The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data[1]

May 16, 2000

Richard J. Patz	Brian W. Junker[2]	Matthew S. Johnson
CTB/McGraw-Hill	Department of Statistics	Department of Statistics
20 Ryan Ranch Road	Carnegie Mellon University	Carnegie Mellon University
Monterey CA 93940	Pittsburgh PA 15213	Pittsburgh PA 15213
rpatz@ctb.com	brian@stat.cmu.edu	masjohns@stat.cmu.edu

[1]This research was supported in part by National Science Foundation grant #DMS-97.05032 to Junker, and by a NAEP Secondary Data Analysis Grant, Award No. R902B980010, from the National Center for Educational Statistics to Patz, Mark Wilson, Junker and Machteld Hoskens. In addition to Wilson and Hoskens, we have benefitted from discussions with Darrell Bock, Lou Mariano, Bob Mislevy, Eiji Muraki and Carol Myford. We also wish to thank the Florida Department of Education for generously making available data from the study of rating modalities in the Florida Comprehensive Assessment Test. A preliminary version of the paper was presented at the Annual Meeting of the American Educational Research Association, April 1999, Montreal Canada.

[2]On leave at the Learning Research and Development Center, 3939 O'Hara Street, University of Pittsburgh, Pittsburgh PA 15260 USA.

Abstract

Single and multiple ratings of test items have become a stock component of standardized educational tests and surveys. For both formative and summative evaluation of raters, a number of multiple-read rating designs are now commonplace (Wilson and Hoskens, 1999), including designs with as many as six raters per item (e.g. Sykes, Heidorn and Lee, 1999). As digital image based distributed rating becomes commonplace, we expect the use of multiple raters as a routine part of test scoring to grow; increasing the number of raters also raises the possibility of improving the precision of examinee proficiency estimates. In this paper we develop Patz's (1996) hierarchical rater model (HRM) for polytomously scored item response data, and show how it can be used, for example, to scale examinees and items, to model aspects of consensus among raters, and to model individual rater severity and consistency effects. The HRM treats examinee responses to open-ended items as *unobserved* discrete variables, and it explicitly models the "proficiency" of raters in assigning accurate scores as well as the proficiency of examinees in providing correct responses. We show how the HRM "fits in" to the generalizability theory framework that has been the traditional analysis tool for rated item response data, and give some relationships between the HRM, the design effects correction of Bock, Brennan and Muraki (1999), and the rater bundles model of Wilson and Hoskens (1999). Using simulated data, we compare analyses using the conventional IRT Facets model for rating data (e.g. Linacre, 1989; Engelhard, 1994, 1996) and illustrate parameter recovery for the HRM. We also analyze data from a study of three different rating modalities intended to support a Grade 5 mathematics exam given in the State of Florida (Sykes, Heidorn and Lee, 1999) to show how the HRM can be used to identify individual raters of poor reliability or excessive severity, how standard errors of estimation of examinee scale scores are affected by multiple reads, and how the HRM scales up to rating designs involving large numbers of raters.

Keywords: Multiple ratings, generalizability, rater consensus, rater consistency, rater severity, latent response model, hierarchical Bayes modeling, item response theory, Markov chain Monte Carlo, MCMC.

1 Introduction

Rated responses to open-ended (or "constructed response") test items have become a standard part of the educational assessment landscape. The inclusion of open-ended items in testing programs is motivated primarily by validity concerns—these items are thought to be direct and "authentic" evaluations of competence and their inclusion is thought to have positive consequences for education (Messick, 1994)—and challenged most frequently on reliability concerns (e.g., Lukhele, Thissen and Wainer, 1994). Assessing the reliability of assessments including rated open-ended items requires replication of the scoring process, leading to multiple ratings of student work. Indeed, multiple ratings of test item responses have become a stock component of standardized educational tests and surveys, from the National Assessment of Educational Progress, to state-level tests aimed at student accountability, to tests developed by commercial test publishers. Examples of multiple ratings include "check-sets" consisting of papers rated in advance by experts and used to monitor rater accuracy during operational scoring, "blind double reads" used to monitor consistency of the scoring process, and "anchor papers" with responses from previous administrations used to monitor year-to-year consistency in the rating process (Wilson and Hoskens, 1999). The increasing use of imaging technology and computer-based scoring makes these multiple ratings designs easier to implement, more effective, and less expensive. With imaging technology as many as six or more truly independent ratings may be routinely gathered for monitoring, evaluation, or experimental purposes (see, for example, Sykes, Heidorn, and Lee, 1999).

Multiple-read designs open up a wide variety of data analysis possibilities—and challenges. Increasing the number of ratings per item immediately affords us the opportunity to improve the precision of estimating examinee proficiency levels, just as increasing the number of items can increase this precision. Multiple ratings also offer the opportunity to directly model aspects of consensus (or its lack) among groups of raters, and—as we shall see—to model bias and consistency *within individual* raters as well. With these possibilities come challenges: when using multiple reads to improve precision of examinee proficiency estimates, we must be sure that the statistical model is appropriately aggregating evidence from the set of ratings for each examinee or item. And in modeling individual rater effects and interactions among raters, we must attend carefully to the tradeoff between having sufficient parameters to model the effects of interest, and having so many parameters that we outstrip the information available in the data.

The purpose of this paper is to introduce and elaborate a statistical model for multiple ratings of test items, first proposed by Patz (1996), called the Hierarchical Rater Model (HRM). The HRM is one of several new approaches (see also Bock, Brennan and Muraki, 1999; Junker and Patz, 1998; Verhelst and Verstralen, 2000; and Wilson and Hoskens, 1999) to correcting a problem in how the Facets model within item response theory (Linacre, 1989) accumulates information in multiple ratings to estimate examinee proficiency. The HRM provides an appropriate way to combine information from multiple raters to learn about student performance, item parameters, etc., because it accounts for marginal dependence between different ratings of the same student work. It makes available tools for assessing the rater component of variability in IRT modeling of rating data corresponding to those available in traditional generalizability models for rating data. In addition, the HRM makes possible calibration and monitoring of individual rater effects such as rater severity and rater consistency.

In the following sections, we develop the HRM for polytomous data. In Section 2 we show some connections between the HRM and some other approaches to rated student performance data by analogy with a simple generalizability theory model; and in Section 3 we give the specific parameterization and estimation methods for the HRM that we use in this paper. In Section 4 we describe two interesting data sets: a data set simulated from the HRM itself to explore parameter recovery and similar issues under the Facets model and the HRM; and a real data set based on a study of multiple raters in the Florida state grade five mathematics assessment (Sykes, Heidorn and Lee, 1999). These data sets are analyzed in Section 5 to show how the

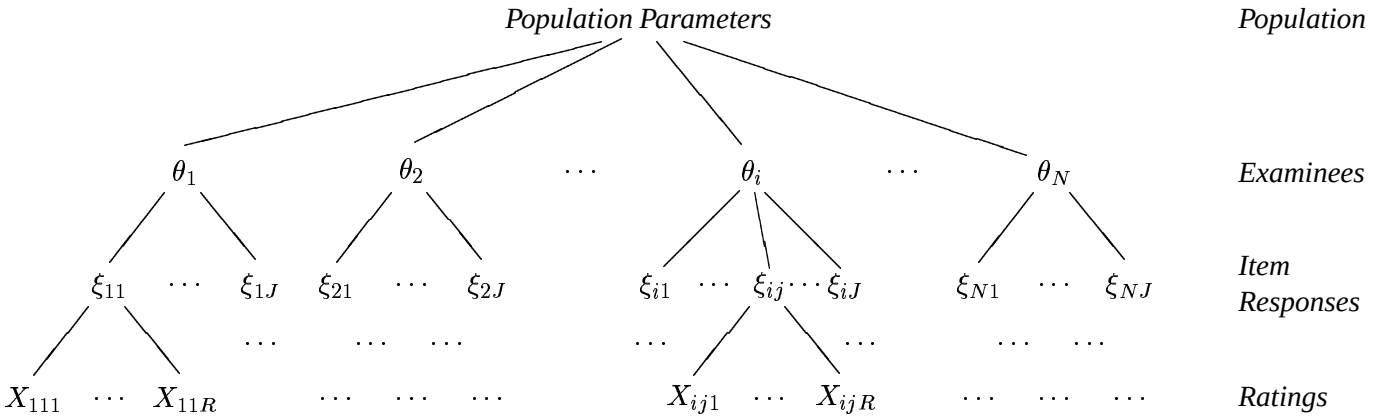

Figure 1: A hierarchical view of a simple generalizability theory model for a situation in which raters, items and examinees are completely crossed. Incompletely crossed and unbalanced designs are all modifications of this setup. The variance components, or facets of variability, are displayed as different levels in the tree. If the branches are modeled with the usual Normal-theory true-score models, one obtains a standard generalizability theory model. If the branches are modeled with IRT and discrete signal detection distributions, one obtains the Hierarchical Rater Model (HRM).

HRM can be used to identify individual raters of poor reliability or excessive severity, how standard errors of estimation of examinee scale scores are affected by multiple reads, and how the HRM scales up to rating designs involving large numbers of raters in loosely connected rating designs. We also briefly discuss overall model fit issues. Some extensions of the HRM, and speculations about the future of multiple rating designs and analyses, can be found in Section 6.

2 Some Models for Multiple Ratings of Test Items

Rater effects have been traditionally modeled and analyzed on the raw score scale using analysis of variance (ANOVA) or generalizability methodology (e.g., Brennan, 1992; Cronbach, Linn, Brennan, and Haertel, 1995; Koretz, Stecher, Klein, and McCaffrey, 1994). When greater measurement precision is required from a test containing rated responses of examinees to open-ended items, we may consider obtaining either 1) responses to additional items (i.e., a longer test with the same rating scheme), or 2) additional ratings per response (i.e., unchanged test length but more extensive ratings). The choice between the two (or of some combination of both) may be considered in a generalizability or variance components framework. By first estimating (in a "G-study") a rater variance component and an item variance component, we can then explore manipulations of the test design (in a "D-study") to make either component arbitrarily small.

Figure 1 presents a hierarchical view of a simple generalizability theory model for a situation in which R raters, J items, and N examinees are completely crossed; incompletely crossed and unbalanced designs are all modifications of this setup. The variance components, or facets of variability, are displayed at different levels in the tree, and labelled at right in Figure 1. The branches of the tree represent probability distributions that relate parameters or observations at each level. As usual in such models, nodes at one level of the tree are conditionally independent, given the "parent" variable(s) that they are connected to at the next higher

level of the tree. If we parameterize the branches in Figure 1 with the usual Normal-theory true score models

$$\begin{aligned} \theta_i &\sim i.i.d.\ N(\mu_{pop}, \sigma^2_\tau),\ i = 1, \ldots, N \\ \xi_{ij} &\sim i.i.d.\ N(\theta_i, \sigma^2_\epsilon),\ j = 1, \ldots, J,\ \text{for each } i \\ X_{ijr} &\sim i.i.d.\ N(\xi_{ij}, \sigma^2_\xi),\ r = 1, \ldots, R,\ \text{for each } i,\ j \end{aligned}$$

we obtain a connection between generalizability theory and hierarchical modeling, that has been noticed several times in the literature (e.g. Lord and Novick, 1968; Mislevy, Beaton, Kaplan and Sheehan, 1992). The expected a-posteriori (EAP) estimates under this Normal theory model of a student proficiency parameter θ, is always expressible as the weighted average of the relevant data mean and prior mean. The generalizability coefficients are the "data weights" in these weighted averages: the larger the generalizability coefficient, the less the data mean is shrunk toward the prior mean in the EAP estimate. For example (see Gelman, Carlin, Stern and Rubin, 1995, pp. 42ff; compare Bock, Brennan and Muraki, 1999), focusing on a a single branch connecting a θ_i to a ξ_{ij} we may compute the posterior mean of θ_i given ξ_{ij} as

$$E[\,\theta_i \mid \xi_{ij}\,] \;=\; \frac{\sigma^2_\epsilon}{\sigma^2_\tau + \sigma^2_\epsilon}\mu_{pop} + \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\epsilon}\xi_{ij} \;=\; (1-\rho)\mu_{pop} + \rho\xi_{ij},$$

where ρ is the usual per-item generalizability. For a set of branches connecting an examinee's θ_i to his/her ideal ratings $\xi_{i1}, \ldots, \xi_{iJ}$, the sufficient statistic for θ_i is $\overline{\xi}_{i\cdot} \sim N(\theta_i, \sigma^2_\epsilon/J)$, so that

$$E[\,\theta_i \mid \overline{\xi}_{i\cdot}\,] \;=\; (1-\rho_J)\mu_{pop} + \rho_J\overline{\xi}_{i\cdot},$$

where

$$\rho_J = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\epsilon/J}$$

is the usual test generalizability. And finally, using a similar analysis,

$$E[\,\theta_i \mid \overline{X}_{i\cdot\cdot}\,] \;=\; (1-\rho_{JR})\mu_{pop} + \rho_{JR}\overline{X}_{i\cdot\cdot},$$

where

$$\rho_{JR} = \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\epsilon/J + \sigma^2_\xi/JR}$$

is a generalizability coefficient for the information in all ratings of examinee i for estimating that examinee's θ_i. Thus, we reduce the item variance component by increasing test length, and we reduce the rater variance component by obtaining additional ratings (see for example, Brennan, 1992; and Cronbach et al. 1995).

This approach, however, has not been sufficiently developed for applications involving nonlinear transformations of raw test scores (Brennan, 1997), individual discrete item responses/ratings, etc., and so has limited ability to quantify the relationships between raters, individual items, and subjects. A currently popular (e.g. Engelhard, 1994, 1996; Myford and Mislevy, 1995; and Wilson and Wang, 1995) item response theory (IRT) based approach to modeling rater effects is the "Facets" model (Linacre, 1989), and has the same mathematical form as the Linear Logistic Test Model (LLTM; Scheiblechner, 1972; Fischer, 1973, 1983). IRT Facets models and their generalizations (e.g. Patz, Wilson and Hoskens, 1997) produce an ANOVA-like decomposition of effects for persons, items and raters on the logit scale, and thus appear to be directly analogous to generalizability analysis on the raw score scale. For example, a Facets model based on the partial credit model (PCM; Masters, 1982) provides additive fixed effects for rater severity, as follows:

$$\text{logit}\, P[X_{ijr} = k | \theta_i, X_{ijr} \in \{k, k-1\}] = \theta_i - \beta_j - \gamma_{jk} - \phi_r, \tag{1}$$

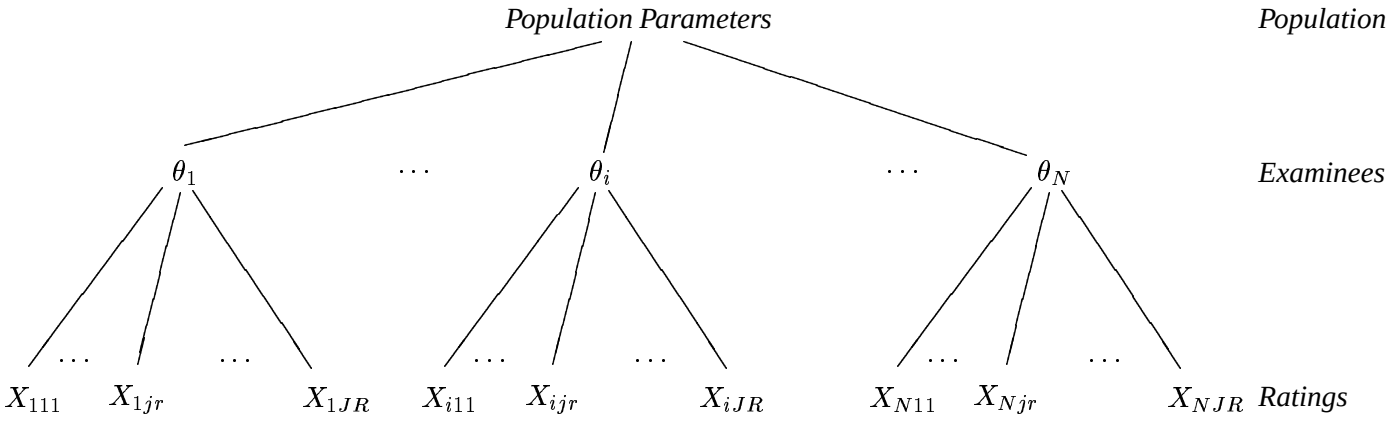

Figure 2: A hierarchical view of the standard Facets model corresponding to Figure 1. The layer of ideal rating variables ξ_{ij}, present in the generalizability and HRM setups, is missing in this model.

---

where X_{ijr} is the integer polytomous rating given to examinee i on item j by rater r, θ_i is the latent proficiency of the examinee, β_j is the item difficulty, γ_{jk} is the item step parameter, and ϕ_r is the rater severity.

The analogy between IRT Facets models and generalizability theory models breaks down in a fundamental and very significant way, however, when multiple measures are obtained from multiple facets. In IRT Facets models the likelihood for the rating data is typically constructed by multiplying together the probabilities displayed in (1) for all observed examinee $\times$ item $\times$ rater combination (e.g. the examples in Wu, Adams and Wilson, 1997). Figure 2 presents a hierarchical view of this model. Essentially, the IRT Facets model removes the layer of ideal rating variables ξ_{ij} in the middle of Figure 1 and make direct connections from each θ_i to the corresponding JR ratings X_{ijr}, typically modeled by (1) and the assumption of local independence between items and ratings of items.

The same argument that says that we can have arbitrarily precise proficiency estimation from a sufficiently long test (e.g., Birnbaum's, 1968, test information increases without bound) shows that as the number of raters per item increases, IRT Facets models appear to give infinitely precise measurement of the examinee's latent scale score θ_i, even though the examinee answers no more items (Patz, 1996; Junker and Patz, 1998). Wilson and Hoskens (1999) and Bock, Brennan and Muraki (1999) have also noted the excessively optimistic standard errors of estimation in IRT Facets models, and simulation work of Donoghue and Hombo (2000) has confirmed empirically that for as few as two raters per item the IRT Facets model can bias standard errors for θ_i well below what would be seen in the corresponding IRT model with no raters. Model fit studies (see Section 5.3 below; as well as Wilson and Hoskens, 1999) also suggest that the linear logistic form of the IRT Facets model may not track the variability in actual rating data as well as models that explicitly take into account the dependence between ratings due to their nesting within raters on the one hand and within examinees on the other.

The hierarchical rater model (HRM, Patz, 1996) corrects the problem of excessively optimistic standard errors in conventional IRT Facets models by breaking the data generation process down into two stages. In the first stage, the HRM posits ideal rating variables ξ_{ij}, describing examinee i's performance on item j, as *unobserved* per-item latent variables. This ideal rating variable may follow, for example, a standard PCM,

$$\text{logit}\, P[\xi_{ij} = \xi | \theta_i, X_{ijr} \in \{\xi, \xi - 1\}] = \theta_i - \beta_j - \gamma_{j\xi}, \tag{2}$$

or any other convenient IRT model. Conceptually, when we define a scoring rubric for an item, we are

		Observed Rating (k)				
		0	1	2	3	4
	0	p_{00r}	p_{01r}	p_{02r}	p_{03r}	p_{04r}
Ideal	1	p_{10r}	p_{11r}	p_{12r}	p_{13r}	p_{14r}
Rating	2	p_{20r}	p_{21r}	p_{22r}	p_{23r}	p_{24r}
(ξ)	3	p_{30r}	p_{31r}	p_{32r}	p_{33r}	p_{34r}
	4	p_{40r}	p_{41r}	p_{42r}	p_{43r}	p_{44r}

Table 1: The matrix of rating probabilities describing the signal detection process modeled in the HRM. $p_{\xi kr} \equiv P[$ Rater r rates $k \mid$ Ideal rating $\xi\ ]$ in each row of this matrix.

defining a map from the space of all possible student responses to an ordinal set of score points; and ξ_{ij} results from an ideal application of this mapping to examinee i's response to item j. Statistically, the ideal rating ξ_{ij} is a modeling device to capture dependence between multiple ratings of the same piece of student work (this is how the HRM corrects the IRT Facets model's underestimation of standard errors); it is related to the latent response variables of Maris (1995) within psychometrics, and to data-augmentation and missing data models (e.g. Tanner, 1996) in applied Bayesian statistics.

In the second stage, one or more raters produces a rating k for the item which may or may not be the same as the ideal rating category. This rating process is modeled as a discrete signal detection problem, using a matrix of rating probabilities $p_{\xi kr} \equiv P[$ Rater r rates $k \mid$ Ideal rating $\xi\ ]$ as displayed in Table 1. The rating probabilities $p_{\xi kr}$ in each row of this matrix must usually be constrained in some way to identify the model, since otherwise we must estimate on the order of K^2 parameters per rater when the items are rated into K categories each. It would be natural, for example, to posit a unimodal (unfolding) discrete distribution in each row of the table, with the location of the mode indicating rater severity and the spread of the distribution indicating rater (un-)reliability. Estimates of ξ_{ij} might then be viewed as a kind of consensus rating for examinee i's work on item j, among the raters who actually rated it. Non-square matrices of rating probabilities—perhaps reflecting a mismatch of the granularity of student performance with the scoring rubric—and interactions between raters and items or examinees can also be modeled in the second stage.

The hierarchical rater model (HRM) can be immediately seen as a reparametrization of the lower two sets of branches in Figure 1:

$$\begin{aligned} \theta_i &\sim \textit{i.i.d. } N(\mu_{pop}, \sigma_\tau^2),\ i = 1, \ldots, N, \text{ (as before);} \\ \xi_{ij} &\sim \text{an IRT model (e.g. PCM), } j = 1, \ldots, J, \text{ for each } i \\ X_{ijr} &\sim \text{the signal detection model in Table 1, } r = 1, \ldots, R, \text{ for each } i,\ j \end{aligned}$$

Thus, the HRM <u>is</u> the generalizability theory model in Figure 1, but with modifications to the distributions that link the facets of variability, to reflect the discrete nature of IRT rating data.

Some authors (e.g. Cronbach et al., 1995, p. 7) view observed examinee performance as developing along a linear continuum, so that ideally a continuous rating would be given to each performance. In this view, categorical or integer ratings are a practical necessity, but they result in a loss of information relative to the ideal continuous rating that is to be minimized, e.g. by using rubrics with many allowable score points, half points, etc. This could be accommodated in the HRM by making the ξ_{ij}'s be continuous latent variables, and the signal detection model a logit or probit response model; however in some ways this conception begs the question of what makes a good rubric. By contrast, we view the ξ_{ij}'s as the result of an ideal use of the given scoring rubric to map the space of all examinee responses—which need not be a continuum and may even have a rich qualitative structure—into a set of ordinal rating categories. This helps clarify what is good

or bad about a rubric (e.g., under- or over-specification), and what is good or bad about a rating (more or less severity, underuse of one or more categories, etc.). It may also help conceptually, to identify changes over time that are good (e.g., more complete specification of the rubric), and that are bad (e.g., increasing individual rater severity, increasing individual rater variability, etc.).

It is also valuable to compare the HRM approach to correcting the IRT Facets model with some other recent approaches. Verhelst and Verstralen (2000) have developed an approach to multiple ratings that is related to the HRM, in which a continuous latent "quality" variable plays a role similar to that of our ξ_{ij}'s, and raters are allowed to vary in severity only. The approach of Bock, Brennan and Muraki (1999) is to compare standard errors for estimating θ under generalizability theory models corresponding to Figures 1 and 2, and to compute a "design effects" correction that approximately corrects the conventional IRT Facets likelihood for omitting the ξ layer. The approach of Wilson and Hoskens (1999) is to build "rater bundles" analogous to Rosenbaum's (1988) item bundles, that explicitly model dependence between multiple reads of the same student work, by replacing the conditional independence model in each subtree of Figure 2 with an appropriate log-linear model. This rater bundle model (RBM) works quite well for modeling a few specific dependencies, between specific pairs of raters, or between specific raters and specific items. The HRM may be viewed as a kind of restriction of Wilson and Hoskens' RBM, that more readily scales up to larger numbers of ratings per item, because it treats raters as *a-priori* exchangeable.

Finally we note that the generalizability coefficients indicated at the beginning of this section do not have direct correspondents in the HRM, because as with most IRT-based models (and in contrast to models motivated from Normal distribution theory), location and scale parameters are tied together, so that the sizes of the variance components that make up the generalizability coefficients change as we move along the latent proficiency and ideal rating scales. However, the HRM makes available analogous tools, such as per-rater estimates of reliability, that in some ways improve our ability to monitor rater uncertainty and incorporate it appropriately into estimates on the latent proficiency scale.

3 Model Specification and Estimation Methods

3.1 The Hierarchical Rater Model

The hierarchical formulation of the HRM that we use begins at the data level with a matrix of rating probabilities (Table 1) for each rater. For the applications of the HRM in this paper we wish to parameterize the rating probabilities in each row of Table 1 so that the model is sensitive to each individual rater's severity and consistency. We do this by making the probabilities $p_{\xi kr} \equiv P[\text{ Rater } r \text{ rates } k \mid \text{Ideal rating } \xi\ ]$ in each row of this matrix proportional to a Normal density in k with location $\xi + \phi_r$ and scale ψ_r:

$$\begin{aligned} p_{\xi kr} &= P[X_{ijr} = k | \xi_{ij} = \xi] \propto \exp\left\{-\frac{1}{2\psi_r^2}[k - (\xi + \phi_r)]^2\right\} \\ & \qquad i = 1, \ldots, N;\ j = 1, \ldots, J;\ r = 1, \ldots, R \end{aligned} \tag{3}$$

This parameterization specifies maximum probability of response for category k when k is nearest to $\xi + \phi_r$. When $\phi_r = 0$, the maximum probability of response is for $k = \xi$, the ideal rating category; when $\phi_r < -0.5$, the maximum probability of response is for $k \leq \xi$, and when $\phi_r > 0.5$ the maximum probability of response is for $k \geq \xi$. Thus, the shift parameter ϕ_r measures individual *rater severity*, with $\phi_r < 0$ indicating greater severity, and $\phi_r > 0$ indicating greater leniency. Similarly, the scale parameter ψ_r controls how quickly the probabilities of response fall to zero as $|k - (\xi + \phi_r)|$ grows; and hence ψ_r is inversely related to individual *rater reliability*: the smaller ψ_r the greater the reliability or consistency of rater r. Thus, raters have established consensus with each other to the extent that both ϕ_r and ψ_r are close to zero across all

raters. Different items may have different numbers of ideal and observed rating categories, and we use this flexibility in analyzing the real data example below.

Since we do not have strong information about any of the rater parameters in our analyses below we took the prior distribution for the ϕ_r to be a relatively uninformative Normal distribution with mean 0 and variance 10, $N(0, 10)$, and for ψ_r a similar log-Normal density, $\log(\psi_r) \sim N(0, 10)$. In practice it is common for raters to qualify for live scoring by performing sufficiently well on examinee responses for which the "ideal rating" has been determined in advance. Information from these so-called "qualifying rounds" and "checksets" may be analyzed in terms of the rating probabilities in Table 1, and these preliminary analyses may support the use of more informative prior probabilities on the rater severity parameters ϕ_r and ψ_r.

We assume in this paper that the ideal ratings ξ_{ij} follow a PCM as in (2), namely

$$P\left[\xi_{ij} = \xi | \theta_i, \beta_j, \gamma_{j\xi}\right] = \frac{\exp\left\{\sum_{k=1}^{\xi}(\theta_i - \beta_j) - \gamma_{jk}\right\}}{\sum_{h=0}^{K}\exp\left\{\sum_{k=1}^{h}(\theta_i - \beta_j) - \gamma_{jk}\right\}} \tag{4}$$

where β_j is the item location (difficulty) parameter, and sums whose indices run from 1 to 0 are defined to be zero (and K need not be the same from one item to the next).

We take the population model for the latent proficiency to be

$$\theta_i \sim i.i.d.\ N(\mu, \sigma^2),\ i = 1, \ldots, N, \tag{5}$$

as indicated above in Section 2. Of course any other plausible population distribution for θ could be used as well. It is convenient to place the prior distribution $\frac{1}{\sigma^2} \sim$ Gamma (α, η) on σ^2, the population variance of θ; and to reflect little prior knowledge about σ^2 we have chosen a flat Gamma distribution with $\alpha = \eta = 1$ for our analyses.

As is usual with the PCM there is a location indeterminacy in that either μ or the β's must be constrained to get an identified model. For the simulated data example below we have taken $\mu = 0$ and allowed the β's to be free, sampled i.i.d. from a relatively uninformative $N(0, 10)$ prior distribution, to facilitate comparisons with an analogous IRT Facets model. For the real-data examples, we have allowed μ to be free with the same uninformative Normal prior, $N(0, 10)$, and forced the β's to satisfy a sum-to-zero constraint by computing $\beta_J = -\sum_1^{J-1}\beta_j$. This has the effect of making the posterior distribution for μ more peaked, since it depends on all the data, which makes our Markov Chain Monte Carlo (MCMC) estimation procedures somewhat more stable (see e.g. Gilks, Richardson and Spiegelhalter, 1995).

Recall that for a K-category item, $\beta_j + \gamma_{jk}$ are the locations of the $K - 1$ points at which adjacent category response curves cross; so only $K - 1$ γ_{jk}'s are formally included in the model. There is still a location indeterminacy in the γ's (add a constant to β_j and subtract the same constant from all the corresponding γ_{jk}). Thus we take the $K - 1$ item step parameters γ_{jk}'s for a K-category to be i.i.d. from a $N(0, 10)$ prior distribution, except that the last γ_{jk} for each item is a linear function of the others, according to a sum-to-zero constraint analogous to the β's. Thus only $K - 2$ item step parameters need to be estimated, for each item.

For incomplete designs, such as the real-data example we consider below, we include only those factors implied by the model specification (3), (4) and (5), that are relevant to the observed data in the likelihood. This has the effect of treating the data missing due to incompleteness of the design as being missing completely at random (MCAR; see for example Mislevy and Wu, 1996). The MCAR assumption is usually correct for data missing by design in straightforward survey and experimental designs where missingness is not informative about the parameters of interest. However, MCAR is not innocuous, and Wilson and

Hoskens (1999) indicate some "multiple-read" designs (such as formative read-behinds by expert raters) in which the presence or absence of a second rating can be quite informative about the quality of the first rating.

3.2 Markov Chain Monte Carlo Estimation

Estimation of the item difficulty parameters β_j, the rater shift (severity) ϕ_r, rater scale (unreliability) ψ_r, and hyper-parameters of the latent proficiency distribution, was carried out using a Markov Chain Monte Carlo (MCMC) algorithm. Given the ideal rating variables ξ_{ij}, MCMC estimation of the posterior of the partial credit model (PCM) is straightforward (see, for example, Patz and Junker, 1999a,b). Johnson, Cohen, and Junker (1999) implement MCMC estimation of the PCM model parameters in BUGS (Spiegelhalter, Thomas, Best and Gilks, 1996), and we extend their MCMC procedure for the PCM to the HRM by adding a step that draws ideal ratings from the relevant complete conditional distributions. The result was programmed in C++.

In the remainder of this subsection we indicate the complete conditional distributions needed to construct a MCMC estimation procedure for the specification of the HRM laid out in Section 3.1. In what follows, the (incomplete) matrix of all observed rating/item/person combinations is denoted $\mathcal{X}$; the notation $f(a|b,c,\ldots)$ is used generically to indicate the density or probability mass function of parameter a given parameters b, c, etc.; and underlining such as "$\underline{a}$" indicates a vector of parameters with similar names in the model.

We begin with the complete conditional distribution for each subject's ideal ratings on each item. The complete conditional posterior for ξ_{ij} $i = 1,\ldots,N;\ j = 1,\ldots,J$ is

$$f(\xi_{ij}|\underline{\theta},\underline{\beta},\underline{\gamma},\underline{\phi},\underline{\psi},\mathcal{X}) \propto \frac{\exp\left\{-\sum_{r\in R_{ij}} \frac{(x_{ijr}-\xi_{ij}-\phi_r)^2}{2\psi_r^2}\right\}}{\prod_{r\in R_{ij}}\sum_{k=0}^{K-1}\exp\left\{-\frac{(k-\xi_{ij}-\phi_r)^2}{2\psi_r^2}\right\}}\exp\left\{\xi_{ij}(\theta_i-\beta_j)-\gamma_{j\xi_{ij}}\right\}$$

where R_{ij} is the set of raters that graded the response of subject i to item j. Similarly the complete conditional posterior distribution for the rater shift parameters ϕ_r is

$$f(\phi_r|\psi_r,\underline{\xi},\mathcal{X}) \propto \frac{\exp\left\{-\sum_{(i,j)\in S_r} \frac{(x_{ijr}-\xi_{ij}-\phi_r)^2}{2\psi_r^2}\right\}}{\prod_{(i,j)\in S_r}\sum_{k=0}^{K-1}\exp\left\{-\frac{(k-\xi_{ij}-\phi_r)^2}{2\psi_r^2}\right\}} f_\Phi(\phi_r)$$

where S_r is the set of subject-item pairs that were rated by rater r, and $f_\Phi(\phi)$ is the prior distribution for each ϕ. The complete conditional posterior for the rater scale parameter ψ_r is almost identical, with $f_\Phi(\phi)$ replaced with $f_\Psi(\psi)$, the prior distribution for each ψ.

Conditional on the ideal ratings $\underline{\xi}$ the PCM parameters are independent of the data $\mathcal{X}$. The complete conditional posterior density for the item difficulty parameter β_j is

$$f(\beta_j|\underline{\xi},\underline{\gamma},\underline{\theta}) \propto \frac{e^{-\xi_{+j}\beta_j}}{\prod_{i=1}^{N}\sum_{k=0}^{K-1}\exp\{k(\theta_i-\beta_j)-\gamma_{jk}\}} f_B(\beta_j)$$

where $\xi_{+j} = \sum_{i=1}^{N}\xi_{ij}$. Similarly the complete conditional distribution for the item-step parameters is

$$f(\gamma_{jk}|\underline{\xi},\beta_j,\underline{\theta}) \propto \frac{e^{n_{jk}\gamma_{jk}}}{\prod_{i=1}^{N}\sum_{k=0}^{K-1}\exp\{k(\theta_i-\beta_j)-\gamma_{jk}\}} f_\Gamma(\gamma_{jk})$$

where $n_{jk} = \sum_{i=1}^{N} I_{\{\xi_{ij}=k\}}$, the number of respondents whose ideal rating category was k on item j. The conditional posterior distribution for the latent proficiency parameter θ_i is

$$f(\theta_i|\underline{\xi},\underline{\beta},\underline{\gamma}) \propto \frac{e^{\xi_{i+}\theta_i-\frac{\theta_i^2}{2\sigma^2}}}{\prod_{j=1}^{J}\sum_{k=0}^{K-1}\exp\{k(\theta_i-\beta_j)-\gamma_{jk}\}}$$

where $\xi_{i+} = \sum_{j=1}^{J} \xi_{ij}$. The complete conditional distribution for the latent proficiency variance is $\sigma^2|\underline{\theta} \sim$ *Inverse-Gamma* $\left(\alpha + N, \eta + \frac{\sum_{i=1}^{N} \theta_i^2}{2}\right)$; cf. e.g. Gelman et al. (1995, pp. 474–475).

The time it takes to complete a sufficiently long run for this MCMC algorithm depends on the number of subjects, items, and raters. The speed of the algorithm is approximately 5000 Markov chain steps per hour on a data set with two raters, eleven items, and 500 subjects, on a Hewlett Packard 9000/770 UNIX workstation; similar times could be expected on a fast PC. Correlations between parameters in the posterior distribution can be large, so to ensure adequate mixing it is necessary to perform moderately long MCMC simulations. In our analyses we have used 20000 Markov chain steps, after a burn-in period of 10000 steps. The algorithm is available in a C++ program from us (contact masjohns@stat.cmu.edu).

The Facets model can be estimated via MCMC using essentially the same software, by first removing the signal-detection (matrix of rating probabilities) level of the algorithm, and then using the PCM level of the algorithm to connect each rater $\times$ item combination directly to examinee proficiencies, as separate "virtual items" (as in Fischer and Ponocny, 1994) with additive effects for raters, item locations and item steps. Thus although the HRM and IRT Facets models are related, they are not nested in the likelihood-ratio testing sense, which complicates model comparisons in Section 5.3 below.

4 The Example Data Sets

We will explore the use of the HRM in two examples. In the first example, we examine data simulated from the HRM as described in Section 3.1, and compare the fit of the HRM itself to the fit of an analogous IRT Facets model. In the second example, we apply the HRM to data from a rating modality study conducted recently by CTB/McGraw-Hill for the Florida Comprehensive Assessment Test (FCAT).

4.1 Simulated Data

Using the HRM as described in Section 3, we simulated $R = 3$ ratings for each of $N = 500$ students on $J = 5$ test items, using five categories per item for both observed and ideal ratings. The examinee proficiencies θ_i, were drawn from a $N(0, 4)$ distribution, and the ideal rating variables ξ_{ij} followed the partial credit model (PCM) with item location (difficulty) parameters $\beta = (2, 1, 0, -1, -2)$, and item-step parameters γ_{jk} drawn from a $N(0, 1)$ distribution. Observed ratings were then simulated according to the matrices of rating probabilities as in Table 1 with rows modeled as in (3). The values of ϕ_r and ψ_r, used to simulate the three raters, $r = 1, 2, 3$, are given in the rightmost column of Table 4. These values were chosen to reflect realistic within-rater severity and reliability, at levels we found in our initial analyses of the FCAT data described below (Section 5.2.1). We will analyze the observed ratings (three per item) only, and treat the ideal ratings as missing data.

4.2 The Grade 5 Florida Mathematics Assessment

These data come from a field study conducted by CTB/McGraw-Hill in support of the Florida Comprehensive Assessment Test (FCAT), described by Sykes, Heidorn and Lee (1999). In the study, responses to open-ended field test items were scored by raters under several designs for assigning examinee responses to raters, using an computer image-based scoring system. Three assignment designs ("scoring modalities") were investigated, and each response was rated twice under each modality. In modality one raters were trained to score the complete set of open-ended items, and they would then score intact student books. In modality two, raters were assigned to score only a single open-ended item at a time. This modality was intended to mitigate the impact of rater severity differences on student scores. In the third scoring modality

Modality 1:	Rater	1	2	3	4	5	6	7			
Responses	Rated	187	2068	1859	2376	2651	1914	1199			
Modality 2:	Rater	9	10	11	12	13	15	16	17	18	19
Responses	Rated	486	412	423	573	530	449	517	537	426	70
	Rater	20	21	22	23	24	25	27	28	29	30
Responses	Rated	550	547	543	915	554	442	433	404	521	502
	Rater	31	32	33	36	37	38				
Responses	Rated	306	459	382	557	466	250				
Modality 3:	Rater	8	9	10	11	12	13	14	15	16	17
Responses	Rated	244	268	596	276	676	1168	800	792	776	1012
	Rater	20	21	23	24	25	26	28	33	34	35
Responses	Rated	1008	676	279	594	465	620	594	405	378	627

Table 2: Distribution of raters among modalities and item responses among raters in the Grade 5 Mathematics Test rating modality study.

Item 9

Score Combination	Rater Combination 12–13	12–16	13–16	Total
0–0	20	9	235	264
0–1	0	1	37	38
0–2	0	0	9	9
0–3	0	0	0	0
0–4	0	0	0	0
1–0	4	1	9	14
1–1	3	4	22	29
1–2	2	0	4	6
1–3	1	0	0	1
1–4	0	0	0	0
2–0	1	0	1	2
2–1	1	0	20	21
2–2	2	6	54	62
2–3	1	1	5	7
2–4	0	0	1	1
3–0	0	0	2	2
3–1	1	0	2	3
3–2	0	0	14	14
3–3	2	1	14	17
3–4	0	0	0	0
4–0	0	0	2	2
4–1	0	0	1	1
4–2	0	0	3	3
4–3	0	0	4	4
4–4	2	4	51	57
Total	40	27	490	557

Item 10

Score Combination	Rater Combination 12–15	12–36	15–36	Total
0–0	45	74	278	397
0–1	0	2	5	7
0–2	0	0	0	0
1–0	0	0	6	6
1–1	4	4	34	42
1–2	4	7	15	26
2–0	0	0	0	0
2–1	0	0	1	1
2–2	3	21	54	78
Total	56	108	393	557

Item 11

Score Combination	Rater Combination 12–37	12–38	36–37	37–38	Total
0–0	231	84	53	146	514
0–1	0	0	0	0	0
0–2	0	0	0	0	0
1–0	0	0	0	0	0
1–1	9	3	0	6	18
1–2	0	0	0	1	1
2–0	0	0	0	0	0
2–1	2	0	0	0	2
2–2	9	4	3	6	22
Total	251	91	56	159	557

Table 3: Cross-tabulations of modality two ratings by pairs of raters, for items 9, 10, and 11 of the Florida grade 5 Mathematics assessment.

raters were assigned to score blocks of items constituting roughly one third of the test. A complete description of the data set and the results of the scoring modality study may be found in Sykes, Heidorn and Lee (1999).

The data consist of scores assigned to open-ended responses from 557 examinees to 11 open-ended items—9 two-point items and 2 four-point items—by 38 raters. As described above each item was rated twice within each of three rating modalities, for a total of six ratings per item. The set of raters who rated in one modality was not necessarily distinct from the set who rated in another modality, as shown in Table 2. The study design is incomplete and unbalanced in the assignment of items to raters, as could be expected to be true of essentially all practical multiple rating situations.

Despite the fact that every piece of student work is rated six times in this study, the data can be extremely sparse, as illustrated by Table 3, which tabulates rating agreements and disagreements among pairs of all raters rating items 9, 10 and 11 in modality two. Considering the Item 9 subtable for example, we see that of the 40 occasions on which raters 12 and 13 both rated an Item 9 response, 20 times they agreed that the response should be rated 0, four times rater 12 rated the response as a 1 and rater 13 rated it as a 0, three times they agreed on a rating of 1, and so forth. For all three items, most of the action is in the low rating categories, indicating that these items are relatively difficult for the examinees.

5 Analyses with the HRM

Our analyses concentrate on the two data sets described in Section 4. In Section 5.1 we describe analyses of the simulated data using both a Facets model with additive effects for items and raters, and the HRM. This allows us to illustrate some qualities of using the HRM when it fits well, and also allows us to examine the Facets model fit when the data clearly contains more dependence than the Facets model can accommodate. In Section 5.2 we use the HRM to examine three subsets the Florida math assessment rater study data: in Section 5.2.1 we examine a small subset of the modality two ratings whose rater $\times$ items design is approximately balanced; in Section 5.2.2, we briefly consider all of the rated items in modality two, which is the same subset of this data that Wilson and Hoskens (1999) used to illustrate the Rater Bundle Model, and in Section 5.2.3 we extend the analysis to all of the rating data from all three rating modalities in the Florida rater study. In Section 5.3 we compare the fits of Facets and HRM models in the simulated and real data sets. Finally in Section 5.4 we illustrate the effects of increasing the number of items and the number of ratings on shrinking interval estimates of examinee proficiency scale scores, by comparing score estimates from various analyses of the Florida data.

By working with a fully Bayesian formulation of the model, as in Patz and Junker (1999a,b), we are able to find the posterior distributions of all the parameters of interest. Below we provide posterior medians (50^{th} posterior percentiles) as point estimates, and equal-tailed 95% credible interval (CI) estimates running from the 2.5^{th} posterior percentile to the 97.5^{th} posterior percentile, for each parameter of interest. We also compute approximate posterior modes (analogous to MLE's from an MML analysis) in the model fit comparison in Section 5.3 below. Because of heavy skewing and other deviations from symmetric unimodal shapes that sometimes occur in IRT posterior distributions, we do not report posterior means and SD's.

5.1 Simulated Data

We have applied both a Facets model and the HRM to the simulated data of 500 examinees scores by three raters in one of five categories on the five items discussed in Section 4.1. The purpose of this small simulation is to illustrate the behavior of the Facets model and HRM on data that has more dependence than the Facets model was designed to accommodate. A more extensive simulation study comparing the performance of the IRT Facets model and the HRM was reported by Donoghue and Hombo (2000).

Table 4 displays the item parameter estimates and proficiency distribution parameter estimates found using the two approaches. Table 5 gives latent proficiency estimates for five simulated examinees spread throughout the range of simulated proficiency values. All parameters for the two models admit comparison—in the sense that they are intended to be sensitive to the same effects on the same scale—*except for* the rater scale parameters ψ_r, which are only estimated in the HRM, and the rater severities ϕ_r. Rater severities are reported for both models for completeness, and to show that at least the direction of the severity estimates is consistent between models. However, the severity parameters are estimated on non-equivalent scales: the IRT Facets model estimates severity as an additive shift in the adjacent rating category logits in equation (1), and the HRM estimates severity as a shift in the modal rating category used by the rater, in equation (3).

We notice in Table 4 that the parameters used to simulate the data are recovered quite well by the HRM; all true parameter values are contained within the corresponding 95% CI. On the other hand, only two of the five item difficulty parameters (β_j's) and eight of the fifteen item step parameters (γ_{jk}'s) were contained in the IRT Facets CI's. In addition, it appears that the item difficulty parameter estimates (β_j's) found using the IRT Facets model have been shrunk toward zero. The item difficulty estimates for the Facets model are on average 0.2 units closer to zero than either the HRM estimates or the true values, with the shrinkage effect more pronounced for the more extreme items #1 and #5. The Facets model also underestimates the latent proficiency variance σ^2.

These estimation biases are to be expected; the IRT Facets model is being fitted to data that was generated from the HRM and therefore has structure that Facets was not designed to accommodate. However, the specific nature of the bias, excessive latent scale shrinkage, is interesting and important to think about. We believe that this shrinkage is exacerbated when individual rater reliability is poor (as it is with raters 1 and 2 in this simulation). When the individual rater reliabilities are low (rater scale parameters are large) then the "observed" ratings from an HRM simulation tend be in more middling categories, ameliorating extreme ideal ratings. The HRM model automatically discounts this since it estimates rater reliability directly along with everything else, but the IRT Facets model effectively assumes that all raters have a fixed standard reliability and thus takes these ameliorated ratings as evidence that the item wasn't so very extremely difficult or extremely easy. Patz, Junker and Johnson (1999) found even more extreme shrinkage effects under the Facets model when raters of even lower reliability were simulated. It is important to keep this behavior of the IRT Facets model in mind, if it is being fitted to data where we suspect low reliabilities of individual raters.

Although rater parameters are not directly comparable in the two models, it is interesting to note that under the HRM, rater scale (ϕ_r) and shift (ψ_r) parameters are estimated with little uncertainty for raters 1 and 2, but with rather high uncertainty for rater 3. We will return to this point, which we believe is also due to low rater reliability (high true ψ_3), in Section 5.2.1.

Table 5 gives posterior median and 95% credible interval estimates for five of the simulated examinees in this simulation. The simulated examinees displayed are located at the minimum, maximum, and quartiles of the simulated θ distribution. Except for the most extreme examinees, both models produce interval estimates that contain the true θ values. However, we note that the estimates of subject ability parameters obtained from the facets model are closer to zero (reflecting again latent scale shrinkage in the IRT Facets model due to rater unreliability), and have substantially narrower 95% intervals, than those from the HRM. Comparing the widths of the 95% CI's for parameters under the Facets model and HRM in Table 4, we also see that there is generally more uncertainty (wider interval estimates) in estimates under the HRM than under the Facets model.

The dramatically narrower interval estimates for estimating θ observed in Table 5 confirm the "double-counting" behavior of the IRT Facets model, relative to the HRM, first demonstrated by Junker and Patz (1998). The double-counting also has the effect of narrowing somewhat the interval estimates of the item parameters (Table 4) of the underlying PCM model. Since the data were simulated from the HRM itself,

Parameter			Facets Fit Median	Facets Fit 95% CI	HRM Fit Median	HRM Fit 95% CI	True Value
Proficiency Mean		μ	0^*	–	-0.13	$(-0.32, 0.05)$	0
Proficiency Variance		σ^2	3.32	$(2.82, 3.89)$	4.25	$(3.12, 5.40)$	4
	Item 1	β_1	-1.79	$(-1.99, -1.57)$	-1.96	$(-2.19, -1.69)$	-2
	Item 2	β_2	-0.98	$(-1.18, -0.78)$	-0.97	$(-1.12, -0.81)$	-1
	Item 3	β_3	-0.25	$(-0.46, -0.04)$	-0.16	$(-0.27, -0.05)$	0
	Item 4	β_4	0.68	$(0.48, 0.87)$	0.96	$(0.82, 1.10)$	1
	Item 5	β_5	1.74	$(1.52, 1.99)$	2.13	$(1.84, 2.37)$	2
Item 1	Step 1	γ_{11}	0.18	$(-0.01, 0.34)$	-0.37	$(-0.81, -0.00)$	-0.26
	Step 2	γ_{12}	-0.21	$(-0.51, 0.05)$	0.34	$(-0.15, 0.82)$	0.25
	Step 3	γ_{13}	-0.02	$(-0.33, 0.35)$	-0.26	$(-0.83, 0.25)$	0.02
Item 2	Step 1	γ_{21}	0.27	$(0.08, 0.44)$	-0.08	$(-0.48, 0.31)$	-0.21
	Step 2	γ_{22}	0.38	$(0.12, 0.62)$	0.66	$(0.22, 1.09)$	0.58
	Step 3	γ_{23}	0.48	$(0.27, 0.75)$	0.62	$(0.18, 1.02)$	0.77
Item 3	Step 1	γ_{31}	0.41	$(0.22, 0.58)$	0.27	$(-0.09, 0.60)$	0.34
	Step 2	γ_{32}	0.15	$(-0.07, 0.38)$	0.17	$(-0.20, 0.60)$	0.12
	Step 3	γ_{33}	0.01	$(-0.22, 0.23)$	-0.04	$(-0.43, 0.43)$	-0.07
Item 4	Step 1	γ_{41}	0.89	$(0.69, 1.07)$	1.03	$(0.66, 1.36)$	0.79
	Step 2	γ_{42}	-0.00	$(-0.21, 0.20)$	-0.14	$(-0.50, 0.19)$	0.03
	Step 3	γ_{43}	-0.48	$(-0.68, -0.26)$	-1.24	$(-1.74, -0.74)$	-1.31
Item 5	Step 1	γ_{51}	0.63	$(0.28, 0.97)$	-0.06	$(-0.74, 0.51)$	0.13
	Step 2	γ_{52}	1.56	$(1.22, 1.85)$	2.21	$(1.41, 2.84)$	2.05
	Step 3	γ_{53}	-0.30	$(-0.46, -0.11)$	-0.36	$(-0.68, -0.06)$	-0.36
Rater 1	Shift	ϕ_1	-0.05	$(-0.12, 0.02)$	-0.08	$(-0.11, -0.06)$	-0.07
	Scale	ψ_1			0.43	$(0.42, 0.44)$	0.43
Rater 2	Shift	ϕ_2	-0.23	$(-0.31, -0.16)$	-0.26	$(-0.29, -0.22)$	-0.25
	Scale	ψ_1			0.73	$(0.70, 0.75)$	0.72
Rater 3	Shift	ϕ_3	0^*	–	0.01	$(-0.40, 0.41)$	-0.02
	Scale	ψ_1			0.01	$(0.0005, 0.20)$	0.06

Table 4: MCMC parameter estimates for the additive Facets model and HRM, using data simulated from the HRM. The posterior median and 95% equal-tailed credible interval (CI) are given for each of the item parameters, the rater parameters and the prior standard deviation of the latent variable. Values marked with a star (*) were fixed at zero to identify the Facets model. True parameter values used to simulate the data are given in the rightmost column.

Simulated Proficiency		Facets Fit Median	Facets Fit 95% Interval	HRM Fit Median	HRM Fit 95% Interval	True Value
Minimum:	θ_{420}	-2.99	$(-4.11, -1.97)$	-3.96	$(-6.46, -2.28)$	-5.64
1^{st} Quartile:	θ_{395}	-1.78	$(-2.67, -0.91)$	-2.05	$(-3.57, -0.80)$	-1.53
Median:	θ_{286}	-0.69	$(-1.36, -0.10)$	-0.83	$(-2.03, 0.15)$	-0.13
3^{rd} Quartile:	θ_{368}	1.39	$(0.81, 2.06)$	1.32	$(0.35, 2.39)$	1.21
Maximum:	θ_{395}	2.72	$(1.98, 3.78)$	3.46	$(1.89, 6.12)$	6.03

Table 5: Estimated examinee proficiencies for the additive Facets model and HRM, using data simulated from the HRM. The simulated examinees displayed are located at the minimum, maximum, and quartiles of the simulated θ distribution. MCMC-based posterior median and 95% equal-tailed credible interval (CI) are given for each simulated examinee. True parameter values used to simulate the data are given in the rightmost column.

---

we know the greater uncertainty represented in the HRM item parameter estimates is more appropriate. The reduction in uncertainty in the IRT Facets parameter estimates is an artifact of that model's assumption, discussed in Section 2 and in Junker and Patz (1998), that response ratings are conditionally independent given examinee proficiencies θ_i. By contrast the HRM assumes that ratings are *dependent* given examinee proficiencies (they are conditionally independent only given the ideal ratings ξ_{ij}); and the extra dependence generally drives up uncertainty of parameter estimates. When similar dependence between ratings exists in real data, the HRM can be used to correct the excessively optimistic standard errors that the IRT Facets model gives. Wilson and Hoskens (1999) demonstrate a similar effect of ignoring dependence between ratings in the IRT Facets model, by showing that the model reliability for their rater bundle model (which also accommodates dependence between raters) was lower than the model reliability of the Facets model, in both simulated and real data.

5.2 The Grade 5 Florida Mathematics Assessment

5.2.1 Items 9, 10, and 11 of the Florida data

We first examine items 9, 10, and 11, scored in modality two in the Florida math assessment rater study, because this data extract exhibited fairly well-balanced rater $\times$ item design (though as illustrated in Table 3 the rater $\times$ examinee balance is not very good); each response was rated by two of seven raters. In Table 6 we report the median and 95% equal-tailed credible intervals (CI's) for HRM parameters for item difficulty, proficiency distribution and rater characteristics, and in Figures 3 and 4 we show histograms of the posterior distributions of the rater shift ϕ_r, and rater scale ψ_r parameters. (For brevity we only show item step parameter estimates for the full data analysis in Section 5.2.3 below).

The item difficulty parameter estimates ($\hat{\beta}$'s) show that item 11 is difficult in comparison to items 9 and 10; indeed item 11's $\hat{\beta}_{11} = 0.84$ is quite far from the latent distribution mean of $\hat{\mu} = -1.31$. The extreme difficulty of item 11 is already evident in the raw data (see Table 3): only 43 out of 557 students were given a non-zero score by at least one of the raters. More generally, we note that the mean $\hat{\mu} = -1.31$ of the latent proficiency distribution is low in comparison to all three item difficulty estimates, confirming the impression from Table 3 that all three items are difficult for the examinees.

Turning to the rater parameter estimates in Table 6 (see also Figures 3 and 4), we see that all seven rater

Parameter		Median	95% CI
Item 9	β_9	-0.53	$(-0.64, -0.41)$
Item 10	β_{10}	-0.31	$(-0.44, -0.19)$
Item 11	β_{11}	0.84	$(0.66, 1.02)$
Mean	μ	-1.31	$(-1.51, -1.15)$
Variance	σ^2	0.84	$(0.56, 1.24)$
Rater 12	Shift (ϕ_{12})	-0.27	$(-0.40, -0.18)$
	Scale (ψ_{12})	0.40	$(0.27, 0.44)$
Rater 13	Shift (ϕ_{13})	-0.07	$(-0.19, 0.05)$
	Scale (ψ_{13})	0.43	$(0.37, 0.49)$
Rater 16	Shift (ϕ_{16})	-0.25	$(-0.36, -0.14)$
	Scale (ψ_{16})	0.72	$(0.65, 0.79)$
Rater 15	Shift (ϕ_{15})	-0.22	$(-0.29, -0.14)$
	Scale (ψ_{15})	0.43	$(0.39, 0.46)$
Rater 36	Shift (ϕ_{36})	-0.01	$(-0.45, 0.44)$
	Scale (ψ_{36})	0.05	$(0.005, 0.26)$
Rater 37	Shift (ϕ_{36})	-0.36	$(-0.49, -0.17)$
	Scale (ψ_{37})	0.24	$(0.07, 0.35)$
Rater 38	Shift (ϕ_{36})	-0.02	$(-0.46, 0.44)$
	Scale (ψ_{38})	0.06	$(0.005, 0.26)$

Table 6: MCMC estimated posterior median and 95% equal-tailed credible intervals (CI's) for the HRM item difficulty, rater, and latent scale prior hyper-parameters based on 557 student responses to items 9, 10, and 11 of the Florida grade 5 Mathematics assessment.

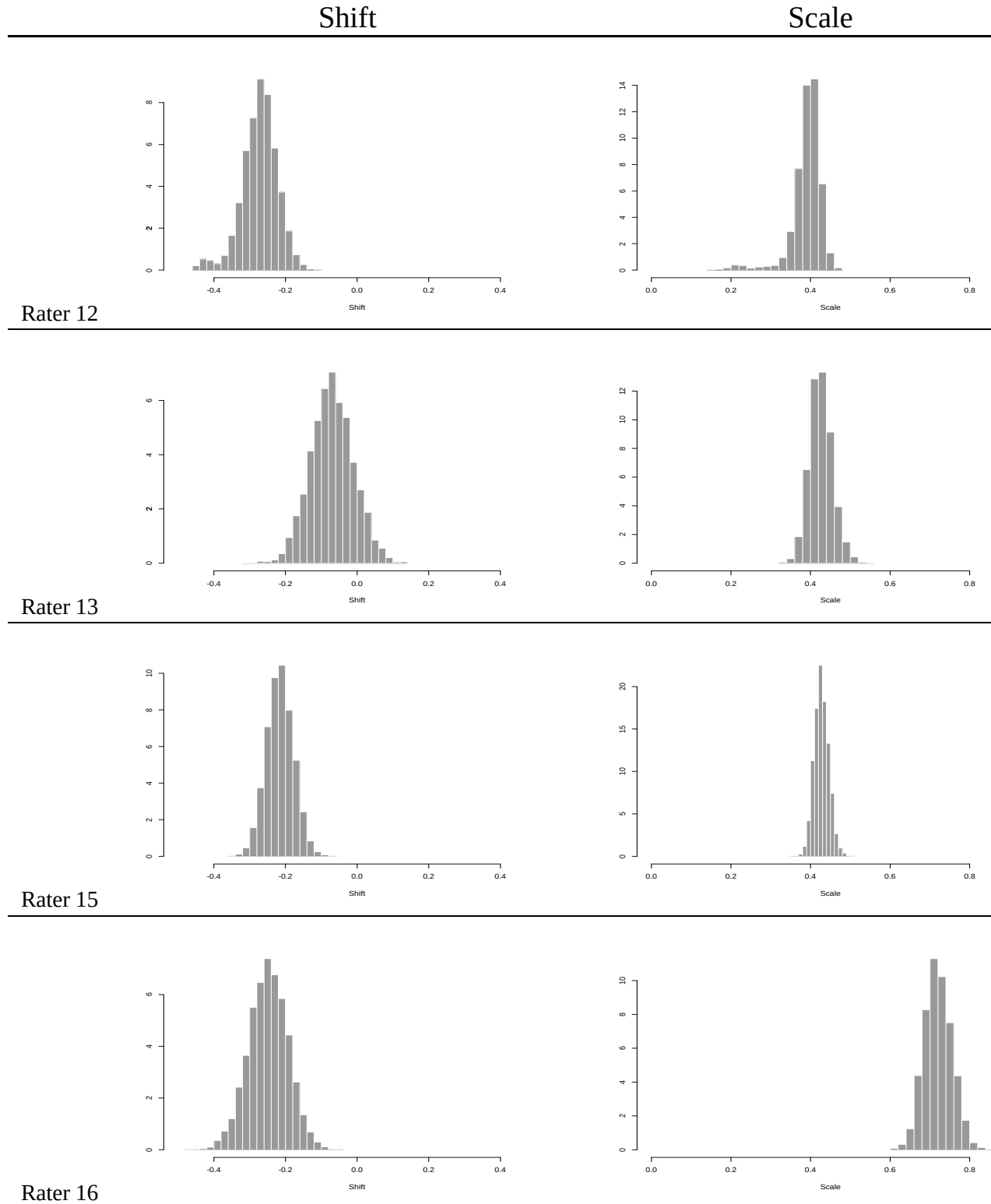

Figure 3: Histograms of the posterior distributions of rater shift and scale parameters based on 557 student responses to items 9, 10, and 11 of the Florida grade 5 Mathematics assessment.

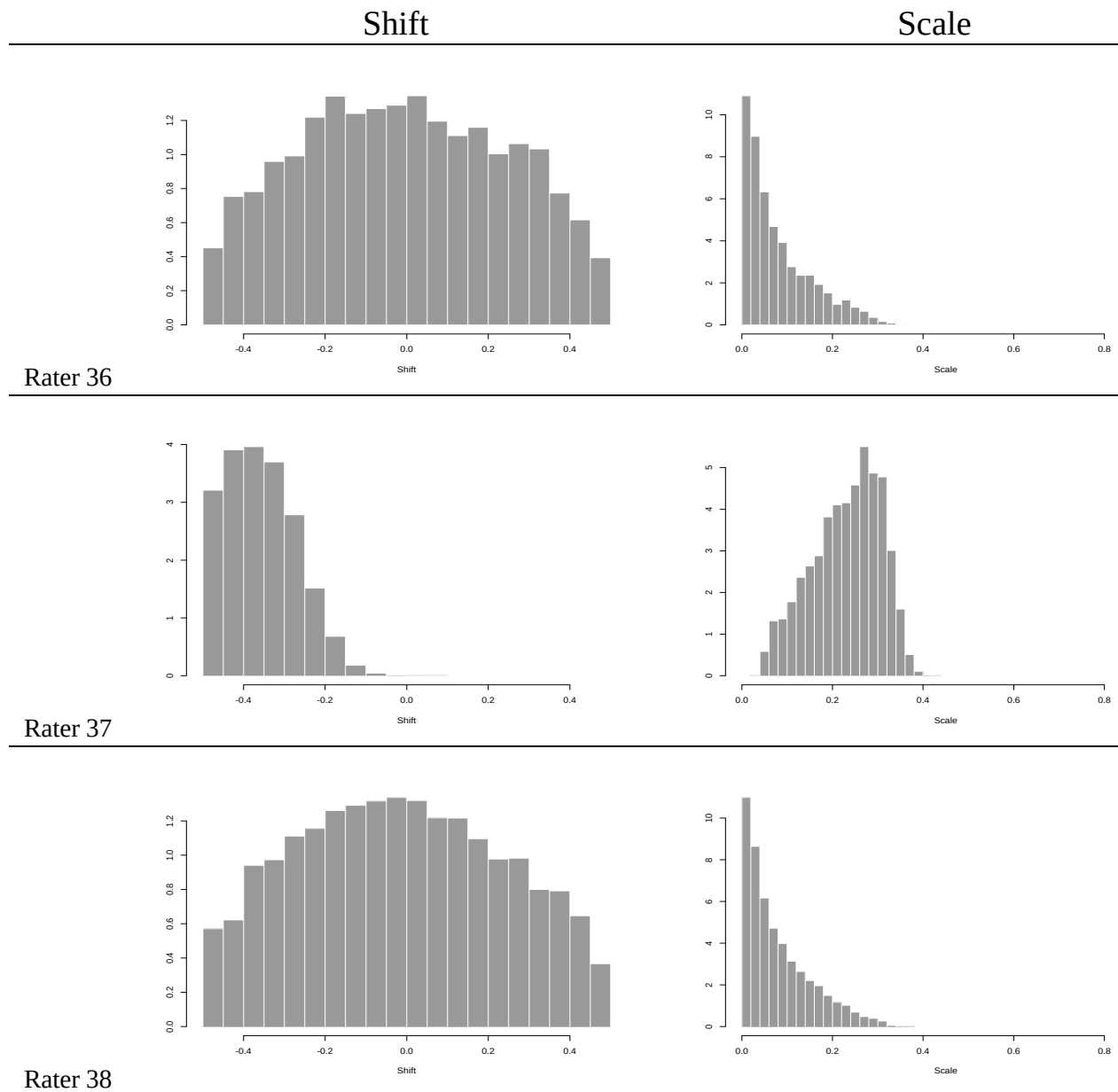

Figure 4: Histograms of the posterior distributions of rater shift and scale parameters based on 557 student responses to items 9, 10, and 11 of the Florida grade 5 Mathematics assessment.

shift parameters satisfy $|\hat{\phi}_r| < 0.5$. This suggests that the raters are approximately equally severe, in the sense that they are each more likely to score an item in the ideal rating category than any other category. If a rater's $|\hat{\phi}_r|$ were greater than 0.5 then the rater would be more likely to give the item a score other than the ideal rating category. Because the ideal rating category is inferred by the HRM from the pooled rating data, the ideal rating category is essentially a "consensus rating", and so the small rater shift parameters suggest that the raters agree on average about how each piece of examinee work should be rated. Despite this agreement on average, the seven raters are not equally reliable: we can see from Table 6 that raters 36 and 38 are quite reliable, with low rater scale estimates of $\hat{\psi}_{36} = 0.05$ and $\hat{\psi}_{38} = 0.06$, respectively. The other raters have rater scale estimates ranging from 0.24 to 0.72.

The rater scale estimate $\hat{\psi}_{16} = 0.72$ for rater 16 is a surprisingly large value, suggesting that this rater is inconsistent in assigning the same score to work of the same quality. The evidence presented in Section 5.1, as well as simulation results not shown here (see Patz, Junker and Johnson, 1999), suggest that this level of inconsistency or unreliability within raters can lead to severe shrinkage in the item difficulty and latent scale estimates in an IRT Facets model, as well as poor θ estimates under either HRM or Facets. Another way to compare the rating performance of these raters is to look at their estimated rating probability matrices under the model. For example, Figure 5 shows bar plots of $p_{\xi kr} = P[\text{ Rater } r \text{ rates category } k \mid \text{ ideal rating } \xi]$, for raters $r = 16$, 13, and 38, based on the point estimates of rater scale and shift in Table 6.

Finally, we see from Table 6—and even more vividly from Figures 3 and 4—that the reliable raters, 36 and 38, have poorly estimated shift parameters: the 95% CI's are quite large and the posterior distributions for ϕ_{36} and ϕ_{38} appear to be nearly uniform in the range -0.5 to $+0.5$. On the other hand the raters with poorer consistency (higher rater scale estimates) have tighter, clearly unimodal distributions for the shift parameters ϕ_r. We believe this is an artifact of using a continuous rating shift parameter ϕ_r to model discrete, whole unit shifts in the observed rating ξ_{ijr} away from the ideal rating category ξ_{ij}. Since raters 36 and 38 essentially always score items in the ideal rating category identified by the HRM, we know their shift parameters ϕ_r must be between -0.5 and $+0.5$; but since they do so with such high consistency, there is essentially no information in the data to determine where in this range their shift parameters ϕ lie. Apparently, nearly equivalent fits could be obtained by constraining the rater severity parameters to a few discrete values, such as 0, ± 1, ± 2, etc. This reduction of the parameter space might also lead to faster convergence of the MCMC estimation algorithm.

5.2.2 All assessment items rated in modality two

We now turn to an analysis of all eleven items graded in modality two. One of the additional items, item 2, was scored in one of the five response categories 0–4, and the remaining 7 were scored in three categories 0–2. A total of 26 raters graded at least one of the eleven items in modality two. The number of ratings per item was two, and the number of items rated by individual raters ranged from one to three, with the most common number of items per rater being one. The same data set was considered by Wilson and Hoskens (1999).

The item difficulty and latent distribution parameter estimates for the partial credit model (PCM) underlying the ideal ratings appear in Table 7, and the rater shift and scale estimates are contained in Table 8. The point estimates for the item difficulties agree quite well with the difficulty estimates under the Facets model as reported by Wilson and Hoskens (1999), after a linear transformation to adjust for different latent proficiency means and variances in the two analyses. The items and raters analyzed in the smaller, more balanced data set in Section 5.2.1, are indicated by asterisks in these tables. Comparing with Table 6 we see very little difference in the estimated rater parameters, and small differences in the item difficulty parameters that seem mostly to be due to the different effects that the sum-to-zero constraint has on them in the model for three items vs. eleven items.

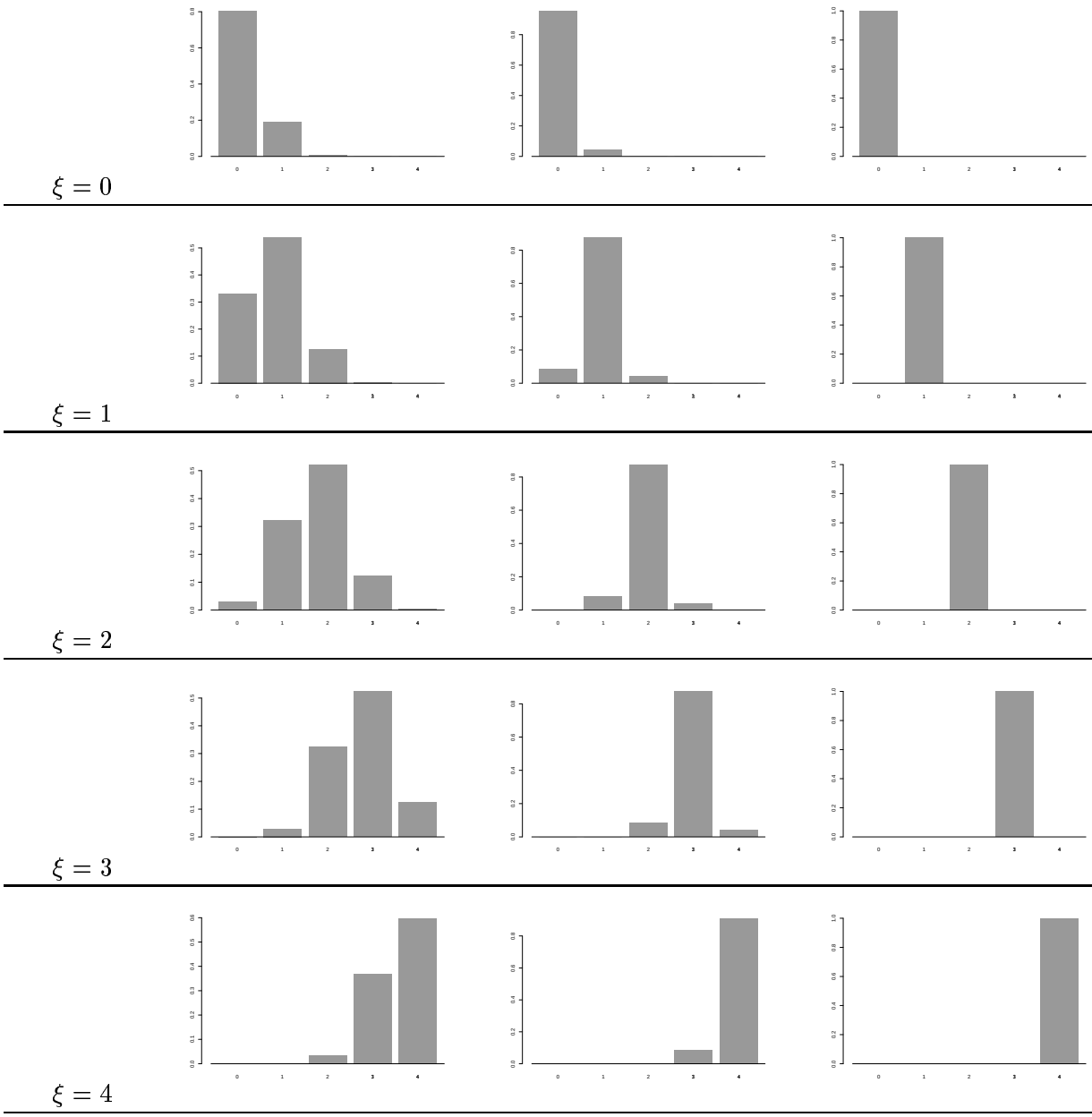

Figure 5: Barplots of the estimated category rating probabilities $p_{\xi kr} = P[$ Rater r rates category $k \mid$ ideal rating $\xi]$, for raters 16, 13 and 38, based on 557 student responses to items 9, 10, and 11 of the Florida grade 5 Mathematics assessment. Rater shift (severity) and scale (unreliability) parameter estimates may be found in Table 6.

Parameter	Median	95% CI
Item 1	-0.06	$(-0.19, 0.07)$
Item 2	-0.25	$(-0.49, 0.12)$
Item 3	0.62	$(0.43, 0.84)$
Item 4	0.08	$(-0.14, 0.30)$
Item 5	-0.68	$(-0.81, -0.56)$
Item 6	0.29	$(0.15, 0.43)$
Item 7	-0.39	$(-0.50, -0.27)$
Item 8	-0.27	$(-0.41, -0.13)$
Item 9*	-0.31	$(-0.41, -0.21)$
Item 10*	-0.08	$(-0.21, 0.04)$
Item 11*	1.03	$(0.83, 1.25)$
Variance	0.73	$(0.61, 0.88)$
Mean	-1.05	$(-1.15, -0.96)$

Table 7: MCMC estimated posterior median and 95% equal-tailed credible intervals (CI's) for the item difficulties, and latent scale mean and variance parameters, of eleven items, based on two ratings in modality two, for each of 557 students responses to eleven items on the Florida Grade 5 Mathematics assessment. Items analyzed in the smaller extract in Section 5.2.1 are indicated by asterisks.

Judging from the PCM estimates in Table 7 we find that items 3, 6, and 11 are the most difficult items; referring to the raw data, these items have, respectively, 422, 443, and 514 students out of 557 who were assigned a score of 0 by both raters. Items 5, 7, 8, and 9 appear to be the least difficult of the Mathematics exam items. For item 5, the easiest of these items as determined by the PCM difficulty parameter estimates, both raters assigned the highest possible score to 140 of the 557 students that they both scored. As noted in the analysis of item 9, 10 and 11 in Section 5.2.1, the mean of the latent scale μ, is found to be quite low, relative to the difficulty of the items. We also note that, as expected, the confidence interval for the latent scale mean is smaller when using all eleven items than when using only the last three items.

Finally we examine the performance of the 26 raters in Modality Two. All raters, with the exception of rater 19, appear to be in agreement with one another, in the sense that their rater shift parameters ϕ_r are all between -0.5 and $+0.5$: they are all more likely to give the student a score equal to the ideal rating category than any other score. The median of the posterior distribution for Rater 19's shift parameter $\phi_{19} = -0.60$. At this value of the shift parameter Rater 19 becomes more likely to score the students' responses one category lower than the ideal rating category.

Rater 11, Rater 36, and Rater 38 are found to be very reliable; their rater scale parameter (ψ_r) estimates are 0.09, 0.05 and 0.06, respectively; this means these raters essentially always score examinee work in the category k nearest to $\xi + \phi$. On the other hand, Raters 16, 17 and Rater 19 have scale estimates that seem high in comparison to the others, again in the same range as our simulated HRM data analyzed in Section 5.1. This suggests that the individual reliability or consistency of these raters is poor: such a rater would be less likely to give consistent ratings on separate reads of equivalent student work.

Rater	Parameter	Median	95% CI
Rater 9	Shift	-0.03	$(-0.15, 0.14)$
	Scale	0.36	$(0.28, 0.40)$
Rater 11	Shift	-0.06	$(-0.46, 0.42)$
	Scale	0.09	$(0.01, 0.35)$
Rater 10	Shift	-0.10	$(-0.26, -0.01)$
	Scale	0.38	$(0.27, 0.41)$
Rater 17	Shift	-0.29	$(-0.44, -0.15)$
	Scale	0.78	$(0.69, 0.88)$
Rater 19	Shift	-0.60	$(-1.16, -0.26)$
	Scale	0.76	$(0.39, 1.20)$
Rater 18	Shift	0.22	$(0.09, 0.36)$
	Scale	0.56	$(0.46, 0.69)$
Rater 20	Shift	-0.22	$(-0.32, -0.10)$
	Scale	0.40	$(0.34, 0.46)$
Rater 21	Shift	-0.44	$(-0.50, -0.32)$
	Scale	0.24	$(0.05, 0.45)$
Rater 22	Shift	0.14	$(-0.02, 0.45)$
	Scale	0.31	$(0.12, 0.38)$
Rater 23	Shift	-0.06	$(-0.13, 0.01)$
	Scale	0.37	$(0.34, 0.39)$
Rater 24	Shift	-0.22	$(-0.29, -0.15)$
	Scale	0.41	$(0.35, 0.45)$
Rater 25	Shift	-0.07	$(-0.16, 0.02)$
	Scale	0.38	$(0.34, 0.42)$
Rater 27	Shift	-0.20	$(-0.46, -0.07)$
	Scale	0.36	$(0.13, 0.41)$
Rater 28	Shift	-0.09	$(-0.48, 0.39)$
	Scale	0.19	$(0.01, 0.36)$
Rater 29	Shift	-0.10	$(-0.18, -0.02)$
	Scale	0.37	$(0.34, 0.42)$
Rater 31	Shift	-0.07	$(-0.37, 0.14)$
	Scale	0.33	$(0.02, 0.38)$
Rater 30	Shift	0.23	$(0.01, 0.47)$
	Scale	0.27	$(0.09, 0.37)$
Rater 32	Shift	0.18	$(0.01, 0.36)$
	Scale	0.34	$(0.22, 0.42)$
Rater 33	Shift	0.04	$(-0.07, 0.17)$
	Scale	0.37	$(0.32, 0.41)$
Rater 12*	Shift	-0.26	$(-0.35, -0.18)$
	Scale	0.40	$(0.33, 0.44)$
Rater 13*	Shift	-0.09	$(-0.19, 0.02)$
	Scale	0.42	$(0.38, 0.48)$
Rater 16*	Shift	-0.26	$(-0.36, -0.15)$
	Scale	0.71	$(0.66, 0.79)$
Rater 15*	Shift	-0.22	$(-0.29, -0.15)$
	Scale	0.43	$(0.39, 0.46)$
Rater 36*	Shift	-0.01	$(-0.45, 0.43)$
	Scale	0.05	$(0.003, 0.27)$
Rater 37*	Shift	-0.36	$(-0.49, -0.17)$
	Scale	0.24	$(0.06, 0.35)$
Rater 38*	Shift	-0.05	$(-0.45, 0.44)$
	Scale	0.06	$(0.007, 0.25)$

Table 8: MCMC estimated posterior median and 95% equal-tailed credible intervals (CI's) for the HRM rater parameters based on two ratings in modality two for each of 557 student responses to eleven items on the Florida Grade 5 Mathematics assessment. Raters analyzed in the smaller extract in Section 5.2.1 are indicated by asterisks.

5.2.3 The full Florida data set

To illustrate how the model scales up to a large number of raters, and a large number of ratings per item, we estimated the HRM using the full Florida data set. Recall that the set of eleven items was rated twice in each of three rating modalities, for a total of six ratings per item using a pool of 38 raters.

As indicated in Section 4.2, the following raters rated in only one modality

- Raters 1–7 are unique to Modality 1 (intact test booklets);
- Raters 18, 19, 22, 27, 29, 30, 31, 32, 36, 37, and 38 are unique to Modality 2 (single items)
- Raters 8, 14, 26, 34, and 35 are unique to Modality 3 (blocks of 1/3 of the items)

and the remaining raters rated (different items) in both modalities 2 and 3. To the extent that only modality affects differences in rater characteristics between modalities, we can examine these groups of raters for possible differences between rating modalities. Table 9 contains the estimated HRM rater parameters for all 38 raters; Table 11 contains estimated item parameters from the full-data fit.

In Table 9, the modalities of those raters who rated in one modality only are identified in bold-face type. In addition, the raters from our initial analysis of items 9, 10, and 11 in modality two only, are indicated again by asterisks. Comparing the starred raters in Table 9 with the parameter estimates in Table 6, and with the starred entries in Table 8, we see that estimates of these raters' parameters are all fairly stable across the three fits, except for rater 36. This rater's shift parameter stays fairly stable, moving only from -0.01 to $+0.05$, but the rater's original scale estimate of 0.05 is now replaced by an estimate of 0.37. This suggests a fair amount of disagreement of rater 36, who only rates in modality two, with raters in modalities 1 and 3, but no real trend in the disagreements toward severity or leniency of rater 36, relative to the other raters.

Since rater 38's shift estimate of -0.02 and scale estimate of 0.06 are fairly similar to the initial estimate of rater 36, and indicate excellent reliability and lack of severity bias, we might wonder how this can happen. A quick look at raters 36 and 38 as they appear in Tables 2 and 3 shows that (a) rater 36 disagreed fairly regularly on item 10 with raters 12 and 15, who also rated in modality 3; (b) rater 36 can be seen to agree in all but 6 or 8 cases with rater 38 on item 11, by linking through rater 37; and finally (c) Rater 38 only rated a relatively small number of responses, 250, compared with most other raters. It thus appears that rater 36 did disagree with many other raters, on other items than item 11, in the full data set; on the other hand rater 38 rated so few examinees/item combinations that there was relatively little opportunity to compare his/her ratings with other raters' ratings.

To summarize the rating characteristics of raters identified in Table 9 who only graded in one modality, we compute mean severity $\overline{\phi}$ and mean unreliability $\overline{\psi}$ for each group, and for all 38 raters as a single group. These are shown in Table 10. Comparing the mean severity for raters nested within each modality suggests that raters in modality 1 were somewhat more lenient on average than those in the other two modalities, and the average reliability of the raters was lowest in modality 2. These findings broadly parallel the findings in Sykes, Heidorn and Lee (1999). In addition, we are able to refer to Table 9 to explore how individual rater severity and (un-)reliability contribute to this overall picture.

The item parameter estimates for the PCM layer of the HRM are listed in Table 11. The item difficulty parameter estimates (β's) in this table are quite similar to those of Table 7, which were estimated from the full modality two data only; the primary difference is that the item difficulties are somewhat more spread out in Table 11, compared to Table 7. All of the item difficulties are above the estimated latent proficiency mean, suggesting that these items were relatively difficult for the examinees. This finding was also suggested by our earlier analyses, and is consistent with results reported by Sykes, Heidorn and Lee (1999).

In addition we have listed the estimated item step parameters for the PCM layer, for each of the items; item step parameters not listed her may be obtained from these via the relevant sum-to-zero constraint (recall

Rater	Param.	Median	95% CI
Rater 1	Shift	-0.10	$(-0.21, 0.01)$
Modality 1	Scale	0.42	$(0.38, 0.47)$
Rater 2	Shift	0.00	$(-0.03, 0.03)$
Modality 1	Scale	0.48	$(0.46, 0.50)$
Rater 3	Shift	-0.13	$(-0.16, -0.09)$
Modality 1	Scale	0.43	$(0.41, 0.44)$
Rater 4	Shift	-0.07	$(-0.10, -0.04)$
Modality 1	Scale	0.51	$(0.49, 0.53)$
Rater 5	Shift	-0.09	$(-0.12, -0.06)$
Modality 1	Scale	0.44	$(0.43, 0.45)$
Rater 6	Shift	-0.05	$(-0.09, -0.02)$
Modality 1	Scale	0.49	$(0.47, 0.51)$
Rater 7	Shift	-0.09	$(-0.13, -0.04)$
Modality 1	Scale	0.43	$(0.41, 0.45)$
Rater 8	Shift	-0.27	$(-0.43, -0.18)$
Modality 3	Scale	0.42	$(0.23, 0.48)$
Rater 9	Shift	-0.22	$(-0.27, -0.17)$
	Scale	0.46	$(0.43, 0.48)$
Rater 10	Shift	-0.04	$(-0.09, 0.01)$
	Scale	0.43	$(0.41, 0.45)$
Rater 11	Shift	-0.02	$(-0.09, 0.04)$
	Scale	0.37	$(0.34, 0.39)$
Rater 12*	Shift	-0.22	$(-0.26, -0.17)$
	Scale	0.47	$(0.45, 0.50)$
Rater 13*	Shift	-0.08	$(-0.12, -0.05)$
	Scale	0.52	$(0.50, 0.54)$
Rater 14	Shift	-0.12	$(-0.17, -0.07)$
Modality 3	Scale	0.53	$(0.51, 0.57)$
Rater 15*	Shift	-0.29	$(-0.33, -0.25)$
	Scale	0.47	$(0.44, 0.49)$
Rater 16*	Shift	-0.22	$(-0.26, -0.18)$
	Scale	0.56	$(0.53, 0.58)$
Rater 17	Shift	-0.30	$(-0.34, -0.27)$
	Scale	0.52	$(0.49, 0.54)$
Rater 18	Shift	-0.01	$(-0.10, 0.06)$
Modality 2	Scale	0.70	$(0.65, 0.76)$
Rater 19	Shift	-0.64	$(-0.88, -0.46)$
Modality 2	Scale	0.66	$(0.53, 0.84)$

Rater	Param.	Median	95% CI
Rater 20	Shift	-0.05	$(-0.09, 0.00)$
	Scale	0.37	$(0.36, 0.39)$
Rater 21	Shift	-0.17	$(-0.22, -0.13)$
	Scale	0.43	$(0.41, 0.45)$
Rater 22	Shift	-0.03	$(-0.10, 0.04)$
Modality 2	Scale	0.35	$(0.33, 0.38)$
Rater 23	Shift	-0.13	$(-0.17, -0.09)$
	Scale	0.40	$(0.39, 0.42)$
Rater 24	Shift	-0.29	$(-0.33, -0.34)$
	Scale	0.48	$(0.45, 0.51)$
Rater 25	Shift	-0.15	$(-0.19, -0.10)$
	Scale	0.48	$(0.45, 0.50)$
Rater 26	Shift	-0.10	$(-0.16, -0.04)$
Modality 3	Scale	0.40	$(0.37, 0.43)$
Rater 27	Shift	-0.20	$(-0.28, -0.11)$
Modality 2	Scale	0.39	$(0.35, 0.43)$
Rater 28	Shift	-0.28	$(-0.34, -0.22)$
	Scale	0.48	$(0.45, 0.51)$
Rater 29	Shift	-0.15	$(-0.22, -0.07)$
Modality 2	Scale	0.41	$(0.38, 0.44)$
Rater 30	Shift	-0.07	$(-0.14, 0.00)$
Modality 2	Scale	0.37	$(0.35, 0.43)$
Rater 31	Shift	-0.10	$(-0.22, 0.02)$
Modality 2	Scale	0.35	$(0.30, 0.40)$
Rater 32	Shift	0.03	$(-0.04, 0.10)$
Modality 2	Scale	0.39	$(0.36, 0.42)$
Rater 33	Shift	-0.05	$(-0.10, -0.00)$
	Scale	0.49	$(0.46, 0.52)$
Rater 34	Shift	-0.31	$(-0.40, -0.22)$
Modality 3	Scale	0.45	$(0.40, 0.50)$
Rater 35	Shift	-0.30	$(-0.38, -0.22)$
Modality 3	Scale	0.53	$(0.49, 0.58)$
Rater 36*	Shift	0.05	$(-0.04, 0.16)$
Modality 2	Scale	0.37	$(0.33, 0.41)$
Rater 37*	Shift	-0.34	$(-0.49, -0.13)$
Modality 2	Scale	0.24	$(0.07, 0.34)$
Rater 38*	Shift	-0.02	$(-0.44, 0.43)$
Modality 2	Scale	0.06	$(0.01, 0.30)$

Table 9: MCMC estimated posterior median and 95% equal-tailed credible intervals (CI's) for the HRM rater parameters based on six ratings for each of 557 student responses to eleven items on the Florida Grade 5 Mathematics assessment. The modality of raters who rated in only one modality is indicated in bold; the other raters rated in both modalities 2 and 3. Raters analyzed in our initial analysis of items 9, 10, and 11 are marked with asterisks.

Parameter		Modality 1	Modality 2	Modality 3	Overall
Shift (severity)	$\overline{\phi}$	-0.08	-0.13	-0.22	-0.15
Scale (unreliability)	$\overline{\psi}$	0.46	0.39	0.47	0.44

Table 10: Mean shift and scale effects for raters in Table 9 who only rated in one modality; and for the entire group of 38 raters.

Item	Parameter	Median	95% CI
Item 1	Difficulty β_1	-0.02	$[-0.16, 0.11]$
	Step 1 γ_{11}	0.05	$[-0.16, 0.26]$
Item 2	Difficulty β_2	-0.66	$[-0.54, -0.77]$
	Step 1 γ_{21}	-1.26	$[-1.53, -0.99]$
	Step 2 γ_{22}	-0.72	$[-0.99, -0.47]$
	Step 3 γ_{23}	0.54	$[0.25, 0.84]$
Item 3	Difficulty β_3	0.84	$[0.65, 1.04]$
	Step 1 γ_{31}	0.09	$[-0.38, 0.21]$
Item 4	Difficulty β_4	-0.00	$[-0.18, 0.20]$
	Step 1 γ_{41}	-1.79	$[-2.01, -1.57]$
Item 5	Difficulty β_5	-0.67	$[-0.79, -0.55]$
	Step 1 γ_{51}	0.06	$[-0.13, 0.25]$
Item 6	Difficulty β_6	0.29	$[0.15, 0.43]$
	Step 1 γ_{61}	-1.43	$[-1.85, -1.07]$
Item 7	Difficulty β_7	-0.36	$[-0.47, -0.25]$
	Step 1 γ_{71}	-2.43	$[-2.97, -1.96]$
Item 8	Difficulty β_8	-0.28	$[-0.41, -0.15]$
	Step 1 γ_{81}	0.41	$[0.22, 0.60]$
Item 9	Difficulty β_9	-0.28	$[-0.38, -0.19]$
	Step 1 γ_{91}	-0.24	$[-0.51, 0.02]$
	Step 2 γ_{92}	0.54	$[0.22, 0.86]$
	Step 3 γ_{93}	-0.64	$[-1.09, -0.20]$
Item 10	Difficulty β_{10}	0.05	$[-0.08, 0.18]$
	Step 1 $\gamma_{10,1}$	-0.96	$[-1.26, -0.67]$
Item 11	Difficulty β_{11}	1.09	$[0.89, 1.32]$
	Step 1 $\gamma_{11,1}$	-1.44	$[-1.96, -0.97]$

Table 11: MCMC estimated posterior median and 95% equal-tailed credible intervals (CI) for the item difficulty and item-step parameters based on six ratings for each of 557 student responses to eleven items on the Florida Grade 5 Mathematics assessment.

from Section 3.1 that we only estimate $K-2$ item step parameters for each K-category item). Several of the items have item step parameters whose magnitudes are roughly in the 1.5 to 2.5 range; this suggests that it is easy for examinees to obtain a nonzero score on the item, but relatively difficult to get a high score. This is another indication that the items were somewhat difficult for the examinees, but perhaps with some suggestion of leniency across all raters at the low end of performance on each item.

We also note that the item step parameter estimate γ_{92} is out of order with the other item step parameter estimates for item nine, suggesting either that few examinee responses met the criterion for category two of the scoring rubric for this item, or that raters could not agree about what constituted a category two response; the partial tabulation of pairs of ratings for item nine in Table 3 reinforces this impression. Thus we may wish to re-evaluate the scoring rubric or the rater training for item nine.

5.3 Model Comparisons

In Table 12 we compare the fits of the IRT Facets model with additive rater effects to the fit of the HRM model, on the simulated data from Section 5.1 and the full Florida rater study data from Section 5.2.3. Since the models are not nested in the usual sense (the IRT Facets model is not obtained by constraining the HRM parameters in a locally linear way; see discussion at the end of Section 3.2), likelihood ratio chi-squared tests cannot be used. Instead, we use a measure of fit known as the Schwarz Criterion, also known as the Bayes Information Criterion (BIC; e.g. Kass and Raftery, 1995). The difference between BIC values for two models approximates the logarithm of the Bayes Factor, which is often used for comparing models in Bayesian statistics; the Bayes Factor can be difficult to compute directly, especially for large models estimated with MCMC methods (see for example DiCiccio, Kass, Raftery and Wasserman, 1997). For a marginal model with p parameters and N examinees, the BIC is given by

$$BIC = -2 \cdot \log\left(\text{marginal model}\right) + p \cdot \log(N),$$

where the marginal model is evaluated at the modal parameter estimates.

Thus, BIC can be interpreted as the usual log-likelihood statistic, penalized for the number of parameters in the model. Any reduction in BIC is considered good, since the penalty $p \cdot \log(N)$ compensates for capitalization on chance; however a commonly used rule of thumb (Kass and Raftery, 1995) for Bayes Factors is that a decrease of 2–6 in this BIC statistic is considered moderately good evidence, and a change of 10 or more is considered very strong evidence, in favor of the model with the lower BIC.

It is no surprise that in Table 12 the HRM fits better than the IRT Facets model in the simulated data, since this data was simulated from the HRM itself. The large change in BIC, a decrease of 4,000 for an increase of only three additional parameters (essentially, the three rater scale parameters) is impressive evidence that the dependence modeled by the HRM cannot somehow be accommodated by the IRT Facets model. Even more impressive is the decrease of over 20,000 for an increase of 38 parameters (again, essentially the rater scale parameters), in favor of the HRM in fitting the Florida mathematics assessment rating study data set. Thus, in the real data too, the HRM is providing a much better model of the dependence structure of the data.

5.4 The Information for Scoring Examinees in Multiple Ratings

To illustrate the additional information available in multiple ratings for estimating examinee scores on the latent scale, we examined the θ estimates of five typical examinees taken from the full data set in Section 5.2.3. In Table 13, we have compared these examinees' posterior median and equal-tailed 95% credible interval estimates, under all three models estimated in Sections 5.2.1, 5.2.2, and 5.2.3. The narrowing of the 95% intervals from Items 9, 10, and 11 within modality two, to the complete eleven-item data set within

HRM Simulated Data

Model	$-2\log$(marginal model)	Parameters	Examinees	BIC
IRT Facets	$14,505$	23	500	$14,648$
HRM	$10,405$	27	500	$10,573$

Florida Data

Model	$-2\log$(marginal model)	Parameters	Examinees	BIC
IRT Facets	$55,256$	65	557	$55,667$
HRM	$33,607$	103	557	$34,258$

Table 12: Model fit comparisons for the HRM-simulated data (Section 5.1) and for the full rater study data set from the Florida Grade 5 mathematics assessment (Section 5.2.3).

modality two, is due mostly to the increased number of items (11 vs. 3), since in both cases there were two ratings per item. The changes in the 95% credible intervals from the complete modality two data to the full data with all three modalities, is due to increasing the number of ratings per item from two to six.

We can see in Table 13 that most of the reduction in uncertainty about θ is obtained by increasing the number of items from three to eleven. Going from the eleven-item, two-rater data from the modality two extract to the eleven-item, six-rater data produced interval estimates the same width or slightly narrower in four cases, and a wider interval in one case. That there is not a greater reduction in θ estimation uncertainty in this example may in part be due to a degree of unreliability in the raters, which is perhaps caused by the multiple-modality design: raters in modality one tended to be more lenient than raters in other modalities for example (see Table 10), and so there is an inherent disagreement across modalities about the quality of the student work. Adding modality one raters to the pool of modality two raters, for example, yields a benefit (more raters reduces uncertainty in estimating examinee proficiencies) and a cost (disagreement across modalities in how to scores student work increases uncertainty in estimating examinee proficiencies), that may cancel each other out, and so adding new raters in different modalities does not necessarily help. In other situations, where we are adding raters who share stronger consensus in how to score student work, we might expect to see a more consistent decrease in the standard errors for examinee latent scale scores.

6 Discussion

In this paper we have implemented Patz's (1996) hierarchical rater model (HRM) for polytomously scored item response data, so that it can be employed with data sets approaching the sizes of those encountered in large-scale educational assessments, or at least in rater studies supporting those assessments. We have shown how the HRM "fits in" to the generalizability theory framework that has been the traditional analysis tool for rated item response data—indeed, the HRM is a standard generalizability theory model for rating data, with IRT distributions replacing the normal theory true score distributions that are usually implicit in inferential applications of the model: observed ratings are related to ideal ratings of each piece of student work through a simple signal detection model that can be further parameterized to be sensitive to individual rater severity and reliability effects, and ideal ratings are related to a latent scale score via a conventional IRT model such as the Partial Credit Model.

In simulated and real data examples, we have shown how the current implementation of the HRM can

Subject	Data Set	Median	95% CI	CI width
Subject 175	Modality 2 (9,10,11)	-0.56	$(-1.63, 0.57)$	2.20
	Modality 2	-0.85	$(-1.62, -0.36)$	1.26
	Full Data	-0.93	$(-1.71, -0.21)$	1.50
Subject 115	Modality 2 (9,10,11)	-1.83	$(-3.23, -0.69)$	2.54
	Modality 2	-1.98	$(-3.21, -1.18)$	2.03
	Full Data	-1.85	$(-2.89, -0.97)$	1.92
Subject 313	Modality 2 (9,10,11)	-1.82	$(-3.33, -0.67)$	2.66
	Modality 2	-0.90	$(-1.76, -0.27)$	1.49
	Full Data	-0.93	$(-1.75, -0.24)$	1.51
Subject 492	Modality 2 (9,10,11)	-1.15	$(-2.33, -0.12)$	2.21
	Modality 2	-1.41	$(-2.26, -0.53)$	1.73
	Full Data	-1.42	$(-2.35, -0.62)$	1.73
Subject 71	Modality 2 (9,10,11)	-1.29	$(-2.77, -0.01)$	2.76
	Modality 2	-1.60	$(-2.71, -0.67)$	2.04
	Full Data	-1.62	$(-2.60, -0.81)$	1.79

Table 13: Comparison of θ estimates for five well-spaced examinees, under each of the three HRM models estimated in Section 5.

be used to scale items and examinees, and learn about rater quality. Using the Schwarz criterion (BIC) we have shown that the HRM fits far better than the IRT Facets model, suggesting that the dependence between multiple ratings of the same student work that the Facets model fails to capture is an important component of multiple-rating assessment data. Using polytomous response data simulated from the HRM, we showed that the HRM is effective at item and rater parameter recovery, and displayed some biases in IRT Facets item parameter estimates that we believe occur when some raters are relatively unreliable or inconsistent in their ratings. Both models produce interval estimates for examinee proficiencies that capture the true θ values, but the IRT Facets model intervals were substantially narrower than the corresponding HRM intervals.

The HRM is one of several current approaches to correcting this underestimation of standard errors for estimating θ, as reported by Patz (1996), Junker and Patz (1998), Donoghue and Hombo (2000), and others. Bock, Brennan and Muraki (1999) construct a generalizability theory based "design effects" correction for the conventional IRT Facets model, and Wilson and Hoskens (1999) replace the conditional independence assumptions of the conventional IRT Facets model with a rater bundle model analogous to Rosenbaum's (1988) item bundles. When prior estimates of the variance components in a traditional generalizability theory model for rating data are available, and we only desire to quickly scale many examinees, the design effects correction of Bock, Brennan and Muraki (1999) may be faster than, and therefore perhaps preferable to, the HRM approach. The HRM may be viewed as a kind of restriction of Wilson and Hoskens' (1999) rater bundles model (RBM). The HRM scales up to multiple reads more readily than the RBM, because it treats raters as *a-priori* exchangeable. On the other hand, if the researcher has a clear idea of the nature of dependence between reads of the same student work, the RBM provides a ready set of tools for zeroing in on that *a-priori* hypothesized dependence. Verhelst and Verstralen's (2000) IRT model for multiple raters is also closely related to the HRM.

We also examined successively larger extracts of a study of three different rating modalities intended to support a Grade 5 mathematics assessment given in the State of Florida (Sykes, Heidorn and Lee 1999). The parameterization of the HRM used in this paper appeared to be very successful at identifying individual raters of poor reliability or excessive severity. Finally we compared interval estimates from the HRM fitted to all three extracts of the Florida mathematics exam data, to show the effects of increasing number of

items and number of raters on standard errors of estimation for examinee scale scores. Our work with this rating study data set also shows that the HRM can easily handle loosely connected rating designs with many possible pairings of raters and severe imbalance in the assignment of raters to items.

The parameterization of the HRM used in this paper emphasizes raters' individual severity and reliability. Other parameterizations of the HRM are also possible, and may be more appropriate in other settings. One set of natural extensions of our parameterization of the HRM would allow us to assess the effects of rater background variables, student background variable, item features, and time, on ratings:

- Are there certain types of students that this rater rates more severely than other types?
- Does rating severity drift (e.g. towards more or less leniency) as a function of time of day?
- Which raters, if any, are giving students very similar ratings on a number of conceptually distinct items (that is, appear to be operating under a halo effect)?

Such analyses within the HRM only require that the relevant covariates (student type, item type, time of day of rating, etc.) be collected, and then incorporated into a model like (3) for the table of rating probabilities in Table 1. For example, we could incorporate mean rater shift and scale estimates across modalities in the Florida mathematics assessment data, by decomposing rater shift parameters ϕ as

$$\phi_{rm} = \phi_m + \delta_{rm}, \quad \sum_r \delta_{rm} = 0$$

for rater r rating under modality m, and similarly for rater scale parameters ψ. Estimates of ϕ_m and ψ_m would then replace our averages $\overline{\phi}$ and $\overline{\psi}$ in Table 10 above. In the setting of centralized scoring sessions, rater table effects might be modeled similarly.

Other questions may be handled by a more radical reparametrization of the probabilities in Table 1. For example,

- If a rater is having trouble applying a rating scale consistently, which scale points are causing the problems? Are the raters having trouble agreeing on what is a good performance, or what is a poor performance, or is there lack of consensus across the whole scale?
- Are any raters using only the inner categories of a rating scale and not venturing out to use the outer categories?

For example the second phenomenon above could be modeled by allowing the rater shift parameter $\phi_{r\xi}$ to depend on the ideal rating category ξ as well as the rater r; a rater exhibiting the "play it safe" strategy would appear to be more severe when ξ was low and more lenient when ξ was high; differential consistency across scale points could be modeled with a similar modification of the rater scale parameters $\psi_{r\xi}$. Either phenomenon might be modeled to operate only for particular items. In addition, the unimodal shape suggested by (3) might be replaced with some other shape.

We expect that as digital imaging technology improves, decentralized online scoring may replace centralized scoring sessions where all raters are in one room, supervised closely at separate tables by table leaders. In these decentralized scoring designs raters, once trained, work on their own in front of a computer terminal at a location of their choosing. A supervisor is online to provide assistance when needed, but much of the valuable qualitative information that supervisors in centralized scoring sessions use to monitor and maintain rating quality—raters' body language, raters flipping back and forth from the rubrics to the student work, discussion with raters of difficult-to-rate cases—is not available in these decentralized scoring environments.

Identifying raters who are having problems from their rating data alone will likely become much more important as a supervisors' direct contact with raters becomes less frequent. Without adequate statistical tools to provide supervisors with useful and timely feedback on each rater, there is greater opportunity for

raters to get off track and less opportunity to quickly bring them back into the fold when they stray. Multiple ratings can provide the data needed for adequate statistical monitoring on the basis of rating data alone, as well as providing some improvement in the precision of estimation of examinee proficiencies. We hope that the hierarchical rater model, and similar approaches such as that of Wilson and Hoskens (1999) and Verhelst and Verstralen (2000) that appropriately account for dependence between ratings, will provide the modeling basis for these statistical monitoring systems.

In order for any of these approaches to be useful in distributed on-line rating systems, they must run fast enough to provide real-time feedback on raters' performance. The MCMC estimation methods for the HRM that we have described in this paper can fit a model "overnight"; this may be improved somewhat by replacing the continuous rater shift parameters ϕ with appropriately-chosen discrete parameters, as suggested in Section 5 above. Further improvements in speed may require other computing techniques, including marginal maximum likelihood (e.g. Donoghue and Hombo, 2000), as well as data summaries that focus on sufficient or "nearly" sufficient statistics for the effects parametrized in the HRM.

References

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

Bock, R. D., Brennan, R. L., and Muraki, E. (1999). The introduction of essay questions to the GRE: Toward a synthesis of item response theory and generalizability theory. Paper presented at the Annual meeting of the American Educational Research Association, April 1999, Montreal Canada.

Brennan, R. L. (1992). *Elements of generalizability theory (revised edition).* Iowa City IA: ACT Publications.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Researcher, 16,* 14–20.

Cronbach, L. J., Linn, R. L., Brennan, R. L., and Haertel, E. (1995). Generalizability analysis for educational assessments. *Evaluation Comment.* Los Angeles: UCLA's Center for the Study of Evaluation and The National Center for Research on Evaluation, Standards and Student Testing. [Online.] `http://www.cse.ucla.edu`. Retrieved 16 May 2000.

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association, 92,* 903–915.

Donoghue, J. R., and Hombo, C. M. (2000). *A comparison of different model assumptions about rater effects.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. April 2000, New Orleans, LA.

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

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

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

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48,* 3-26.

Fischer, G. H., and Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika, 59,* 177–192.

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

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Practical Markov chain Monte Carlo*. New York: Chapman and Hall.

Holland, P. W. (1990). The Dutch Identity: A new tool for the study of item response models. *Psychometrika, 55,* 5–18.

Johnson, M. S., Cohen, W. and Junker, B. W. (1999). Measuring Appropriability in Research and Development with Item Response Models. CMU Statistics Department Technical Report #690. [Online.] URL `http://www.stat.cmu.edu/cmu-stats/tr`. Retrieved 16 May 2000.

Junker, B. W. and Patz, R. J. (June 1998). *The hierarchical rater model for rated test items.* Presented at the Annual North American Meeting of the Psychometric Society, June 17–21, 1998, Champaign-Urbana IL, USA.

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

Koretz, D., Stecher, B., Klein, S., and McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: findings and implications. *Educational Measurement: Issues and Practice, 13,* 5–16.

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

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Massachusetts: Addison-Wesley.

Lukhele, R., Thissen, D., and Wainer, H. (1994). On the relative value of multiple choice, constructed response, and examinee selected items on two achievement tests. *Journal of Educational Measurement, 31,* 234–250.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Mislevy, R. J., Beaton, A. E., Kaplan, B., Sheehan, K. M. (1992). Estimating population characteristics from a sparse matrix sample of item responses. *Journal of Educational Measurement, 29,* 131–154.

Mislevy, R. J. and Wu, P. K. (1996). Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing. ETS Technical Report, RR-96-30-ONR.

Myford, C. M., and Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system.* Center for Performance Assessment Research Report. Princeton, NJ: Educational Testing Service.

Patz, R. J. (1996). Markov Chain Monte Carlo Methods for Item Response Theory Models with Applications for the National Assessment of Educational Progress. Ph.D. Dissertation, Carnegie Mellon University.

Patz, R. J. and 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. and Junker, B. W. (1999b). Fitting item response models to incomplete, multiple item type educational assessment data using Markov chain Monte Carlo methods. In press, *Journal of Educational and Behavioral Statistics.*

Patz, R. J., Junker, B. W. and Johnson, M. S. (1999). *The hierarchical rater model for rated test items and its application to large-scale educational assessment data.* Paper presented April 23, 1999 at the Annual Meeting of the American Educational Research Association, Montreal Canada.

Patz, R. J., Wilson, M., and Hoskens, M. (1997). Optimal rating procedures for NAEP open-ended items. Final report to the National Center for Education Statistics under the redesign of NAEP.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349–359.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. [The learning and solving of complex reasoning items.] *Zeitschrift für Experimentelle und Angewandte Psychologie, 3,* 456–506.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). *BUGS 0.5: Bayesian inference using Gibbs Sampling, Version ii*. Technical report of the MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK. [Online.] `http://www.mrc-bsu.cam.ac.uk/bugs/`. Retrieved 16 May 2000.

Sykes, R. C., Heidorn, M., and Lee, G. (1999). *The assignment of raters to items: controlling for rater bias.* Paper presented at the annual meeting of the National Council on Measurement in Education. April 1999, Montreal Canada.

Tanner, M. A. (1996). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions. 3^{rd} Edition.* New York: Springer-Verlag.

Verhelst, N. D., and Verstralen, H. H. F. M. (2000). IRT models for multiple raters. In A. Boomsma, T. Snijders, and M. Van Duijn (Eds.), *Essays in Item Response Modeling* (pp. xxx–xxx). New York: Springer-Verlag. (to appear, September 2000).

Wilson. M. and Hoskens, M. (1999). The rater bundle model. Submitted for publication.

Wilson, M., and Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19(1),* 51-72.

Wu, M. L., Adams, R. J, and Wilson, M. R. (1997). *ConQuest: Generalized item response modeling software.* ACER.