Some thoughts about raters and parametric item response theory[a]

Educational Testing Service
Princeton NJ
July 9, 1998

Brian W. Junker
Department of Statistics
Carnegie Mellon University
Pittsburgh PA 15213
brian@stat.cmu.edu

Richard M. Patz
CTB/McGraw-Hill
20 Ryan Ranch Road
Monterey CA 93940
rpatz@ctb.com

[a] This paper is based on portions of Chapter 6 of Richard J. Patz's Ph.D. dissertation at Carnegie Mellon University, which was supported by a Spencer Foundation Dissertation Fellowship in Educational Research. Junker's work was supported by NSF grant #DMS-97.05032.

Talk Outline:

- Review of the LLTM/FACETS style models
- Asymptotics for LLTM/FACETS: *too good!*
- Motivating and defining the HRM:
  - *First* the examinee answers the question;
  - *Then* the rater rates it.
- Asymptotics for HRM: *just right!*
- MCMC fits of very simple made-up data
- Speculations about variations, applications, future work, etc.

Review of the LLTM/FACETS style models

- Rated responses are a standard part of testing.
- Most models based on LLTM/FACETS formulation
  - ⋆ $i = 1 \ldots N$ Examinees
  - ⋆ $j = 1 \ldots J$ Items
  - ⋆ $r = 1 \ldots R$ Raters
  - ⋆ $X_{ijr} = 1 \Leftrightarrow$ rater r rates item j correct for examinee i; otherwise $X_{ijr} = 0$
  - ⋆ $P_{jr}(\theta_i) = P[X_{ijr} = 1|\theta_i]$:

$$\log \frac{P_{jr}(\theta_i)}{1 - P_{jr}(\theta_i)} = \theta_i - \beta_j - \rho_r$$

    plus LI and identifiability constraints

- Fischer's LLTM (1973, 1983); Linacre's (1989) FACETS. Genralize to polytomous, e.g. via PCM.

  <u>*Note*</u>*: Fischer has argued against using the LLTM for this purpose (Glas, priv. comm.)*

LLTM/Facets models: Advantages

- Analogous to Generalizability Models (Brennan, 1983; Cronbach et al., 1995)
- Additive logit-scale decomposition for each factor (subjects, items, raters, . . .).
- Easy to generalize estimation methods from standard IRT models, e.g.
  - CML: Fischer. . . ; FACETS (Linacre, 1989)
  - MML: Parscale (Muraki, 1995); ConQuest (Wilson et al., 1997)

  and easy to handle incomplete rater designs.
- Easy to generalize to account for some observed response/rating phenomena, e.g.:
  - GLLTM (Patz, 1996): rated items of varying discrimination;
  - 2PLRM (Wolfe, 1998): varying "spread" in raters' scale usage.

LLTM/Facets models: Disadvantages

- Individual (within-) rater reliability is unmodelled.

- Raters are usually treated as fixed effects, so even inter-rater reliability is not in the model.

  Note: *As I understand it, this is a big part of Fischer's objection to using LLTM for raters.*

- Difficult to model situations in which raters use more or fewer categories than examinee responses.

- Excessively optimistic about SE(θ).

Asymptotics for LLTM/FACETS:

Suppose items (β_j) and raters (ρ_r) have been pre-calibrated and these parameters are known.

- By usual arguments (Lord, 1980), the standard error for estimating θ in this model is $SE(\theta) = \sqrt{1/I_{JR}(\theta)}$, where

$$I_{JR}(\theta) = \sum_{j=1}^{J} \sum_{r=1}^{R} \frac{[P'_{jr}(\theta)]^2}{P_{jr}(\theta)[1 - P_{jr}(\theta)]}$$

  as long as the β's and ρ's are bounded, $I_{JR}(\theta) \to \infty$, and hence $SE(\theta) \to 0$, as $J \to \infty$.

  This is good!

- However, *under the same assumptions*,

$$\lim_{R \to \infty} I_{JR}(\theta) = \infty, \text{ and } \lim_{R \to \infty} SE(\theta) = 0$$

  so it appears that we can reduce estimation error for θ by increasing the number of raters R, keeping the number of items J fixed.

  This is too good!

Asymptotics for LLTM/FACETS: What's wrong?

- Increasing the number of raters should tell us much more about the examinee's "true" responses to the items, than about the examinee's θ.

- *The problem:* The usual LLTM/FACETS model *assumes LI across raters* as well as across items.

  The information in a single item,

  $$I_j(\theta) = \sum_{r=1}^{R} \frac{[P'_{jr}(\theta)]^2}{P_{jr}(\theta)[1 - P_{jr}(\theta)]}$$

  grows linearly in the number of raters R assigned to the problem!

Asymptotics for LLTM/FACETS: How to fix it?

- *If the ρ's are treated as random effects*, the information will not increase linearly with R.

  This would solve the information problem, but at the expense of never being able to calibrate raters.

- By *reformulating IRT-based rater models* we can
  - Solve the information problem,
  - Allow the possibility of calibrating raters, and
  - Open the door to modeling within-rater reliability, etc.

Motivating and defining the HRM:

- *First* the examinee answers the question;
- *Then* the rater rates it.

Definition:

- Let $\xi_{i1}, \ldots, \xi_{iJ}$ be the *true responses* of the examinee:

  $$\xi_{ij} = \begin{cases} 1 & \text{if examinee has answered correctly} \\ 0 & \text{else} \end{cases}$$

- The *true ICC's* are

  $$P_j(\theta) = P[\xi_j = 1|\theta]$$

  and can follow any convenient IRT model.

- Define X_{ijr} as before, but now governed by the parameters

$P[X_{ijr} = x \mid \xi_{ij} = \xi]$	$x = 1$	$x = 0$
$\xi = 1$	p_{11ijr}	$1 - p_{11ijr}$
$\xi = 0$	p_{01ijr}	$1 - p_{01ijr}$

<u>The HRM: Some details</u>

- The true ICC's $P_j(\theta)$ may follow any convenient IRT model. For specificity we use the Rasch model below.

- One may also wish to further model the $p's$. For specificity, we will assume

$$\begin{aligned} p_{01ijr} &= p_{01r}, \text{ rater } r\text{'s false positive rate;} \\ p_{11ijr} &= p_{11r}, \text{ rater } r\text{'s true positive rate.} \end{aligned}$$

  This framework allows us to model various aspects of rating behavior.

- The full ICC for an observable item response under the HRM is

$$\begin{aligned} P_{jr}(\theta_i) &= E\{P[X_{ijr} = 1|\theta_i, \xi_{ijr}]|\theta\} \\ &= P_j(\theta_i)p_{11r} + (1 - P_j(\theta_i))p_{01r} \end{aligned}$$

- Generalization to polytomous responses/ratings is clear. Also, note that the number of categories of ξ need not equal the number of categories of x.

Asymptotics for HRM:

To calculate $I_j(\theta)$ for the HRM, we need to borrow a tool used in the E-M literature, to compute Fisher information.

Missing information principle:

$$\begin{aligned}\frac{-\partial^2 \log p(\theta|\underline{X}_j)}{\partial\theta^2} &= -E\left[\left.\frac{-\partial^2 \log p(\theta|\underline{X}_j,\xi_j)}{\partial\theta^2}\right|\theta,\underline{X}_j\right] \\ &\quad - \text{Var}\left[\left.\frac{-\partial \log p(\theta|\underline{X}_j,\xi_j)}{\partial\theta}\right|\theta,\underline{X}_j\right]\end{aligned}$$

where $\underline{X}_j$ is the set of ratings for item j, ξ_j is the true response to item j, and θ is to be estimated (e.g., Tanner, 1996; Louis, 1992).

Because of local independence, information will still add across items, so that

$$I_{JR}(\theta) = \sum_{j=1}^{J} I_j(\theta) = \sum_{j=1}^{J} E\left[\left.\frac{-\partial^2 \log p(\theta|\underline{X}_j)}{\partial\theta^2}\right|\theta\right]$$

Asymptotics for HRM: Calculation...

Using the missing information principle,

$$
\begin{aligned}
&\frac{-\partial^2 \log p(\theta|\underline{X}_j)}{\partial\theta^2} \\
&\quad= \; I_j^{true}(\theta) \; - \; \lambda_j''(\theta)[\tilde{P}_j(\theta;\underline{X}_j) - P_j(\theta)] \\
&\qquad\quad - \; [\lambda_j'(\theta)]^2 \tilde{P}_j(\theta;\underline{X}_j)(1 - \tilde{P}_j(\theta;\underline{X}_j)) \\
&\quad\equiv \; I_j^{true}(\theta) - A_j - B_j
\end{aligned}
$$

where

$$
\begin{aligned}
I_j^{true} &= \frac{[P_j'(\theta)]^2}{P_j(\theta)[1 - P_j(\theta)]} \\
\lambda_j(\theta) &= \log \frac{P_j(\theta)}{1 - P_j(\theta)}
\end{aligned}
$$

and

$$
\tilde{P}_j(\theta;\underline{X}_j) \;=\; P[\xi_j = 1|\theta,\underline{X}_j]
$$

It also follows that

$$
\begin{aligned}
E[A_j|\theta] &= 0 \\
E[B_j|\theta] &= [\lambda_j'(\theta)]^2 \Big\{ P_j(\theta)(1 - P_j(\theta)) \\
&\qquad - \; \mathrm{Var}\left[\tilde{P}_j(\theta;\underline{X}_j) \middle| \theta \right] \Big\}
\end{aligned}
$$

Asymptotics for HRM: just right!

With some effort, one can show

$$\lim_{R\to\infty} \mathrm{Var}\left[\tilde{P}_j(\theta; \underline{X}_j) \middle| \theta \right] = P_j(\theta)(1 - P_j(\theta))$$

[for example, this can be calculated explicitly in case $p_{01r} \equiv p_{01}$ and $p_{11r} = p_{11}$ for all r].

It follows that

$$\lim_{R\to\infty} I_j(\theta) = I_j^{true}(\theta)$$

and hence

$$\lim_{R\to\infty} I_{JR}(\theta) = \lim_{R\to\infty} \sum_{j=1}^{J} I_j(\theta) = \sum_{j=1}^{J} I_j^{true}(\theta) = I_J^{true}(\theta)$$

Hence, under the HRM,

- *As the number R of raters grows, we learn only as much about θ as we would if the test had been perfectly scored with no raters.*

MCMC fits of very simple made-up data

As a simple example we extract of 100 examinees from the famous LSAT data set considered by Bock and Aitken (1981). In our extract there are

- $N = 100$ examinees
- $J = 5$ items
- $R = 3$ raters (simulated)

We will estimate three models:

- The Rasch model applied to the perfectly scored, UNRATED 100×5 item response matrix.
- The FACETS model applied to the full $100 \times 5 \times 3$ data set.
- The HRM model in which the true responses are modeled with the Rasch model and each rater r has rating parameters p_{01r} and p_{11r}, again applied to the full $100 \times 5 \times 3$ data set.

MCMC details I...

We will estimate all three models in a Bayesian framework since then straightforward MCMC IRT estimation methods (Patz and Junker, 1997a,b) can be used.

It is enough to describe the HRM (Rasch and FACETS are special cases):

- $X_{ijr}|\xi_{ij} \;\sim\; Bern(p_{11r}\xi_{ij} + p_{01r}(1-\xi_{ij}))$;
- $p_{01r} \;\sim\; Beta(\alpha_0, \beta_0)$ and $p_{11r} \;\sim\; Beta(\alpha_1, \beta_1)$;
- $\xi_{ij}|\theta_i, \beta_j \;\sim\; Rasch(\theta_i, \beta_j)$;
- $\beta_j \;\sim\; Normal(0, \tau^2)$;
- $\theta_i \;\sim\; Normal(0, \sigma^2)$;
- $\sigma^2 \;\sim\; InvGamma(\alpha, \beta)$.

and we will take all free α's, β's $= 1$, and $\tau^2 = 2$.

MCMC details II. . .

Need to simulate from the "complete conditional" distributions

$$p(\text{param}|\text{rest})$$

where "param" stands for each free parameter or latent variable in the model and "rest" stands for all other parameters, latent variables, and the observed data:

- $p_{01r}|\text{rest} \sim Beta(\sum_i \sum_j \xi_{ij} X_{ijr} + \alpha_0,$
  $\sum_i \sum_j \xi_{ij}(1 - X_{ijr}) + \beta_0);$
- $p_{11r}|\text{rest} \sim Beta(\sum_i \sum_j (1-\xi_{ij}) X_{ijr} + \alpha_1,$
  $\sum_i \sum_j (1-\xi_{ij})(1 - X_{ijr}) + \beta_1);$
- $\xi_{ij}|\text{rest} \sim Bern(\tilde{P}_j(\theta_i; \underline{X}_{ij}));$
- $p(\beta_j|\text{rest}) \propto \prod_i P_j(\theta_i)^{\xi_{ij}}(1-P_j(\theta_i))^{1-\xi_{ij}}\, n(\beta_j|0,\tau^2);$
- $p(\theta_i|\text{rest}) \propto \prod_j P_j(\theta_i)^{\xi_{ij}}(1-P_j(\theta_i))^{1-\xi_{ij}}\, n(\theta_i|0,\sigma^2);$
- $\sigma^2|\text{rest} \sim InvGamma(\alpha + N/2, \beta')$, where

$$\beta' = \left[1/\beta + 1/2 \sum_i \theta_i^2\right]^{-1}$$

MCMC results

As might be expected,

For θ Estimates: 95% interval widths are ordered FACETS < UNRATED < HRM; σ^2 estimates are ordered similarly.

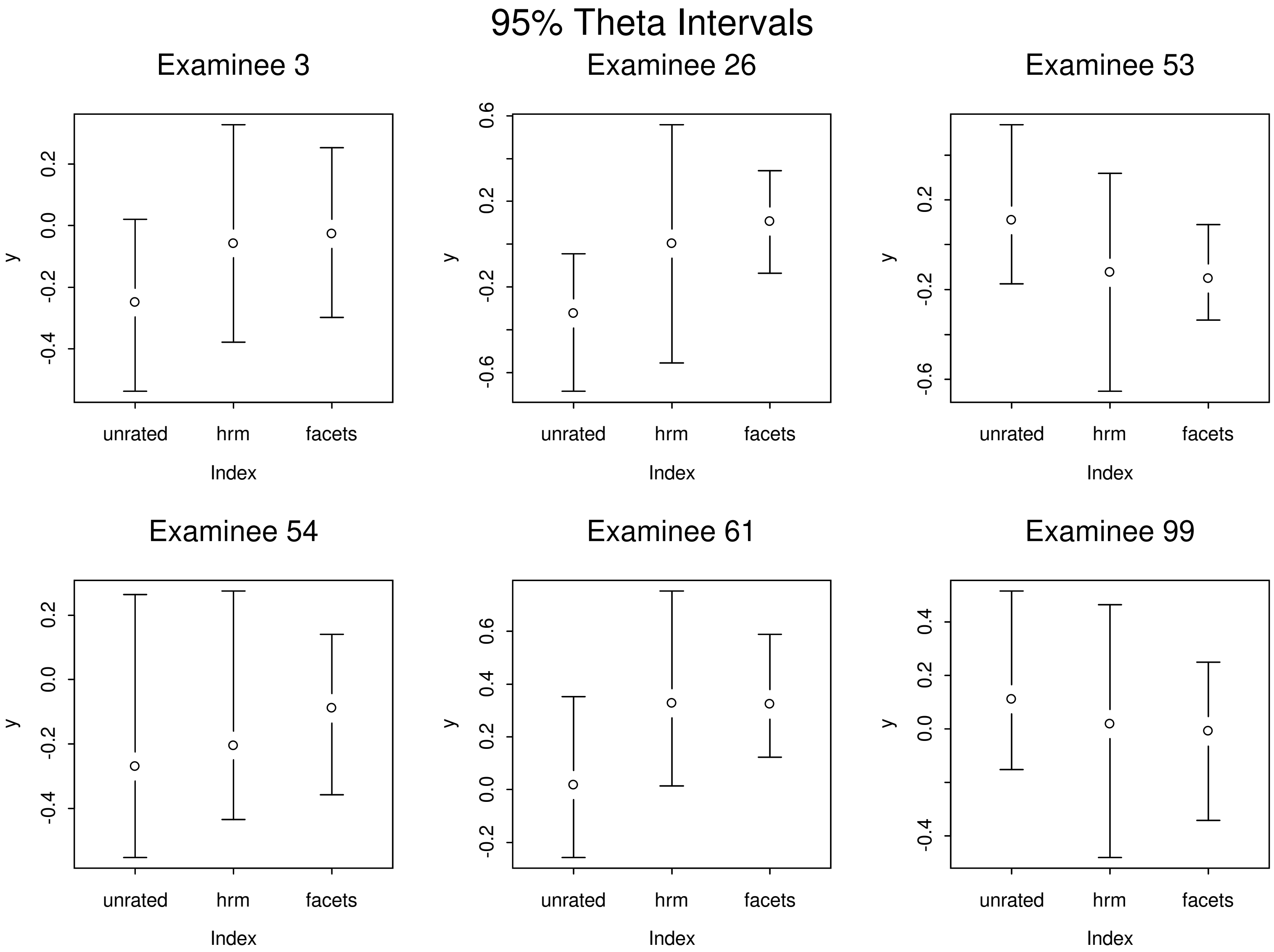
For β Estimates: 95% intervals are much wider for HRM than for UNRATED.

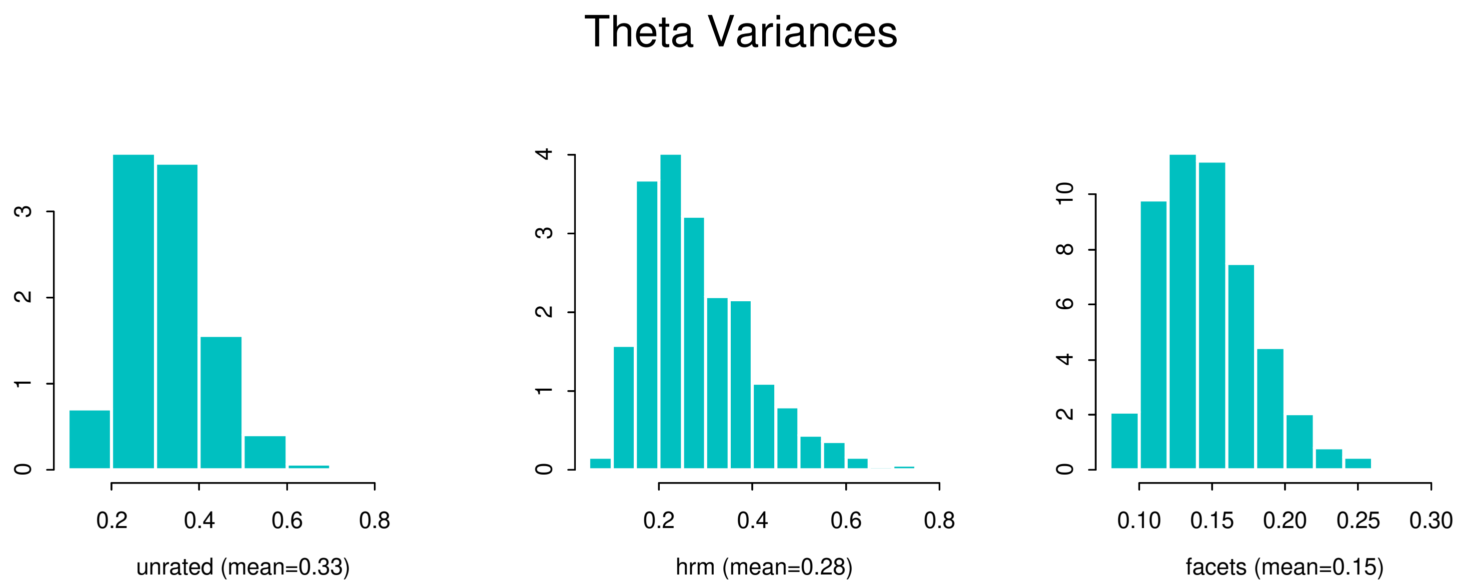
For Rater Effects: The simulated raters had the following parameter values:

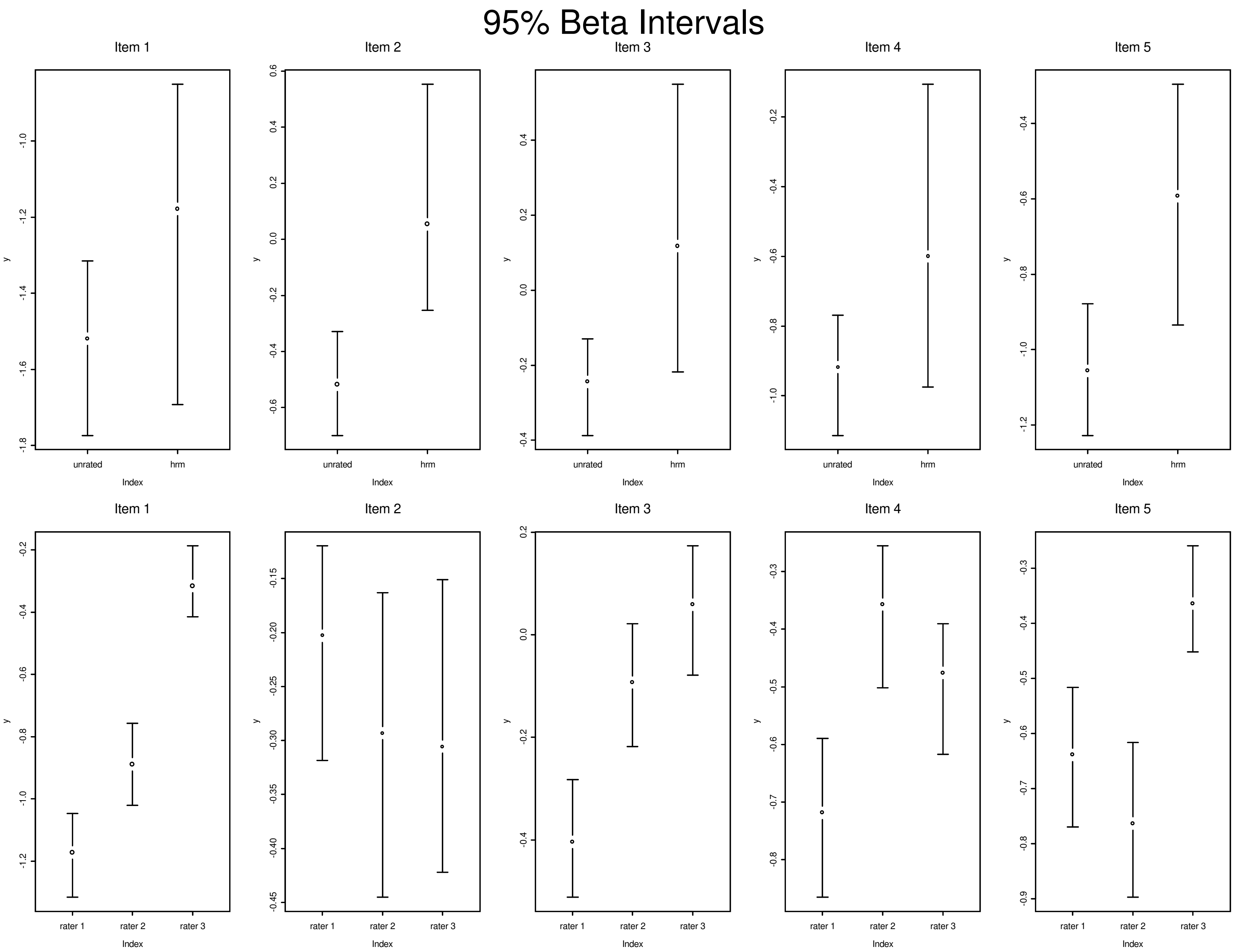
	Rater 1	Rater 2	Rater 3
p_{01r}	0.1	0.2	0.3
p_{11r}	0.9	0.8	0.7

were largely recovered by HRM.

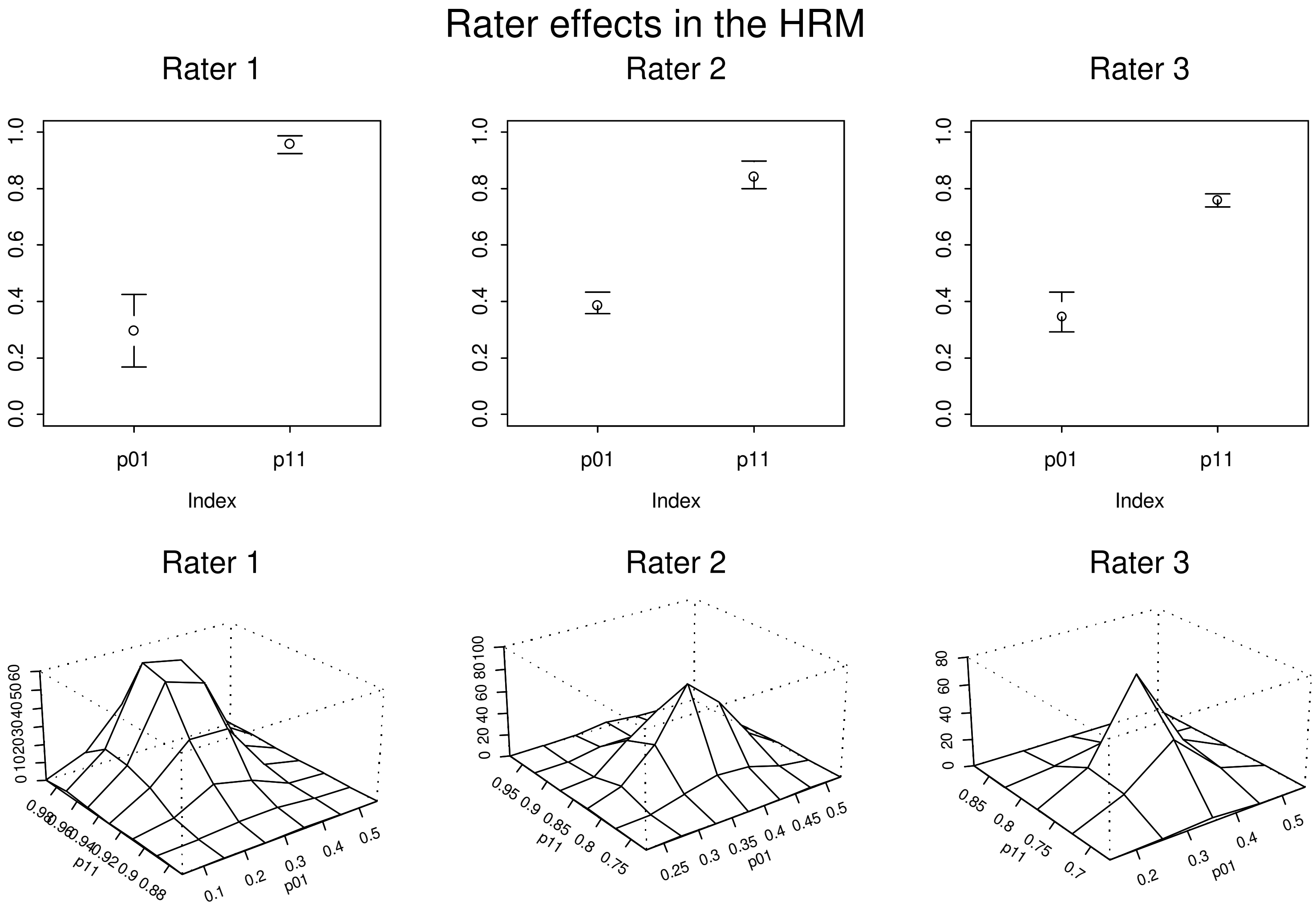
These show up as artifical interactions between raters and items in the FACETS model.


95% Theta Intervals
Examinee 3
Examinee 26
Examinee 53
Examinee 54
Examinee 61
Examinee 99
y
Index
unrated
hrm
facets
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6


Theta Variances
unrated (mean=0.33)
hrm (mean=0.28)
facets (mean=0.15)



95% Beta Intervals
Item 1
Item 2
Item 3
Item 4
Item 5
unrated
hrm
rater 1
rater 2
rater 3
Index
y



Rater effects in the HRM
Rater 1
Rater 2
Rater 3
0.0
0.2
0.4
0.6
0.8
1.0
p01
p11
Index
Rater 1
Rater 2
Rater 3
p11
p01

Speculations about variations, applications, future work, etc. I

- Current IRT-based rater models imply that perfect measurement can be achieved by increasing the number of raters.

  This can be illustrated with fairly small data sets.

- Current IRT-based rater models cannot easily model within-rater reliability and other rater behaviors.

  Rater inaccuracy can be misidentified as rater-by-item interactions in LLTM/FACETS models.

- The hierarchical rater model (HRM) addresses both concerns above.

- The HRM can be estimated using the straightforward MCMC methodology of Patz and Junker (1997a,b).

Speculations about variations, applications, future work, etc. II

- The HRM fully separates modeling rater behavior from modeling examinee performance.

- The HRM MCMC steps can be "pasted on" to any existing MCMC-based IRT estimation program, for example.

- The HRM can be generalized to polytomous response/rating data.

- The HRM can be used in situations where the "ideal" number of categories does not equal the number of rating categories.

- Much, *much* more work needs to be done to implement and test the HRM in various situations, compare its performance with FACETS models in situations that favor FACETS, to develop more detailed models for the HRM rater parameters, etc.

References and Notes

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. http://www.cse.ucla.edu.

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.

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.

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. (1997a). A straightforward approach to Markov chain Monte Carlo methods for item response models. In press, *Journal of Educational and Behavioral Statistics.*

Patz, R. J. and Junker, B. W. (1997b). Fitting item response models to incomplete, multiple item type educational assessment data using Markov chain Monte Carlo methods. Submitted for publication.

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.

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