Order-Constrained Bayes Inference for Dichotomous Models of Unidimensional Nonparametric IRT

George Karabatsos[1] and Ching-Fan Sheu[2]
[1]**University of Illinois at Chicago and** [2]**DePaul University**

This study introduces an order-constrained Bayes inference framework useful for analyzing data containing dichotomous-scored item responses, under the assumptions of either the monotone homogeneity model or the double monotonicity model of nonparametric item response theory (NIRT). The framework involves the implementation of Gibbs sampling to estimate order-constrained parameters, followed by inference with the posterior-predictive distribution to test the monotonicity, invariant item ordering, and local independence assumptions of NIRT. The Bayes framework is demonstrated through the analysis of real test data, and possible extensions of it are discussed. *Index terms: nonparametric item response theory; order-restricted inference; inequality constraints; Bayesian inference.*

Central to the paradigm of item response theory is the item response function (IRF), denoted by $P_j(\theta) = \Pr(X_j = 1|\theta) = 1 - \Pr(X_j = 0|\theta)$, which refers to the probability of a positive (e.g., correct) response to item j, given the value of a respondent's latent trait θ. Typically, parametric item response theory (PIRT) models assume a particular shape for the IRFs, such as the shape assumed by the logistic distribution. A well-known example is the two-parameter logistic model, given by $P_j(\theta) = [1 + \exp(-a_j(\theta - b_j))]^{-1}$, with $\theta \in \mathrm{Re}$ and parameters $a_j \geq 0$ and $b_j \in \mathrm{Re}$ (where Re refers to the domain of real numbers). In contrast (see Junker & Sijtsma, 2001, p. 211), nonparametric IRT (NIRT) models offer more flexible alternatives because they do not commit to any particular shape of the IRFs. The availability of such alternatives is useful, especially when a family of PIRT models has been shown to poorly fit a data set. Also, NIRT has a greater capability than PIRT for analyzing data sets containing either a small sample of respondents or items. Furthermore, because NIRT analysis focuses on testing key properties that underlie measurement, described next, they can provide a deep understanding of what PIRT models do.

Three basic assumptions characterize NIRT models (e.g., Junker & Sijtsma, 2001, pp. 211-212). They are

1. *Unidimensionality:* the latent trait takes values satisfying $\theta \in \mathrm{Re}$,
2. *Monotonicity:* $P_j(\theta)$ is a nondecreasing function of θ, and

Applied Psychological Measurement, Vol. 28 No. 2, March 2004, 110–125
DOI: 10.1177/0146621603260678

3. *Local independence:* postulates that a unidimensional latent trait θ exists, with the joint conditional probability of J item responses satisfying

$$P\,(X_1 = x_1, \ldots, X_J = x_J \mid \theta) = \prod_{j=1}^{J} P\,(X_1 = 1 \mid \theta)^{x_j}\,[1 - P\,(X_1 = 1 \mid \theta)]^{1-x_j}. \quad (1)$$

The current study focuses on unidimensional NIRT models for dichotomous item scores, although it is straightforward to modify these assumptions to address multidimensional NIRT models or NIRT models for polytomous item scores (e.g., Junker & Sijtsma, 2001).

For dichotomous item scores, there are two well-known unidimensional NIRT models. The first is the monotone homogeneity (MH) model, characterized exactly by these three assumptions (Holland & Rosenbaum, 1986; Junker, 1993; Meredith, 1965; Mokken, 1971; Mokken & Lewis, 1982). The one-, two-, three-, and four-parameter logistic models of PIRT are all special cases of the MH model (Sijtsma, 1998). The second model is a special case of MH, known as the double monotonicity (DM) model, which adds the assumption that each of the J item response functions $P(X_j = 1|\theta)$ are nonintersecting over the entire range of θ. This nonintersection property implies that the J items have an invariant order of difficulty and is called *invariant item ordering* (IIO) (Sijtsma & Junker, 1996). In fact, the one-parameter logistic model is a special case of DM (Sijtsma, 1998). Invariant item ordering is practical for many testing situations, such as intelligence testing, the analysis of differential item functioning, test equating, item banking, and exploring hypotheses about which cognitive operations are acquired by children (e.g., Sijtsma & Junker, 1996; Scheiblechner, 1995).

When either the MH or DM model fits a data set of dichotomous items, it is possible to measure respondents by the total test score $X_+ = \sum_{j=1}^{J} X_j$. Grayson (1988; Huynh, 1994) showed that $\Pr(X_+ = x_+|\theta)$ satisfies the monotone likelihood ratio property. This implies two important stochastic ordering properties—namely, the stochastic ordering of the manifest score ($\Pr(X_+ > x_+|\theta)$ is nondecreasing in θ for each x_+) and stochastic ordering of the latent trait ($\Pr(\theta > t|X_+ = x_+)$ is nondecreasing in x_+ for each t) (Hemker, Sijtsma, Molenaar, & Junker, 1997; Junker & Sijtsma, 2000). Furthermore, Stout (1990, Theorem 3.2) showed that X_+ is an ordinally consistent estimator of θ, whereas Hoijtink and Molenaar (1997) and Van Onna (2002) showed that the continuous latent trait θ can be well approximated by a (small) discrete number of latent classes.

The monotonicity and IIO assumptions of NIRT each imply qualitative order constraints on the IRFs, and actually, these constraints resemble (Karabatsos, 2001; Scheiblechner, 1995, 1999) the order-independence axioms of conjoint measurement theory (Krantz, Luce, Suppes, & Tversky, 1971, chap. 6). From a statistical perspective, it is natural to perform inference on the MH and DM models with order-constrained statistical inference. In fact, Sijtsma (1988, chap. 2) and Scheiblechner (1995, 1999) considered a classical isotonic regression approach (e.g., Barlow, Bartholomew, Bremner, & Brunk, 1972; Robertson, Wright, & Dykstra, 1988) for NIRT, which is couched in the classical frequentist framework of statistical inference. Ideas of order-constrained inference are also present in the so-called order-restricted latent class models of psychometrics. Croon (1991) and Vermunt (2001) investigated such models in the context of frequentist inference, whereas Hoijtink and Molenaar (1997) and Van Onna (2002) considered them in the context of Bayesian inference, with posterior inference based on the Gibbs sampler (e.g., Gelfand & Smith, 1990).

For practical reasons, a Bayesian inference framework may be preferable for order-constrained inference. Because such order restrictions render a complicated parameter space, it is very difficult to quantify uncertainty of the parameter estimates (e.g., with standard errors) and to derive the null sampling distributions necessary for performing global and detailed hypothesis tests of model fit. In contrast, an order-constrained Bayes approach, based on the Gibbs sampler, routinely addresses all these practical issues of the classical framework (e.g., Gelfand, Smith, & Lee, 1992).

Continuing the previous work by Karabatsos (2001) and Karabatsos and Ullrich (2002), this study presents a general order-constrained Bayes inference framework, useful for testing the monotonicity, invariant item ordering, and local independence assumptions of NIRT. These tests can be rigorously performed using the posterior-predictive distribution, based on Gibbs algorithms that generate samples from order-constrained posterior distributions of parameters. Tests that conclude violations of either monotonicity or local independence are evidence for the rejection of the MH and the DM model, and tests that conclude violations of invariant item ordering suggest evidence against the DM model.

The statistical framework for the order-constrained Bayesian model is presented in the next section, along with an illustration of the framework through its application on a real test data set consisting of dichotomous items. The last conclusion section contains comments on the proposed framework and describes possible extensions of it.

A Bayesian Framework for Nonparametric IRT

Two Data Matrices

From a given set of test data, two different data matrices are considered in the order-constrained Bayes inference framework. First, in testing the monotonicity (M) assumption, the proportion-data matrix of interest is

$$\mathbf{p}^{(M)} = \left(\hat{p}_{ij}^{(M)} = y_{ij}^{(M)} \Big/ N_{ij}^{(M)} \mid i = 0, \ldots, I;\ j = 1, \ldots, J;\ I = J - 1\right), \tag{2}$$

with $\mathbf{y}^{(M)} = \left(y_{ij}^{(M)} \mid i = 0, \ldots, I; j = 1, \ldots, J\right)$ and $\mathbf{N}^{(M)} = (N_{ij}^{(M)} | i = 0, \ldots, I;\ j = 1, \ldots, J)$. In (2), $y_{ij}^{(M)} \epsilon \{0, \ldots, N_{ij}^{(M)}\}$ is the number of correct responses on item j, obtained by the $N_{ij}^{(M)} \geq 0$ respondents who have rest score $R_{(j)} = (X_+ - X_j) = i \in \{0, 1, \ldots, I = J - 1\}$. Later, it is described how $\mathbf{p}^{(M)}$ is used to test for local independence. Second, to test the IIO assumption of the DM model, the proportion-data matrix of interest is

$$\mathbf{p}^{(IIO)} = \left(\hat{p}_{ij}^{(IIO)} = y_{ij}^{(IIO)} \Big/ N_{ij}^{(IIO)} \mid i = 0, \ldots, I;\ j = 1, \ldots, J; I = J\right), \tag{3}$$

where $\mathbf{y}^{(IIO)} = (y_{ij}^{(IIO)} | i = 0, \ldots, I; j = 1, \ldots, J)$ and $\mathbf{N}^{(IIO)} = (N_{ij}^{(IIO)} | i = 0, \ldots, I;\ j = 1, \ldots, J)$. In (3), $y_{ij}^{(IIO)} \in \{0, \ldots, N_{ij}^{(IIO)}\}$ is the number of correct responses on item j, obtained by the $N_{ij}^{(IIO)} \geq 0$ respondents with total test score $X_+ = i \in \{0, 1, \ldots, I = J\}$. Furthermore, in each of the matrices (2) and (3), the items are indexed $j = 1, \ldots, J$ to correspond with the ordering of the item proportion correct $\hat{p}_j = \frac{1}{N}\sum_{n=1}^{N} x_{nj}$ over the total number N of test respondents. In particular, item $j = 1$ is the most difficult item (lowest $\hat{p}_j$) and $j = J$ is the easiest (highest $\hat{p}_j$).

These two different matrices are now illustrated from a data set (see Patz & Junker, 1999) containing the responses of a random sample of 100 fourth-grade students (from a total sample of 3,000). Each examinee responded to a six-item reading test, and each item response is coded 1 = correct and 0 = incorrect. The test was part of the 1992 Trial State Assessment in Reading at Grade 4, conducted by the National Assessments of Educational Progress (NAEP).

Table 1 presents the NAEP data in the form of $\mathbf{p}^{(M)}$, where the rows $i = 0, 1, \ldots, I$ correspond to different levels of the rest score $i = R_{(j)}$; the six items $j = 1, \ldots, J$ are ordered by proportion correct $\hat{p}_j$; and each cell of the matrix contains $\hat{p}_{ij}^{(M)}$, the proportion of respondents in rest

Table 1

Proportion Data $\mathbf{p}^{(M)}$, Containing the Responses of 100 Examinees to Six Items of the National Assessments of Educational Progress (NAEP) Reading Test

	Item					
Rest score	1	2	3	4	5	6
0	.00 (3)	.40 (5)	.40 (5)	.00 (3)	.63 (8)	.57 (7)
1	.23 (17)	.00 (11)	.15 (13)	.28 (18)	.58 (19)	.57 (21)
2	.23 (17)	.19 (21)	.44 (27)	.45 (22)	.75 (24)	.78 (23)
3	.38 (29)	.47 (34)	.52 (21)	.66 (35)	.81 (22)	.84 (25)
4	.42 (24)	.55 (20)	.52 (29)	.89 (18)	**.63** (19)	**.80** (20)
5	**.40** (10)	**.44** (9)	.80 (5)	1.00 (4)	**.50** (8)	1.00 (4)
Item proportion correct	.33	.37	.46	.58	.68	.75

Note. Each cell indicates the proportion of the N_{ij} respondents (in parentheses), with rest score $R_{(j)}$, who answered item j correctly. The proportions in bold suggest areas of the data that contradict monotonicity.

score group $i = R_{(j)}$ answering item j correctly. Notice that for each of the six items (columns) in Table 1, some of the proportions (indicated in bold) decrease as a function of increasing rest score, so there are apparent violations of monotonicity in the data (see Junker & Sijtsma, 2000). However, it is not clear yet whether any of these reflect statistically significant deviations of monotonicity.

In Table 2, the NAEP data are in the form of the matrix $\mathbf{p}^{(IIO)}$. In this matrix, the rows $i = 0, 1, \ldots, I$ correspond to different levels of the total test score $i = X_+$; the six items $j = 1, \ldots, J$ are ordered by proportion correct $\hat{p}_j$; and each cell of the matrix contains $\hat{p}_{ij}^{(IIO)}$, the proportion of respondents in score group $i = X_+$ answering item j correctly. For each of the seven score groups (rows) in Table 2, representing different levels of the latent trait θ, some of the proportions (indicated in bold) decrease as a function of increasing j, item easiness, so there are apparent violations of invariant item ordering. Again, it is not clear yet whether any of these reflect statistically significant deviations of invariant item ordering.

The order-constrained Bayes inference framework, presented next, provides a basis on which to test data for statistically significant deviations from monotonicity (with $\mathbf{p}^{(M)}$) or invariant item ordering (with $\mathbf{p}^{(IIO)}$). Later, the framework is applied to test the fit of the NAEP data (in Tables 1 and 2) to these two properties, respectively.

Table 2

Proportion Data $\mathbf{p}^{(IIO)}$ Containing the Responses of 100 Examinees to Six Items of the National Assessments of Educational Progress (NAEP) Reading Test

Total score group (X_+)	Item 1	2	3	4	5	6	Number of Respondents
0	.00	.00	.00	.00	.00	.00	3
1	.00	.15	.15	**.00**	.39	**.31**	13
2	**.24**	**.00**	.12	.29	.65	.71	17
3	.18	.18	.55	**.46**	.82	.82	22
4	.44	.64	**.44**	.92	**.72**	.84	25
5	.63	.69	.94	1.00	**.75**	1.00	16
6	1.00	1.00	1.00	1.00	1.00	1.00	4
Item proportion correct	.33	.37	.46	.58	.68	.75	

Note. The proportions in bold indicate suggested areas of the data $\mathbf{p}^{(IIO)}$ that violate invariant item ordering (IIO).

The Order-Constrained Bayes Inference Framework

Corresponding to the elements of a data matrix $(\mathbf{p}, \mathbf{N}, \mathbf{y}) \in \{(\mathbf{p}^{(M)}, \mathbf{N}^{(M)}, \mathbf{y}^{(M)}), (\mathbf{p}^{(IIO)}, \mathbf{N}^{(IIO)}, \mathbf{y}^{(IIO)})\}$ are the parameters:

$$\Theta \quad = (\theta_{ij} \,|\, i = 0, \ldots, I;\ j = 1, \ldots, J) \in (0, 1)^{(I+1)J}, \tag{4}$$

where a single parameter θ_{ij} represents the probability of a correct response by group i on item j. The parameters in Θ are specified to have an order-constrained posterior distribution:

$$\pi\left(\Theta \mid \mathbf{p}\right) = \frac{L\left(\mathbf{p} \mid \Theta\right)\pi\left(\Theta\right)}{\int_\Omega L\left(\mathbf{p} \mid \Theta\right)\pi\left(\Theta\right)d\Theta}, \tag{5}$$

where L refers to the likelihood of the data given the parameters in Θ; $\pi(\Theta)$ is the prior distribution of these parameters, representing the order constraints that restrict Θ to lie within a proper subset Ω of $(0,1)^{(I+1)J}$; and the integral in the denominator refers to the marginal density of a data set $\mathbf{p} \in \{\mathbf{p}^{(M)}, \mathbf{p}^{(IIO)}\}$.

On one hand, the likelihood $L(\mathbf{p} \,|\Theta)$ is assumed to be a product of $(I+1)J$ independent binomial probability mass functions:

$$L\left(\mathbf{p} \mid \Theta\right) = \prod_{i=0}^{I}\prod_{j=1}^{J}\binom{N_{ij}}{y_{ij}}\theta_{ij}^{y_{ij}}\left(1-\theta_{ij}\right)^{N_{ij}-y_{ij}}, \tag{6}$$

with $(\mathbf{N}, \mathbf{y}) \in \{(\mathbf{N}^{(M)}, \mathbf{y}^{(M)}), (\mathbf{N}^{(IIO)}, \mathbf{y}^{(IIO)})\}$. Clearly, (6) assumes independence of the data $\mathbf{p} \in \{\mathbf{p}^{(M)}, \mathbf{p}^{(IIO)}\}$ conditional on the parameters in Θ; in particular, the product operation over the J items directly corresponds to the local independence assumption described in equation (1). On the other hand, the order-constraining prior distribution $\pi(\Theta)$ has the form

$$\pi(\Theta) = \begin{cases} > 0 \text{ iff } \Theta \in \Omega \\ 0 \text{ iff } \Theta \notin \Omega. \end{cases} \tag{7}$$

In this study, two different order-constraining prior distributions are considered for $\pi(\Theta)$, and they are described as follows.

In testing the fit of the observed data $\mathbf{p}^{(M)}$ to monotonicity, the prior distribution $\pi(\Theta_M)$ is implemented for inference with the order-constrained posterior distribution $\pi(\Theta|\ \mathbf{p}^{(M)})$. The monotonicity assumption shared by both the MH model and the DM model implies that θ_{ij} is nondecreasing over the rest score $R_{(j)}$ (Junker, 1993; Junker & Sijtsma, 2000). So with this prior distribution $\pi(\Theta_M)$, θ_{ij} is specified to be nondecreasing over the rest score $i = R_{(j)} \in \{0, 1, \ldots, I = J - 1\}$ by placing the order constraints $\{\theta_{0j} \leq \cdots \leq \theta_{ij} \leq \cdots \leq \theta_{Ij}\}$ for each item $j \in \{1, \ldots, J\}$.

In testing the fit of the observed data $\mathbf{p}^{(IIO)}$ to IIO, the prior distribution $\pi(\Theta_{IIO})$ is implemented for inference with the order-constrained posterior distribution $\pi(\Theta|\mathbf{p}^{(IIO)})$. So with this prior distribution $\pi(\Theta_{IIO})$, θ_{ij} is specified to be nondecreasing over j (i.e., item easiness) by placing the order constraints $\{\theta_{i1} \leq \cdots \leq \theta_{ij} \leq \cdots \leq \theta_{iJ}\}$ for each row i containing respondents with a common total test score X_+. Each row i is assumed to represent the different levels of the latent trait θ because X_+ is an ordinally consistent estimator of θ when monotonicity holds (e.g., Stout, 1990, Theorem 3.2).

Note that the posterior distribution $\pi(\Theta|\ \mathbf{p}^{(IIO)})$ provides a heuristic approach for testing invariant item ordering because it implies linear order constraints between the J items for each row $i = X_+ \in \{0, 1, \ldots, J\}$ corresponding to the value of the total test score. Sijtsma and Junker (1996) consider a more theoretically correct approach to testing IIO, which involves comparing the responses for all pairs of items j,k for each possible *rest score* group $R_{(jk)} = [X_+ - (X_j + X_k)] \in \{0, 1, \ldots, J - 2\}$. However, their approach requires the data analyst to interpret a large number of separate statistical tests—$(J(J - 1)/2)(J - 1)$, to be exact. With this many tests, it is not clear how to summarize the fit of each item to IIO or to globally test the fit of all items to IIO. As shown later in the applications of the Bayes inference framework to real data, in the implementation of the prior distribution $\pi(\Theta_{IIO})$, the resulting posterior inference $\pi(\Theta|\ \mathbf{p}^{(IIO)})$ provides a test to fit each item to IIO and to globally test the fit of all items to IIO (albeit these are heuristic tests of IIO).

Estimating an Order-Constrained Posterior Distribution

The prior (7) leads to the marginal density (integral) in (5) that cannot be evaluated analytically for the order-constrained posterior distribution. However, as Gelfand et al. (1992) show, if the form of the *unconstrained* version of the posterior distribution (5) is known, then the Gibbs sampling algorithm can be routinely applied to estimate an order-constrained posterior distribution (5) while ignoring the intractable integral.

So let $\Theta_u = (\theta_{(u)ij} | i = 0, \ldots, I;\ j = 1, \ldots, J) \in (0, 1)^{(I+1)J}$ denote the set of *unconstrained* parameters. These parameters have the same posterior form as (5), with the exception that the denominator integrates over the entire space $(0,1)^{(I+1)J}$ because the prior distribution $\pi(\Theta_u)$ ignores the constraint Ω. This unconstrained posterior distribution $\pi(\Theta_u|\ \mathbf{p})$ can be modeled in terms of a beta distribution (e.g., Johnson & Kotz, 1970), with a beta density prior $\pi(\theta_{(u)ij}|\ \mathbf{p})$ specified independently for each and every $\theta_{(u)ij}$, as this density is conjugate to the binomial likelihood $L(p_{ij}|\theta_{(u)ij})$ (e.g., Carlin & Louis, 1996, p. 51). In terms of a cumulative distribution function (c.d.f.), the beta posterior of the unconstrained parameter $\theta_{(u)ij}$, fully conditioning out the remaining parameters of Θ_u, is given by

$$F_{ij}(w) = \frac{\Gamma\left(y_{ij} + \alpha_{ij} + N_{ij} - y_{ij} + \beta_{ij}\right)}{\Gamma\left(y_{ij} + \alpha_{ij}\right)\,\Gamma\left(N_{ij} - y_{ij} + \beta_{ij}\right)} \int_0^w \theta_{(u)ij}^{y_{ij}+\alpha_{ij}-1}\left(1 - \theta_{(u)ij}\right)^{N_{ij}-y_{ij}+\beta_{ij}-1} d\theta_{(u)ij}. \quad (8)$$

The symbol Γ refers to the gamma function, the data y_{ij} and N_{ij} are contributions to the binomial likelihood $L(p_{ij}|\theta_{(u)ij})$, and the "shape parameters" $\alpha_{ij}, \beta_{ij} > 0$ are contributions of the beta prior $\pi(\theta_{(u)ij})$, with $w \in [0, 1]$. The shape parameters α_{ij}, β_{ij} may be viewed as pre-experimental information about the number of "successes" and "failures," respectively.

The unconstraining prior distribution $\pi(\Theta_u)$ is specified as being based on no prior information. For each single parameter $\theta_{(u)ij}$, a natural choice of a noninformative prior is given by the uniform distribution on (0,1), obtained by setting $\alpha_{ij} = \beta_{ij} = 1$. But as is well known, the uniform prior is not invariant over one-to-one transformations of $\theta_{(u)ij}$. For example, although the uniform distribution leads to prior $\pi(\theta_{(u)ij})$ that is noninformative for the binomial parameter $\theta_{(u)ij}$, it also leads to a nonuniform and *informative* prior distribution for $\pi(\rho_{ij})$, after $\theta_{(u)ij}$ is reparameterized to the odds ratio $\rho_{ij} = \theta_{(u)ij}/(1 - \theta_{(u)ij})$ (see Robert, 2001, p. 128). So with the interests of specifying a prior distribution for $\pi(\Theta_{(u)})$ that is noninformative for any reparameterization of $\theta_{(u)ij}$, the so-called reference prior is chosen for this study (Berger & Bernardo, 1992; Bernardo, 1979). The reference prior is formulated by separating parameters of interest from nuisance parameters. But when no such ordering can be proposed, as in the current study, a reference noninformative prior can be specified by considering each component of Θ_u seperately (Berger & Bernardo, 1992). Accordingly, with respect to the unconstrained posterior c.d.f. in (8), $\pi(\Theta_u)$ is specified as this reference prior distribution by setting $\alpha_{ij} = \beta_{ij} = 1/2$ for all ij (see Polson & Wasserman, 1990; Yang & Berger, 1997). For each parameter $\theta_{(u)ij}$, it can be shown that this prior distribution is U-shaped over the domain (0,1).

Two Gibbs algorithms were developed to sample from the unconstrained posterior c.d.f. (8) under the reference prior distribution. The first algorithm is used to generate T samples $\left(\Theta^{(t)} \mid t = 1, \ldots, T\right)$ from the order-constrained posterior distribution $\pi(\Theta|\, \mathbf{p}^{(M)})$, according to the informative prior distribution $\pi(\Theta_M)$. The prior $\pi(\Theta_M)$ places each parameter θ_{ij} under the constraint $\theta_{i-1,j} \leq \theta_{ij} \leq \theta_{i+1,j}$ for each item j (with $\theta_{-1,j} = \theta_{i,0} \equiv 0$ and $\theta_{I+1,j} = \theta_{i,J+1} \equiv 1$), so an order-constrained sample from the full-conditional posterior distribution,

$$\pi\left(\theta_{ij} \,\middle|\, \theta_{i-1,j}^{(t)}, \theta_{i+1,j}^{(t)}, \mathbf{p}^{(M)}\right), \quad (9)$$

is generated by:

$$\theta_{ij}^{(t)} = F_{ij}^{-1}\left[F_{ij}\left(\theta_{i-1,j}^{(t)}\right) + u_{ij}^{(t)}\left(F_{ij}\left(\theta_{i+1,j}^{(t)}\right) - F_{ij}\left(\theta_{i-1,j}^{(t)}\right)\right)\right], \quad (10)$$

where $u_{ij}^{(t)}$ is a random draw from (0,1) for parameter ij on iteration t of the Gibbs algorithm. The second Gibbs algorithm is used to generate samples $\left(\Theta^{(t)} \mid t = 1, \ldots, T\right)$ from the order-constrained posterior distribution $\pi(\Theta|\, \mathbf{p}^{(IIO)})$, according to the informative prior distribution $\pi(\Theta_{IIO})$. The prior $\pi(\Theta_{IIO})$ places each parameter θ_{ij} under the constraint $\theta_{i,j-1} \leq \theta_{ij} \leq \theta_{i,j+1}$ for each i (with $\theta_{-1,j} = \theta_{i,0} \equiv 0$ and $\theta_{I+1,j} = \theta_{i,J+1} \equiv 1$), so an order-constrained sample from the full-conditional posterior distribution,

$$\pi\left(\theta_{ij} \,\middle|\, \theta_{i,j-1}^{(t)}, \theta_{i,j+1}^{(t)}, \mathbf{p}^{(IIO)}\right), \quad (11)$$

is generated by:

$$\theta_{ij}^{(t)} = F_{ij}^{-1}\left[F_{ij}\left(\theta_{i,j-1}^{(t)}\right) + u_{ij}^{(t)}\left(F_{ij}\left(\theta_{i,j+1}^{(t)}\right) - F_{ij}\left(\theta_{i,j-1}^{(t)}\right)\right)\right]. \tag{12}$$

In equations (9) through (12), the method of Gelfand et al. (1992, p. 524; from Devroye, 1986, p. 38) is adapted to sample from order-constrained posterior distributions. The following details the two Gibbs algorithms, and both assume appropriate starting values $\Theta^{(t=0)} \in \Omega$.

Algorithm 1: Gibbs Sampler for Testing Monotonicity

Perform the following two steps in iteration $t \in \{1, \ldots, T\}$:
Step 1. For odd i and all j, generate a sample $\theta_{ij}^{(t)}$ from the full-conditional posterior distribution $\pi\left(\theta_{ij}\,\middle|\,\theta_{i-1,j}^{(t-1)}, \theta_{i+1,j}^{(t-1)}, \mathbf{p}^{(M)}\right)$.

Step 2. For even i and all j, generate a sample $\theta_{ij}^{(t)}$ from the full-conditional posterior distribution $\pi\left(\theta_{ij}\,\middle|\,\theta_{i-1,j}^{(t)}, \theta_{i+1,j}^{(t)}, \mathbf{p}^{(M)}\right)$.

Algorithm 2: Gibbs Sampler for Testing Invariant Item Ordering

Perform the following two steps in iteration $t \in \{1, \ldots, T\}$:
Step 1. For all i and odd j, generate a sample $\theta_{ij}^{(t)}$ from the full-conditional posterior distribution $\pi\left(\theta_{ij}\,\middle|\,\theta_{i,j-1}^{(t-1)}, \theta_{i,j+1}^{(t-1)}, \mathbf{p}^{(IIO)}\right)$.

Step 2. For all i and even j, generate a sample $\theta_{ij}^{(t)}$ from the full-conditional posterior distribution $\pi\left(\theta_{ij}\,\middle|\,\theta_{i,j-1}^{(t)}, \theta_{i,j+1}^{(t)}, \mathbf{p}^{(IIO)}\right)$.

As shown above, the "blocking" approach of Liu, Wong, and Kong (1994) is adapted to sample many parameters at a time in each step of the Gibbs algorithms. This is because in the full-conditional posterior distribution of a parameter of interest given all the remaining parameters, denoted by $\pi\left(\theta_{ij}\,\middle|\,\theta_{ij} \notin \Theta, \mathbf{p}\right)$, the conditioning parameters are reduced to a smaller set (in particular, $\{\theta_{i-1,j}, \theta_{i+1,j}\}$ for Algorithm 1 and $\{\theta_{i,j-1}, \theta_{i,j+1}\}$ for Algorithm 2). The above Gibbs algorithms seem faster than the approach of Gelfand et al. (1992), which involves sampling one order-constrained parameter at a time. See see Dunson and Neelon (2003) for other new approaches to sampling order-constrained parameters in the context of generalized linear models.

To estimate an order-constrained posterior distribution $\pi(\Theta|\mathbf{p})$ from a data matrix $\mathbf{p} \in \{\mathbf{p}^{(M)}, \mathbf{p}^{(IIO)}\}$, the task is to repeat the Gibbs sampling algorithm for a "large" number of T times because under mild regularity conditions (see Tierney, 1994, for details), the sequence $\left(\Theta^{(t)} \mid t = 1, \ldots, T\right)$ converges to a sample from the posterior $\pi(\Theta|\mathbf{p})$ as T approaches infinity. From the generated sample of any parameter, $\left(\theta_{ij}^{(t)}|t = 1, \ldots, T\right)$, point estimates of Θ are directly calculated. For example, the posterior mean $\bar{\theta}_{ij}$ is estimated by the sample average, and the .025 and .975 quantiles of the sample bracket a 95% posterior interval of θ_{ij}.

Before calculating *any* estimate from a posterior sample $\left(\Theta^{(t)}|t = 1, \ldots, T\right)$, it is wise to discard the first B "burn-in" samples $\left(\Theta^{(t)}|1 \leq t \leq B < T\right)$ because these samples usually depend on the potentially arbitrary starting values $\Theta^{(t=0)} \in \Omega$ needed to initiate the Gibbs algorithm. Geyer (1992) observed that B is likely to be less than 1% of a T large enough for adequate precision in the posterior estimates. In the next subsection, for notational simplicity, assume T to be the number of Gibbs iterations *after* removing the burn-in samples.

Testing Monotonicity, Invariant Item Ordering, and Local Independence

Given the two Gibbs sampling algorithms that generate samples from $\pi(\Theta^{(M)}| \mathbf{p}^{(M)})$ and from $\pi(\Theta^{(IIO)}| \mathbf{p}^{(IIO)})$, it is possible to test the fit of the data $\mathbf{p} = \mathbf{p}^{(M)}$ to monotonicity and local independence, or test the fit of the data $\mathbf{p} = \mathbf{p}^{(IIO)}$ to invariant item ordering, by comparing $\mathbf{p}$ to the posterior-predictive distribution (Gelman, Meng, & Stern, 1996; Meng, 1994; Rubin, 1984). Conditional on the most recent observation $\mathbf{p} \in \{\mathbf{p}^{(M)}, \mathbf{p}^{(IIO)}\}$, the posterior-predictive probability of a possible future value $\mathbf{p}^{rep}$ (fixing N_{ij} for all i and j) is:

$$\pi\left(\mathbf{p}^{rep} \mid \mathbf{p}\right) = \int_{\Omega} \pi\left(\mathbf{p}^{rep} \mid \Theta\right) \pi\left(\Theta \mid \mathbf{p}\right) d\Theta. \tag{13}$$

The distribution of $\mathbf{p}^{rep}$, via (13), is estimated as a simple by-product of the Gibbs sampler used to estimate the posterior distribution $\pi(\Theta| \mathbf{p})$. Specifically, after the values $\left(\theta_{ij}^{(t)} \mid i = 0, \ldots, I;\ j = 1, \ldots, J\right)$ are sampled in a given iteration t, for each ij, proportions $\left(p_{ij}^{rep(t)} = y_{ij}^{rep(t)} \big/ N_{ij} \mid i = 0, \ldots, I;\ j = 1, \ldots, J\right)$ are independently drawn with sample size N_{ij} and success probability $\theta_{ij}^{(t)}$. The sequence $\left(\mathbf{p}^{rep(t)} \mid t = 1, \ldots, T\right)$ generated in this way provides an estimate of the posterior-predictive distribution (13).

The distribution of any discrepancy statistic can be estimated from the predictive distribution (13) to evaluate the fit of the data $\mathbf{p}^{(M)}$ to monotonicity or evaluate the fit of the data $\mathbf{p}^{(IIO)}$ to invariant item ordering. For $\mathbf{p} \in \{\mathbf{p}^{(M)}, \mathbf{p}^{(IIO)}\}$, a useful choice is the chi-square discrepancy,

$$\chi^2\left(\mathbf{p}\,;\ \Theta\right) = \sum_{i=0}^{I}\sum_{j=1}^{J}\left[\left(N_{ij}p_{ij} - N_{ij}\theta_{ij}\right)^2 \big/ N_{ij}\theta_{ij}\right]. \tag{14}$$

With this, the fit of a data set $\mathbf{p} \in \{\mathbf{p}^{(M)}, \mathbf{p}^{(IIO)}\}$ is checked by the posterior-predictive p value, given by:

$$\begin{aligned} p\text{ value}\left(\mathbf{p} \mid \Theta\right) &= \Pr\left[\chi^2\left(\mathbf{p}^{rep};\ \Theta\right) \geq \chi^2\left(\mathbf{p}\,;\ \Theta\right) \mid \mathbf{p}\right] \\ &= \int\int_{\Omega} I\left[\chi^2\left(\mathbf{p}^{rep};\ \Theta\right) \geq \chi^2\left(\mathbf{p}\,;\ \Theta\right)\right] p\left(\mathbf{p}^{rep} \mid \Theta\right) p\left(\Theta \mid \mathbf{p}\right) d\mathbf{p}^{rep} d\Theta, \end{aligned} \tag{15}$$

where $I(\cdot)$ is the indicator function. Notice that the null distribution of the chi-square statistic is inferred through (15) by integrating $\chi^2(\mathbf{p}^{rep}; \Theta)$ over the posterior distribution $p(\Theta|\mathbf{p})$ and over the posterior- predictive distribution $p(\mathbf{p}^{rep}|\Theta)$. So the p value is calculated by comparing this null distribution to the distribution of the chi-square $\chi^2(\mathbf{p}; \Theta)$ that is based on the observation $\mathbf{p}$, where this distribution is obtained by integrating that chi-square over the posterior $p(\Theta| \mathbf{p})$.

Low p values—say, less than .15—indicate model misfit. As (15) suggests, the posterior-predictive p value is the posterior mean of the classical p value, averaged over the posterior distribution of the model parameters (Meng, 1994, p. 1142). Given the set of Gibbs samples $(\Theta^{(t)}|t = 1, \ldots, T)$, (15) is estimated by

$$\frac{1}{T}\sum_{t=1}^{T} I\left(\chi^2\left(\mathbf{p}^{rep(t)}; \Theta^{(t)}\right) \geq \chi^2\left(\mathbf{p}; \Theta^{(t)}\right)\right).$$

Equations (14) and (15) provide devices for testing the global fit of $\mathbf{p}^{(M)}$ to monotonicity and for testing the global fit of $\mathbf{p}^{(IIO)}$ to invariant item ordering. Of course, the discrepancy statistic (14) can be easily revised to evaluate the fit of any region of a data set $\mathbf{p} \in \{\mathbf{p}^{(M)}, \mathbf{p}^{(IIO)}\}$. For example, for testing the fit of an item j, (14) is modified to sum over I score groups within j. Similarly, to test the fit of a single cell ij, (14) is revised to focus on that cell (i.e., using no summation over I or over J).

As already mentioned, one important assumption of item response theory is local independence, and it is implied by the likelihood function in (6). A statistic known to test local independence in parametric IRT models is Yen's Q statistic (e.g., Yen, 1993), which was formulated within the frequentist framework of classical statistics. This statistic investigates possible sources of local dependence by analyzing residual correlations between all pairs of items. The Q statistic is adapted into the current Bayesian framework to test for local independence on the data $\mathbf{p}^{(M)}$, where the possible rest scores $i = R_{(j)} = 0, 1, \ldots, I = J - 1$ are treated as ordinally consistent discrete estimators of the latent trait θ (e.g., Stout, 1990, Theorem 3.2). The Q statistic is adapted into the Bayesian framework by calculating the residual correlation between items j and k, given by $Q_{jk}^{(t)} = Corr\left[\mathbf{r}_j^{(t)}, \mathbf{r}_k^{(t)}\right]$, where $\mathbf{r}_j^{(t)} = \left(\hat{p}_{ij} - \theta_{ij}^{(t)} \mid i = 0, \ldots, I\right)$, and also calculating the same statistic based on posterior-predictive data replications, $Q_{jk}^{rep(t)} = Corr\left[\mathbf{r}_j^{*(t)}, \mathbf{r}_k^{*(t)}\right]$, with $\mathbf{r}_j^{rep(t)} = \left(p_{ij}^{rep(M)(t)} - \theta_{ij}^{(t)} \mid i = 0, \ldots, I\right)$. Then, a two-tailed posterior-predictive p value for the test of local independence between any pair of items j and k is estimated by $\frac{1}{T} I \left(\left|Q_{jk}^{rep(t)}\right| \geq \left|Q_{jk}^{(t)}\right|\right)$.

For similar applications of the posterior-predictive framework in a different psychometric context, see Karabatsos and Batchelder (2003). More generally, in the posterior-predictive approach (via Gibbs sampling) (e.g., equation (15)), the p value can be based on the estimated sampling distribution of any arbitrary discrepancy statistic (e.g., chi-square or Yen's Q). This estimated sampling distribution could be any shape, and one need not a priori assume a particular sampling distribution for a given statistic. These features are important to rigorously check the fit of the data to a given set of order constraints, such as the order constraints implied by the prior distribution $\pi(\Theta_M)$ of monotonicity, or the order constraints implied by the prior distribution $\pi(\Theta_{IIO})$ of invariant item ordering.

Application of the Bayes Framework to Test Data

The order-constrained Bayes framework is now applied to analyze the NAEP data matrices $\mathbf{p}^{(M)}$ and $\mathbf{p}^{(IIO)}$ of Tables 1 and 2, respectively, to test the fit of the data to the monotonicity, invariant item ordering, and local independence assumptions of NIRT. This data analysis pertains to the MH and DM models of nonparametric item response theory.

To analyze the NAEP data matrix $\mathbf{p}^{(M)}$ (Table 1), the posterior distribution $\pi(\Theta|\mathbf{p}^{(M)})$ was estimated under the order-constraining prior $\pi(\Theta_M)$ pertaining to monotonicity, based on 15,000 samples generated by the first Gibbs sampler (after discarding the first 5,000 "burn-in" Gibbs samples). Table 3 contains, for each parameter θ_{ij}, estimates of the posterior mean and of the 95% posterior interval (based on the estimated 2.5% and 97.5% quantiles of the posterior). Within each of the items (columns), the posterior means are nondecreasing over rest score $i = R_{(j)}$ because of the order-constraining prior $\pi(\Theta_M)$.

Table 4 contains the posterior-predictive p value for each item and each cell ij (using the chi-square discrepancy measure) in testing the fit of the NAEP data matrix $\mathbf{p}^{(M)}$ (Table 1) to monotonicity. As shown in the table, Item 2 significantly violates monotonicity (p = .07), as do cells 0,2 and 5,5 (p = .09 and .08, respectively). The global p value was .17, so it seems the NAEP data matrix $\mathbf{p}^{(M)}$ does not violate monotonicity.

Table 5 contains the results of the posterior-predictive p values based on Yen's Q statistic. There is no evidence in the NAEP data matrix $\mathbf{p}^{(M)}$ that any of the items violate local independence; .36 is the lowest p value among all the item pairs.

Table 3

The Posterior Mean Estimates of the Parameters Constrained to Be Nondecreasing Over the Rest Score $R_{(j)}$, Along With Estimates of the 95% Posterior Interval (in parentheses) Based on the 2.5% and 97.5% Posterior Quantiles

	Item					
Rest score	1	2	3	4	5	6
0	.05 (.00, .20)	.09 (.02, .21)	.17 (.04, .35)	.07 (.00, .28)	.45 (.02, .63)	.44 (.19, .65)
1	.19 (.07, .32)	.13 (.03, .26)	.25 (.10, .41)	.28 (.12, .45)	.57 (.42, .71)	.59 (.41, .74)
2	.26 (.15, .39)	.23 (.10, .39)	.41 (.27, .54)	.46 (.30, .63)	.67 (.55, .77)	.73 (.60, .85)
3	.36 (.24, .49)	.43 (.30, .56)	.50 (.37, .63)	.66 (.51, .79)	.72 (.61, .81)	.81 (.70, .90)
4	.44 (.32, .58)	.53 (.39, .68)	.58 (.45, .71)	.86 (.71, .97)	.75 (.65, .84)	.86 (.76, .94)
5	.55 (.38, .75)	.63 (.46, .81)	.81 (.57, .98)	.96 (.84, 1.00)	.80 (.69, .91)	.96 (.86, 1.00)

Table 4

Estimated Posterior-Predictive p Values That Detail the Fit of the Data $\mathbf{p}^{(M)}$ to Monotonicity (According to the Chi-Square Discrepancy Measure)

	Item					
Rest score	1	2	3	4	5	6
0	.98	**.09**	.30	.94	.49	.60
1	.60	.39	.56	.62	.70	.63
2	.72	.62	.62	.62	.51	.63
3	.66	.57	.70	.58	.42	.69
4	.68	.69	.54	.64	.33	.50
5	.49	.39	.74	.98	**.08**	.98
Item fit	.67	**.07**	.35	.62	.22	.57

Note. The low p values in bold indicate the areas of the data $\mathbf{p}^{(M)}$ that significantly violate monotonicity (either a cell ij or an item j).

Table 5

Estimated Posterior-Predictive p Values, Based on Yen's Q Local-Independence Tests Between All Pairs of Items, With Respect to the Data Set $\mathbf{p}^{(M)}$

	Item				
Item	2	3	4	5	6
1	.61	.61	.52	.47	.54
2		.58	.55	.50	.58
3			.55	.39	.36
4				.52	.53
5					.37

Note. The low p values in bold indicate the item pairs that significantly violate local independence.

The posterior-predictive p values of Table 6 (according to the chi-square discrepancy measure) help address the question as to whether any part of the NAEP data matrix $\mathbf{p}^{(IIO)}$ (Table 2) has statistically significant violations of invariant item ordering. These p values are based on the posterior distribution $\pi(\Theta|\, \mathbf{p}^{(IIO)})$, estimated under the order-constraining prior $\pi(\Theta_{IIO})$ pertaining to invariant item ordering. The posterior was estimated from 15,000 Gibbs samples, after discarding the first 5,000 "burn-in" Gibbs samples. In Table 6, it is shown that Item 1 does not conform well to invariant item ordering ($p = .18$), as do cells 2,1, 4,4, and 5,5 ($p = .12$, .14, and .10, respectively). The global p value was .16, so the NAEP data matrix $\mathbf{p}^{(IIO)}$ globally does not violate invariant item ordering.

According to Tables 1 and 2, the NAEP data apparently displayed violations of monotonicity and invariant item ordering. However, the p values pertaining to the global tests of monotonicity and invariant item ordering indicate that, in general, both the MH and DM models of nonparametric item response theory fit the NAEP data set. So there is evidence in the NAEP data that each respondent's latent trait θ is measurable by the total score X_+, and furthermore, there is evidence that the difficulty order among the items $j = 1, \ldots, J$, which was hypothesized to be $1 \leq 2 \leq 3 \leq 4 \leq 5 \leq 6$, is invariant over all values of θ.

Conclusions

This study demonstrated that the order-constrained Bayes framework, for dichotomous item response data, provides a general way to routinely perform statistical inference with either the MH or DM model of NIRT. In fact, the corresponding author has written a program, in S-PLUS (1995) code, that can perform order-constrained data analysis of binomial data for an $I \times J$ matrix of any size. A free copy of the program can be obtained by e-mailing the author (georgek@uic.edu).

Two extensions of the order-constrained Bayes framework should be investigated in future studies of the framework, and they are described as follows. First, it would be useful to consider other methods of estimating posterior-predictive p values, which may offer more power to detect data

Table 6

Estimated Posterior-Predictive p Values That Detail the Fit of the Data $\mathbf{p}^{(IIO)}$ to Invariant Item Ordering (IIO) (According to the Chi-Square Discrepancy Measure)

Total score group (X_+)	Item 1	2	3	4	5	6
0	1.00	1.00	.98	.91	.75	.45
1	.94	.46	.72	.20	.61	.51
2	**.12**	.25	.62	.62	.61	.60
3	.59	.59	.49	.49	.59	.59
4	.60	.36	.26	**.14**	.39	.58
5	.61	.68	.44	.29	**.10**	.90
6	.47	.76	.92	.98	1.00	1.00
Item fit	**.18**	.26	.47	.21	.40	.37

Note. The low p values in bold indicate the areas of the data $\mathbf{p}^{(IIO)}$ that significantly violate invariant item ordering (either a cell *ij* or an item j).

violations of a nonparametric item response theory model (Bayarri & Berger, 2000; Bayarri & Castellanos, 2000; Evans, 1997; Robins, van der Vaart, & Ventura, 2000). Such other methods are already being considered by Karabatsos (2003, in press) in the context of testing axioms of measurement and decision.

Second, although the Bayes inference framework model was demonstrated for the case of dichotomous item scores, it is possible to extend the framework for the analysis of polytomous items. Consider any particular sequence of response thresholds $\mathbf{c} = (c_1, \ldots, c_j, \ldots, c_J)$ over the J test items, where each item j can have several thresholds $c_j \in \left\{c_j, c'_j, c''_j, \ldots\right\}$. A threshold may refer to an ordinal response category (i.e., Likert response) or a "cutoff" point in the case when item responses are from a continuous scale. Under the nonparametric graded-response model (the polytomous response analog of the MH model), the monotonicity of $\Pr(X_j > c_j|\theta)$ over θ does imply the monotonicity of $\Pr(X_j > c_j|R_{(j),\mathbf{c}})$ over all possible $\mathbf{c}$, where $R_{(j),\mathbf{c}} = \sum_{k\neq j}^{J} Y_{kc_j}$ is the rest score corresponding to a particular choice of the sequence $\mathbf{c}$, and $Y_{jc_j} = I\left(X_j > c_j\right)$ is the dichotomized response variable (see Junker, 1991; Junker & Ellis, 1997; Junker & Sijtsma, 2000, p. 80; Samejima, 1969; Scheiblechner, 1995). So in applying the Bayes framework to polyomous response data, given a choice of $\mathbf{c}$, the parameter θ_{ijc_j} is estimated to be nondecreasing over $i = R_{(j),\mathbf{c}}$, where θ_{ijc_j} is the probability of an item response that equals or exceeds a threshold c_j. Furthermore, Bayesian tests of the isotonic ordinal probabilistic model (Scheiblechner, 1995) are possible by estimating θ_{ijc_j} to be nondecreasing over j for each test score group $i = X_{+,\mathbf{c}}$. Of course, in all these proposed Bayesian tests of polytomous item responses, one would need to consider all the many possible sequences of thresholds $\mathbf{c} = (c_1, \ldots, c_j, \ldots, c_J)$. Each of these tests, however, should be stochastically dependent on one another over the possible sequences. So perhaps a careful search strategy could be used to find the few sequences that are nonredundant (Junker & Sijtsma, 2000, p. 80).

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions*. New York: John Wiley.

Bayarri, M. J., & Berger, J. O. (2000). *P*-values for composite null models. *Journal of the American Statistical Association*, *95*, 1127-1142.

Bayarri, M. J., & Castellanos, M. E. (2000). *A comparison of p-values for goodness-of-fit checking* (Working Paper Series, 00-40, Institute for Statistics and Decision Sciences). Durham, NC: Duke University. Available: http://ftp.isds.duke.edu/WorkingPapers/00-40.ps

Berger, J. O., & Bernardo, J. M. (1992). On the development of the reference prior method. In J. M. Bernardo, J. O. Berger, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 4*. Oxford, UK: Oxford University Press.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society, Series B*, *41*, 113-147.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis* (Reprint of 1st ed.). New York: hapman & Hall/CRC.

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology*, *44*, 315-341.

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer-Verlag.

Dunson, D. B., & Neelon, B. (2003). Bayesian inferences on order constrained parameters in generalized linear models. *Biometrics*, *5*, 286-295.

Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics: Theory and Methods*, *26*, 1125-1143.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*, 523-532.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistic Sinica*, *6*, 733-807.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, *7*, 473-511.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383-392.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331-347.

Hoijtink, H. J. A., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, *62*, 171-189.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics*, *14*, 1523-1543.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika*, *59*, 77-79.

Johnson, N. L., & Kotz, S. (1970). *Continuous univariate distributions* (Vol. 2). Boston: Houghton-Mifflin.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, *56*, 255-278.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Annals of Statistics*, *21*, 1359-1378.

Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *Annals of Statistics*, *25*, 1327-1343.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65-81.

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, *25*, 211-220.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, *2*, 389-423.

Karabatsos, G. (2003). *A Bayesian Dirichlet model for testing deterministic axioms of measurement and decision*. Manuscript under review.

Karabatsos, G. (in press). Additivity tests. In B. Everitt & B. C. Howell (Eds.), *Encyclopedia of behavioral statistics*. New York: John Wiley.

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain Monte-Carlo estimation for test theory without an answer key. *Psychometrika*, *68*, 373-389.

Karabatsos, G., & Ullrich, J. (2002). Enumerating and testing conjoint measurement models. *Mathematical Social Sciences*, *43*, 487-505.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.

Liu, J. S., Wong, W. H., & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, *81*, 27-40.

Meng, X.-L. (1994). Posterior predictive p-values. *Annals of Statistics*, *22*, 1142-1160.

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika*, *30*, 419-440.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Paris: Mouton.

Mokken, R. J., & Lewis, C. (1982). A non-parametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417-430.

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

Polson, N., & Wasserman, L. (1990). Prior distributions for the bivariate binomial. *Biometrika*, *77*, 901-904.

Robert, C. P. (2001). *The Bayesian choice: From decision-theoretic foundations to computer implementation* (2nd ed.). New York: Springer.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference*. New York: John Wiley.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). The asymptotic distribution of *p*-values for composite null models. *Journal of the American Statistical Association*, *95*, 1143-1156.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151-1172.

S-PLUS. (1995). *S-PLUS documentation*. Seattle, WA: Statistical Sciences.

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

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models. *Psychometrika*, *60*, 281-304.

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models. *Psychometrika*, *64*, 295-316.

Sijtsma, K. (1988). *Contributions to Mokken's nonparametric item response theory*. Amsterdam: Free University Press.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, *22*, 3-31.

Sijtsma, K., & Junker, B.W. (1996). A survey for theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*, 79-105.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293-325.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22*, 1701-1728.

Van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika*, *67*, 519-538.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement*, *25*, 283-294.

Yang, R., & Berger, J. O. (1997). *A catalog of noninformative priors* (Working Paper Series, 97-42, Institute for Statistics and Decision Sciences, Duke University) [Online]. Available: http://ftp.isds.duke.edu/WorkingPapers/97-42.html

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213.

Acknowledgments

This research is supported by National Science Foundation grant SES-0242030 and in part by a Spencer Foundation research grant SG20010020. The authors thank Brian Junker for several helpful conversations and the three anonymous referees for many helpful comments on earlier versions of this manuscript.

Author's Address

George Karabatsos, University of Illinois–Chicago, College of Education, 1040 W. Harrison St. (MC 147), Chicago, IL 60607; phone: (312) 413-1816; fax: (312) 996-5651; e-mail: georgek@uic.edu.