - \eta_2$ and the two environment variance parameters $\tau_1 - \tau_2$.

There is a small but detectable difference in the biases. However, this difference is small enough to be of little operational significance. With individual rater

TABLE 3
Posterior Median and 95% Equal-Tailed Credible Interval (CI) Estimates of the Item Parameters Golden State Exam Image-Scoring Pilot Study

Parameter	Median	95% CI	Parameter	Median	95% CI
Constructed response					
α_1	1.142	(1.071, 1.215)	γ_{11}	−1.432	(−1.495, −1.372)
β_1	1.468	(1.406, 1.534)	γ_{12}	0.757	(0.663, 0.859)
			γ_{13}	0.673	(0.560, 0.782)
Multiple-choice					
β_2	−0.666	(−0.724, −0.612)	α_2	1.072	(0.996, 1.148)
β_2	−0.532	(−0.595, −0.473)	α_3	0.913	(0.847, 0.982)
β_4	−0.198	(−0.242, −0.153)	α_4	1.210	(1.134, 1.288)
β_5	0.662	(0.581, 0.751)	α_5	0.693	(0.635, 0.749)
β_6	−0.447	(−0.529, −0.372)	α_6	0.659	(0.600, 0.717)
β_7	0.897	(0.779, 1.013)	α_7	0.525	(0.474, 0.579)
β_8	−0.751	(−0.823, −0.680)	α_8	0.854	(0.787, 0.922)
β_9	1.834	(1.557, 2.262)	α_9	0.293	(0.240, 0.342)
β_{10}	0.352	(0.281, 0.428)	α_{10}	0.682	(0.623, 0.737)
β_{11}	0.227	(0.147, 0.310)	α_{11}	0.580	(0.525, 0.635)
β_{12}	−0.188	(−0.244, −0.130)	α_{12}	0.836	(0.776, 0.898)
β_{13}	2.802	(2.134, 3.788)	α_{13}	0.149	(0.110, 0.193)
β_{14}	0.625	(0.459, 0.842)	α_{14}	0.268	(0.218, 0.317)
β_{15}	−0.163	(−0.265, −0.064)	α_{15}	0.444	(0.395, 0.495)
β_{16}	2.521	(1.814, 4.132)	α_{16}	0.125	(0.078, 0.170)
β_{17}	0.362	(0.269, 0.462)	α_{17}	0.495	(0.442, 0.548)
β_{18}	0.927	(0.806, 1.060)	α_{18}	0.477	(0.428, 0.531)
β_{19}	−0.251	(−0.309, −0.190)	α_{19}	0.837	(0.774, 0.899)
β_{20}	−0.522	(−0.594, −0.456)	α_{20}	0.775	(0.714, 0.837)
β_{21}	1.107	(0.891, 1.365)	α_{21}	0.275	(0.231, 0.324)
β_{22}	0.381	(0.301, 0.468)	α_{22}	0.580	(0.527, 0.635)
β_{23}	0.050	(−0.022, 0.121)	α_{23}	0.633	(0.581, 0.688)
β_{24}	0.203	(0.137, 0.267)	α_{24}	0.731	(0.671, 0.790)
β_{25}	0.246	(0.175, 0.325)	α_{25}	0.639	(0.584, 0.695)
β_{26}	−0.674	(−0.751, −0.604)	α_{26}	0.788	(0.723, 0.853)
β_{27}	−0.425	(−0.494, −0.359)	α_{27}	0.723	(0.664, 0.785)
β_{28}	0.205	(0.137, 0.272)	α_{28}	0.698	(0.641, 0.756)
β_{29}	0.953	(0.815, 1.124)	α_{29}	0.412	(0.362, 0.463)
β_{30}	0.070	(0.011, 0.128)	α_{30}	0.775	(0.717, 0.834)
β_{31}	0.235	(0.188, 0.284)	α_{31}	1.071	(1.004, 1.140)

bias and variability held fixed at $\phi_r = 0$ and $\psi_r = 1$, respectively, the associated signal detection models do not contain any cell differences greater than .037.

TABLE 4

Posterior Median and 95% Equal-Tailed Credible Interval (CI) Estimates of the Covariate Effects of Rating Environment (Image vs. Paper) Present in the Golden State Exam Image-Scoring Pilot Study as Fit Using the HRMoc With Fixed Rating Effects

		Bias			Variability	
Covariate		Median	95% CI		Median	95% CI
Image	η_1	−0.154	(−0.179, −0.129)	τ_1	0.426	(0.410, 0.439)
Paper	η_2	−0.126	(−0.153, −0.098)	τ_2	0.412	(0.396, 0.424)
	$\eta_1 - \eta_2$	−0.028	(−0.049, −0.007)	$\tau_1 - \tau_2$	0.015	(−0.004, 0.033)

Note: Sum-to-zero constraints were imposed on the raters. HRMoc = Hierarchical Rater Model for both objectively scored and constructed response items.

The variability attributable to the two rater sources is comparable. Thus, from the perspective of practicality, no important operational differences have been detected by imaging scoring instead of paper scoring when performed within a common setting for the raters.

Example: Random Rating Effects

To illustrate the random rating effects option described in Section 3, we use the GSE rating study data to examine the covariate effects of rating environment, rater source, and rater table assignment. We again use the HRMoc model (Equations 26 and 27) with $J_c = 1$ constructed response item and $J - J_c = 30$ objective/multiple-choice items, this time treating the rater covariate effects as random. Recall that this random effects model is implemented by retaining the pseudorater bias ρ_v and variability ω_v terms at the level of the rating data (as in Equation 11) and, one level below in the model hierarchy, decomposing the pseudorater bias and variability into individual rater and covariate effects using a linear model for each (Equations 16 and 17):

$$\rho_v \sim \mathrm{N}(Y_v \eta, \sigma_\rho^2), \qquad \omega_v \sim \mathrm{N}(Y_v \tau, \sigma_\omega^2).$$

The rating table assignments of each of the $Q = 28$ individual raters, as well as their sources, are displayed in Table 5. Although the raters all rated in both the image-scoring and paper-scoring environments, they are also nested both within rater source and within rating table, both of which remained constant throughout the rating process. This design produces $V = 2{*}28 = 56$ Rater × Covariate combinations (56 pseudoraters). Considering a model that includes all three of these covariates, there are eight covariate factor effects to consider. The full design matrix for the rater and covariate effects has 36 columns: The first 28 of these are rater indicators, the next 4 are table indicators, followed by 2 source indicators, and finally, 2 environment indicators. Let $Y_{r,t,s,e}$ indicate the row of the

TABLE 5
Rater Source and Table Assignments Present in the Golden State Exam Image-Scoring Pilot Study

Rating Table 1		Rating Table 2		Rating Table 3		Rating Table 4	
Rater	Source	Rater	Source	Rater	Source	Rater	Source
1	A	8	A	15	A	22	A
2	A	9	A	16	A	23	A
3	A	10	B	17	B	24	B
4	A	11	A	18	B	25	A
5	A	12	A	19	A	26	A
6	B	13	A	20	A	27	A
7	B	14	B	21	B	28	B

Note: Each rater scored under both the image-scoring and paper-scoring environments.

design matrix Y corresponding to rater r, who is from source s and seated at table t, when rating in environment e. Here, $e = 1$ indicates image scoring, $e = 2$ denotes paper scoring, and $s = 1$ indicates that the rater was provided by source A, whereas $s = 2$ indicates a source B rater. We have, for example, the row for rater 9 when paper scoring:

$$Y_{9,2,1,2} = (\underbrace{0000000010000000000000000000}_{\text{rater}}\overbrace{0100}^{\text{table}}\ \underbrace{10}_{\text{source}}\ \overbrace{01}^{\text{env.}}).$$

Similarly, let η_{t1}, η_{t2}, η_{t3}, η_{t4} represent the table bias effects of tables 1 through 4, respectively, η_{s1} and η_{s2} the source A and source B bias effects, respectively, and η_{e1} and η_{e2} the rater bias effects of image and paper scoring, respectively, with analogous parsing of the variability effects τ.

The design matrix described above is not of full rank; it has 36 columns but is only of rank 29. As was the case with the fixed effects example, to expose the differences in rating effects that might exist between different levels of a covariate, we need to constrain the design matrix. Given the rank of the design, seven constraints will need to be imposed. Because the covariate effects are of interest, we first constrain the rater parameters by imposing sum-to-zero constraints over the raters within each of the four tables and also within each of the two sources. Notice that this only imposes five constraints and not six: Because the rater within-table constraints imply that the entire set of rater effects sums to zero, imposing an additional sum-to-zero constraint on the source A raters forces the source B rater effects to sum to zero as well. This leaves two additional constraints to be imposed over the three covariates, which may be accomplished by again summing to zero across any two of the three covariates, leaving the final covariate free. See Section 5 for additional discussion of model identifiability and alternative constraints.

One procedure often used to implement linear constraints across levels of a categorical variable contained in a design matrix is to exclude a column from the matrix that corresponds to one of the factor levels. For example, in the case at hand, one might impose a constraint across the rating tables by deleting the column corresponding to table 4. However, placing the multiple constraints on the rater effects (summing to zero both within tables and within source) creates ambiguity when the method of excluding columns is employed. If we were to delete the column corresponding to the first rater, this could be a constraint either on table 1 or on source A, and multiple solutions exist when transforming this version of a constrained design matrix back to the full design Y to expose the coefficients of all the factors.

An alternative strategy that avoids such ambiguities is to append additional rows to the design matrix to indicate which factors will be constrained (e.g., Kirk, 1982, p. 212); the constrained factors each receive a value of 1, and all other columns receive a 0. These rows are accompanied by corresponding values of zero appended in the vector of response variables. For example, to impose a sum-to-zero constraint across the rating tables, we append the following row vector to our original design matrix Y,

$$(\underbrace{0000000000000000000000000000}_{\text{rater}}\overbrace{1111}^{\text{table}}\;\underbrace{00}_{\text{source}}\;\overbrace{00}^{\text{env.}}),$$

and append a zero to our vectors of pseudorater bias ρ and variability ω parameters, which are our response variables in Equations 16 and 17, respectively. Additional rows are appended for each of our seven rating design constraints. We employ this strategy in the analysis below.

Note that an effect of defining the constraints by adding additional rows to the design matrix is that the constraint is implemented in Equation 22, not Equation 25, meaning that the constraint affects the means of the rating effects instead of the rating effects themselves.

With the design matrix constrained as described above, the HRMoc was fit including covariates to the GSE data using the random rating effects option, as described in Section 3. Prior distributions on the item parameters were set to $\beta_j \sim \text{Normal}(0, 10)$, $\ln \alpha_j \sim \text{Normal}(0, 10)$, and $\gamma_{1k} \sim \text{Normal}(0, 3)$. On both of the linear model components, we used the unweighted form with standard noninformative priors (e.g., Gelman et al., 1995) $f(\eta, \sigma_\rho^2 | Y) \propto \sigma_\rho^{-2}$ and $f(\tau, \sigma_\omega^2 | Y) \propto \sigma_\omega^{-2}$. We again programmed the MCMC in C++, with resulting parameter estimates based on 10,000 iterations of the HRMoc taken from five MCMC chains, using a burn-in of 8,000 iterations each. Posterior estimates of the item parameters were similar to those from the fixed effects example (see Table 3). Tables 6 through 8 display differences in the effects of each covariate,

TABLE 6

Posterior Median and 95% Equal-Tailed Credible Interval (CI) Estimates of the Covariate Effects of Rater Source Present in Part 1 of the Golden State Exam Image-Scoring Pilot Study as Fit Using the HRMoc With Random Rating Effects

		Bias			Variability	
		Median	95% CI		Median	95% CI
Source A	η_{S1}	−0.132	(−0.229, −0.035)	τ_{s1}	0.384	(0.355, 0.408)
Source B	η_{S2}	−0.168	(−0.273, −0.063)	τ_{s2}	0.438	(0.403, 0.471)
	$\eta_{S1} - \eta_{S2}$	0.037	(−0.039, 0.114)	$\tau_{s1} - \tau_{s2}$	−0.054	(−0.091, −0.020)

Note: Sum-to-zero constraints were imposed on the raters, both within tables and within source, and across rater table and environment. HRMoc = Hierarchical Rater Model for both objectively scored and constructed response items.

when the covariate of interest is the one left unconstrained. For all versions, the bias regression error variance has a posterior median estimate of $\sigma^2_\rho = 0.061$ with a 95% credible interval estimate of (0.033, 0.129), and the variability regression error variance has a posterior median estimate of $\sigma^2_\omega = 0.014$ with a 95% credible interval estimate of (0.006, 0.034).

Table 6 reveals no discernible difference in rating bias attributable to the source of the rater. The results for the variability attributable to rater source are a bit more interesting, with raters from source A demonstrating more consistency in applying the scoring guidelines. However, with a posterior median of − 0.054, this difference is not consequential at an operational level.

The estimates shown in Table 7 indicate a disparity among the rating tables, both bias and variability. There is a difference in bias between rating table 1 and rating table 2. In addition, the 95% equal-tailed interval estimate of the difference in bias between rating table 1 and rating table 4 (−0.182, 0.013) barely contains zero; alternate constructions of a 95% interval could exclude zero completely. Discernible differences in variability exist between rating tables 1 and 2 and between tables 1 and 4. These results indicate that the raters at table 1 are more harsh and less consistent in applying the scoring guidelines than are those raters at tables 2 and 4. Had these data applied to an actual assessment scoring session, instead of a special rating study, such results could be addressed with table 1's raters and table leader for correction in future ratings; estimating examinee proficiency θ_i with this version of the HRM would automatically account for such differences in the performance of the rating tables.

Table 8 displays a 95% equal-tailed interval estimate of the difference in bias of the ratings when scoring by the paper or image methods that barely contains zero; again, alternate 95% constructions could exclude zero completely, which would coincide with the results of the fixed rating effects example illustrated in Table 4. However, this interval estimate is wider than the corresponding interval in the fixed rating effects example (0.116 vs. 0.047), which may be attributable

TABLE 7
Posterior Median and 95% Equal-Tailed Credible Interval (CI) Estimates of the Covariate Effects of Table Assignment Present in Part 1 of the Golden State Exam Image-Scoring Pilot Study as Fit Using the HRMoc With Random Rating Effects

		Bias			Variability	
Covariate		Median	95% CI		Median	95% CI
Table 1	η_{t1}	−0.206	(−0.333, −0.090)	τ_{t1}	0.453	(0.409, 0.491)
Table 2	η_{t2}	−0.097	(−0.220, 0.022)	τ_{t2}	0.393	(0.347, 0.431)
Table 3	η_{t3}	−0.173	(−0.290, −0.055)	τ_{t3}	0.418	(0.373, 0.457)
Table 4	η_{t4}	−0.122	(−0.255, 0.015)	τ_{t4}	0.378	(0.330, 0.428)
	$\eta_{t1} - \eta_{t2}$	−0.110	(−0.205, −0.023)	$\tau_{t1} - \tau_{t2}$	0.060	(0.016, 0.109)
	$\eta_{t1} - \eta_{t3}$	−0.034	(−0.117, 0.045)	$\tau_{t1} - \tau_{t3}$	0.035	(−0.015, 0.086)
	$\eta_{t1} - \eta_{t4}$	−0.086	(−0.182, 0.013)	$\tau_{t1} - \tau_{t4}$	0.074	(0.018, 0.124)
	$\eta_{t2} - \eta_{t3}$	0.076	(−0.007, 0.161)	$\tau_{t2} - \tau_{t3}$	−0.026	(−0.082, 0.032)
	$\eta_{t2} - \eta_{t4}$	0.025	(−0.078, 0.123)	$\tau_{t2} - \tau_{t4}$	0.016	(−0.054, 0.065)
	$\eta_{t3} - \eta_{t4}$	−0.053	(−0.149, 0.050)	$\tau_{t3} - \tau_{t4}$	0.038	(−0.011, 0.089)

Note: Sum-to-zero constraints were imposed on the raters, both within tables and within source, and across rater source and environment. HRMoc = Hierarchical Rater Model for both objectively scored and constructed response items.

to the extra variability allowed in the random effects model. The small increase in the median difference from that of the previous example (−0.032 vs. −0.028) may be attributable to also accounting for table and source effects in this version of the model, Monte Carlo error, or both.

The 95% equal-tailed credible interval estimate of the difference in rating environment variability $\tau_{e1} - \tau_{e2}$ reports a difference that was undetected in the previous example. However, with the posterior median estimate indicating variability under image scoring only 0.046 higher, this difference is of little practical significance. Also note that the interval lower bound for both of these examples is essentially zero. As mentioned in Section 3, the rating variability was parameterized differently in the fixed and random rating effects versions. A parameterization that lends to a more direct comparison is, of course, possible; here, we chose to highlight the flexibility of the modeling options instead of the specific results of the example.

Model Comparison

Patz et al. (2002) provide evidence for the improved fit of the HRM over Facets for multiply rated constructed response data in education assessment, using both simulated and live data. Mariano (2002) provides a theoretical basis for the use of the HRM over Facets in modeling multiple ratings of individual responses. For the GSE data, following the format of Patz et al. (2002), we

TABLE 8

Posterior Median and 95% Equal-Tailed Credible Interval (CI) Estimates of the Covariate Effects of Image Versus Paper Scoring Present in Part 1 of the Golden State Exam Image-Scoring Pilot Study as Fit Using the HRMoc With Random Rating Effects

		Bias			Variability	
Covariate		Median	95% CI		Median	95% CI
Image	η_{e1}	−0.166	(−0.262, −0.071)	τ_{e1}	0.434	(0.402, 0.465)
Paper	η_{e2}	−0.134	(−0.236, −0.031)	τ_{e2}	0.388	(0.347, 0.421)
	$\eta_{e1} - \eta_{e2}$	−0.032	(−0.091, 0.025)	$\tau_{e1} - \tau_{e2}$	0.046	(0.002, 0.096)

Note: Sum-to-zero constraints were imposed on the raters, both within tables and within source, and across rater source and table. HRMoc = Hierarchical Rater Model for both objectively scored and constructed response items.

compare the fit of the HRM with and without covariates. To compare model fit, we employ the Bayesian Information Criterion (BIC; e.g., Kass & Raftery, 1995; Schwarz, 1978),

$$\text{BIC} = -2\ \ln\ (\text{maximum marginal likelihood}) + p\ \ln(\text{N}).$$

For this application, the marginal likelihood is taken after integrating over θ and ξ, *p* is the number of parameters in the model, and *N* is the number of examinees. The BIC serves as an approximation to the Bayes Factor (e.g., Kass & Raftery, 1995), which compares the marginal density of the data among candidate models but is difficult to calculate directly. A lower BIC indicates better model fit.

Table 9 displays the BIC for three different versions of the HRMoc fit to the GSE data: no covariates, a fixed effects environment covariate, and a random effects environment covariate. The HRMoc without the rating covariates offers the lowest BIC, indicating the best fit of the three. The random effects covariate version had the largest maximum marginal likelihood (i.e., smallest when multiplied by −2); however, the larger number of parameters inflated the penalty term, *p* ln(N), yielding a higher BIC. The higher BIC for the fixed effects version, as compared with the no covariates version, is also attributable to the penalty term.

The fixed effects version produced a posterior median estimate for the difference in rating bias between environments of −.028, with a 95% posterior interval upper bound of merely −0.007. A wider interval percentage width would cross zero (i.e., the MCMC chain for the difference in bias between environments does cross zero postconvergence). The BIC procedure is favoring model parsimony over these relatively tiny potential effects.

TABLE 9
Bayesian Information Criterion (BIC) Statistics for the HRMoc, With and Without Covariates, as Fit to Golden State Exam Image-Scoring Pilot Study Data

Model	BIC
HRMoc, no covariates	400, 157
HRMoc, environment covariate, fixed effects	400, 176
HRMoc, environment covariate, random effects	401, 160

Note: HRMoc = Hierarchical Rater Model for both objectively scored and constructed response items.

Section 5: Discussion

Section 3 describes a general format for the treatment of measurement covariates and two methods for including the covariates into the structure of three different Bayesian hierarchical models for rating data. Although the method for including covariates of the rating process is similar, these three models demonstrate sharp contrasts in their structure and assumptions, which will govern choice among them. As the facets framework treats all ratings as independent given the latent trait, it is best suited for the analysis of ratings of multiple responses to the same item. The HRM (Patz et al., 2002) and MMR (Verhelst & Verstralen, 2001) both treat the ratings as independent given the response to the item and thus are better suited for multiple ratings of the same response. Mariano (2002) details the error in estimating the latent trait that arises when applying the Facets model to multiple ratings of the same response. The HRM and MMR also differ by the form in which they characterize the quality of the response, with the MMR using a continuous scale and the HRM using a discrete scoring scale similar to that of the scoring rubric.

These three models are further distinguished by the form in which the effects of the rating process enter the model. Hierarchical facets and the MMR characterize the bias attributable to the rating process as a shift in the item difficulty attributable to the conditions under which it was scored (i.e., shifting the item response function). Meanwhile, the HRM treats the rating bias as a shift in the location of the conditional probability distribution of the ratings given the quality of response, which may be interpreted as placing an asymptotic bound (Junker & Patz, 1998) on the item response function.

The alternative forms of the rating effect have natural consequences for the interpretation of the included covariates. For example, when examining the rating environment covariate example in Section 4 under the HRM, we are examining how much more likely a response of a particular quality (i.e., ideal score) is to be rated in category k versus $k-1$, when rated on paper versus viewing a scanned image. Here, environment is a covariate affecting the rating process. If we were to examine the same environment covariate under the MMR, we would be examining how much more difficult it is for the examinee to score in category k versus $k-1$ on item j when being rated on paper versus a scanned image. In this scenario, environment is a covariate affecting item difficulty, a characteristic of the item. Thus, although the process by which the rating covariates enter the models is the same, the meaning that they take on is different, and this difference must be considered when choosing a model for the covariate and the associated inference we wish to draw.

In considering the actual inclusion of the rating covariates, Section 3 presents options for both fixed and random effects models. Both versions feature a rating design matrix that may need to be constrained for model identifiability. Ultimately, we are interested in understanding the differences, if any, between rating effects for different levels of a covariate. Although the design matrix may be less than full rank, these differences are well defined. The challenge is to expose the differences, either by employing reparameterization techniques to directly estimate the differences (e.g., Bock, 1975) or, as demonstrated in Section 4, by constraining the rating design matrix such that the individual covariate effects are identified. A choice to constrain the rating design matrix Y in a particular way imparts a particular definition to the individual rating covariate parameters unique to that constraint, and care must be taken if such parameters are interpreted individually.

Finally, we note the flexibility of the expanded HRM in this application. Whereas the GSE example employs a GPCM for the IRT level of the HRM, the general specification of the HRM in Equation 2 allows for the use of any parametric or nonparametric partial credit model (e.g., Hemker et al., 1996), including Samejima's (1969) GRM. Alternate parameterizations may also aide in accelerating MCMC convergence (e.g., Nandram & Chen, 1996). Portraying the contributions of the rating covariates in linear form is a familiar choice for understanding these effects, but certainly not the only possibility. Alternate adaptations using the rating effects structure of Section 3 are easy to envision and may prove more appropriate in certain cases.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: American College Testing.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*, 327–335.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with many-faceted Rasch models. *Journal of Educational Measurement*, *31*, 93–112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*, 56–70.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, *61*, 679–693.

Junker, B. W., & Patz, R. J. (1998, June). *The hierarchical rater model for rated test items*. Presented at the annual North American meeting of the Psychometric Society, Urbana-Champaign, IL.

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Monterey, CA: Brooks/Cole.

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

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

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*, 523–547.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer-Verlag.

Nandram, B., & Chen, M. (1996). Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation*, *54*, 129–144.

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, Department of Statistics, Carnegie Mellon University.

Patz, R. J., Awamleh, J., & Kelly, R. (1999). *Golden State Exam Image-Scoring Pilot Study* (Tech. rep.). Monterey, CA.

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

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

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.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Samejima, F. (1969). *Estimation of latent trait ability using a response pattern of graded scores*. Psychometrika Monograph, No. 17.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–549.

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

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

Authors

LOUIS T. MARIANO is a statistician at RAND Corporation, 1200 S. Hayes Street, Arlington, VA 22202; Lou_Mariano@rand.org. His research interests include Bayesian hierarchical models, with applications to student assessment and behavioral data.

BRIAN W. JUNKER is a professor in the Department of Statistics at Carnegie Mellon University, Pittsburgh, PA 15213; brian@stat.cmu.edu. His research interests include the statistical foundations of latent variable models for measurement, as well as applications of latent variable modeling in the design and analysis of standardized tests, cognitive diagnosis, and large-scale educational surveys such as the National Assessment of Educational Progress.

Manuscript received March 9, 2004
Accepted February 11, 2006