PSYCHOMETRIKA—VOL. 68, NO. 3, 373–389
SEPTEMBER 2003

MARKOV CHAIN ESTIMATION FOR TEST THEORY WITHOUT AN ANSWER KEY

GEORGE KARABATSOS

UNIVERSITY OF ILLINOIS, CHICAGO

WILLIAM H. BATCHELDER

UNIVERSITY OF CALIFORNIA, IRVINE

This study develops Markov Chain Monte Carlo (MCMC) estimation theory for the General Condorcet Model (GCM), an item response model for dichotomous response data which does not presume the analyst knows the correct answers to the test a priori (answer key). In addition to the answer key, respondent ability, guessing bias, and difficulty parameters are estimated. With respect to data-fit, the study compares between the possible GCM formulations, using MCMC-based methods for model assessment and model selection. Real data applications and a simulation study show that the GCM can accurately reconstruct the answer key from a small number of respondents.

Key words: consensus theory, Bayesian inference, Markov Chain Monte Carlo, posterior predictive model evaluation, Bayesian model selection.

Introduction

Cultural Consensus Theory (CCT) models are developed for situations where the researcher can write test or questionnaire items for a group of respondents who share common knowledge. Unlike traditional applications of Item Response Theory (IRT), the CCT researcher does not assume a priori knowledge of the "true" or "correct" response category for each item. Instead, the answer key is treated as a set of parameters representing the cultural beliefs of a respondent group. Thus, the data for CCT analysis are the respondents' un-scored item responses. An additional practical feature is that often CCT models can lead to accurate answer key estimates, its main goal, with very small numbers of respondents (Batchelder & Romney, 1988), whereas accurate IRT inference concerning respondent ability and item difficulty parameters may require hundreds of respondents (e.g., Lord, 1983). CCT models have been successfully applied in anthropology, social psychology, cross-cultural psychology, social networks, and the analysis of expert opinion (see the review by Romney & Batchelder, 1999).

Batchelder and Romney (1986, 1988, 1989) and Romney, Weller, and Batchelder (1986) proposed and developed a model for dichotomous (Yes/No) items, called the General Condorcet Model (GCM). The GCM was motivated by earlier work modeling dichotomous items in information pooling problems (Grofman & Owen, 1986). In the cited references, Batchelder and Romney introduced respondent "ability" and respondent guessing bias parameters into the model using an approach from signal detection theory (e.g. Green & Swets, 1966), and provided substantial rationale for the model. They also proved (Observation 5, Batchelder & Romney, 1988)

This study was supported in part by Spencer Foundation grant SG2001000020, George Karabatsos, Principal Investigator, and also in part by NSF Renewal Grant SES-0001550 to A.K. Romney and W.H. Batchelder, Co-Principal Investigators. The second author acknowledges the kind support of the Santa Fe Institute, where he worked on aspects of this paper as a Visiting Professor in the fall of 2001. Both authors appreciate the detailed comments offered by the Editor and two referees on an earlier version of the manuscript. Requests for reprints should be sent to George Karabatsos, University of Illinois-Chicago, College of Education, 1040 W. Harrison Street (MC 147), Chicago, IL 60607, E-Mail: georgek@uic.edu

0033-3123/2003-3/2001-0887-A $00.75/0

that the GCM is structurally isomorphic to the two-class latent structure model (Clogg, 1981; Lazarsfeld & Henry, 1968, chap. 2) with the roles of respondents and items interchanged.

Batchelder and Romney (1988) examined two versions of the GCM. They developed method-of-moments estimation for the first version that allows respondent inhomogeneity in ability, but assumes homogeneous item difficulties and no guessing response bias. They proposed a more general GCM (described later) that adds inhomogeneity in item difficulty, however, estimation theory for this model awaited further development. Despite the availability of standard IRT estimation procedures (e.g., Baker, 1992), developing estimation theory for the GCM is not straightforward, because the unknown answer key is discrete (two-valued). For example, simulated annealing (Aarts & Kours, 1989) was applied to estimate a GCM version (Batchelder, Kumbasar, & Boyd, 1997) to overcome the combinatorially explosive number of possible answer keys.

Fortunately, Markov Chain Monte Carlo (MCMC) estimation theory provides a flexible framework to handle the estimation of even highly complex models (e.g., Carlin & Louis, 1998). Recently, Patz and Junker (1999) introduce a general MCMC estimation procedure for IRT, which can be applied to complex IRT models.

The general objective of the current study is to use MCMC's flexibility to further examine the GCM family. This study develops a MCMC estimation algorithm for the GCM, and shows that the algorithm provides a natural way to estimate a model involving both discrete and continuous parameters. The study also presents MCMC-based statistical methods for testing GCM assumptions, in addition to performing model selection between the GCM versions. The GCM and MCMC framework is examined through applications to two real data sets as well as a simulation study. Before these applications, it is first necessary in the next section to review formal aspects of the GCM, and for the section that follows, to describe the MCMC statistical analysis framework.

The Single-Culture GCM for Dichotomous Data

Two-way designs represent a common data structure in CCT, containing N respondents, $\{i = 1, \ldots, N\}$, and M questionnaire items, $\{k = 1, \ldots, M\}$. Three classes of random variables describe the case of dichotomous test items:

1. *Response profile:* $\mathbf{X} = \{X_{ik}; i = 1, \ldots, N, k = 1, \ldots, M\}$.

$$X_{ik} = \begin{cases} 1 & \text{if respondent } i \text{ responds "Yes" to item } k \\ 0 & \text{if respondent } i \text{ responds "No" to item } k. \end{cases} \tag{1}$$

2. *Answer key:* $\mathbf{Z} = (Z_1, \ldots, Z_M)$.

$$Z_k = \begin{cases} 1 & \text{if correct response to item } k \text{ is "Yes"} \\ 0 & \text{if correct response to item } k \text{ is "No".} \end{cases} \tag{2}$$

3. *Performance profile:* $\mathbf{Y} = \{Y_{ik}; i = 1, \ldots, N, k = 1, \ldots, M\}$

$$Y_{ik} = \begin{cases} 1 & \text{if respondent } i \text{ is "correct" on item } k,\ X_{ik} = Z_k \\ 0 & \text{if respondent } i \text{ is "incorrect" on item } k,\ X_{ik} \neq Z_k. \end{cases} \tag{3}$$

The performance profile in (3) does not refer to correct and incorrect in the "objective" sense, but instead refers to the consistency of response X_{ik} to the item key Z_k believed by the respondent group. In the CCT paradigm the analyst is only provided $\mathbf{X}$, not $\mathbf{Y}$, and the task is to use a model to estimate the answer key $\mathbf{Z}$, where $\mathbf{Z}$ represents the respondents' consensus belief about the correct item responses. This contrasts with typical IRT applications, where the analyst assumes a-priori knowledge of $\mathbf{Z}$ and $\mathbf{Y}$, and uses it to estimate IRT model parameters, such as respondent ability and item difficulty.

The following axioms describe the GCM (Batchelder & Romney, 1989):

Axiom 1: Single culture. All respondents belong to the same cultural group, characterized by a correct answer $z_k \in \{0, 1\}$ for each item $k = 1, \ldots, M$.

Axiom 2: Conditional independence. The respondent-item response random variables satisfy conditional independence, given by

$$\Pr[\mathbf{X} = \mathbf{x}|\mathbf{Z} = \mathbf{z}] = \prod_{i=1}^{N}\prod_{k=1}^{M} \Pr(X_{ik} = x_{ik}|Z_k = z_k), \tag{4}$$

for all possible response profile matrices $\mathbf{x}$ and all possible answer keys $\mathbf{z}$.

Axiom 3: Item inhomogeneity. Each respondent $i = 1, \ldots, N$ has latent hit rates H_{ik} and false-alarm rates F_{ik} that vary over M items, where

$$\Pr[X_{ik} = x_{ik}|Z_k = z_k] = \begin{cases} H_{ik} & \text{if } z_k = 1 \\ F_{ik} & \text{if } z_k = 0, \end{cases} \tag{5}$$

subject to the constraint $0 < F_{ik} < H_{ik} < 1$.

The model, as described by the axioms, is not identifiable because it has been intentionally overparameterized to permit various useful submodels to be defined. The constraint, $F_{ik} < H_{ik}$, is needed to identify some of the submodels (see Batchelder & Romney, 1988, Observation 4), described later.

As is well known in models of signal detection (Swets, 1996), the hidden variable of response bias renders it difficult to directly interpret data in terms of hit and false-alarm rates. Consequently, a special re-parameterization of the model is considered (Batchelder & Romney, 1988, 1989) by introducing two new parameters. First, $D_{ik} \in (0, 1)$ is the probability that respondent i "knows" the correct response to item k, and $g_i \in (0, 1)$ is the probability of guessing "Yes" when the correct answer is unknown. The model is defined and restricted by:

$$H_{ik} = D_{ik} + (1 - D_{ik})g_i, \tag{6a}$$

$$F_{ik} = (1 - D_{ik})g_i, \tag{6b}$$

where (6) is a version of the so-called double-high threshold model (Macmillan & Creelman, 1991). Specifically, (6a) is the probability of a correct response to an item with a correct answer of "Yes" ($Z_k = 1$). This is modeled as a function of the probability of knowing the correct answer, D_{ik}, and the probability of not knowing the correct answer but obtaining it through guessing, $(1 - D_{ik})g_i$. On the other hand, (6b) is the probability of incorrectly guessing "Yes" when the correct answer is "No" ($Z_k = 0$). Note that (6) does satisfy the restriction $0 < F_{ik} < H_{ik} < 1$.

Relative to NM observations of $\mathbf{X}$, the model in (6) has $N(M + 1)$ continuous parameters, and M answer key parameters. Batchelder and Romney (1988) develop and apply a simplified model $\{H_{ik} = H_i, F_{ik} = F_i; i = 1, \ldots, N, k = 1, \ldots, M\}$ that postulates item homogeneity. They also note in passing that (6) can be further specified while retaining item inhomogeneity, with the formulation:

$$D_{ik} = \frac{\theta_i(1 - \delta_k)}{\theta_i(1 - \delta_k) + (1 - \theta_i)\delta_k}, \tag{7}$$

for $i = 1, \ldots, N$ and $k = 1, \ldots, M$, where $\theta_i, \delta_k \in (0, 1)$ are the respondent ability parameter and item difficulty parameters, respectively. Equation (7) is a version of the Rasch model (Roskam & Jansen, 1984) studied by Crowther, Batchelder, and Hu (1995) in detail. As with all other parameterizations of the Rasch (1960) model (e.g., logistic), (7) is nonidentifiable, as more

than one pair of values $\{\theta_i, \delta_k\}$ can yield the same $D_{ik} \in (0, 1)$. The current study identifies (7) by setting $\delta_k = \frac{1}{2}$ for a single chosen item $k \in \{1, \ldots, M\}$, though any single item can be selected and set to a constant in (0, 1).

The fully parameterized submodel in (6) and (7) is called the GCM3 and has parameters

$$\theta = \{\theta_1, \ldots, \theta_N\}, g = \{g_1, \ldots, g_N\}, \quad \text{and} \quad \delta = \{\delta_1, \ldots, \delta_M\},$$

relative to the NM observations of $\mathbf{X}$. This article also investigates three nested versions of the GCM3. The GCM2g assumes neutral guessing bias, $\{g_i = .5; i = 1, \ldots, N\}$. The GCM2$\delta$ assumes item homogeneity by setting $\{\delta_k = d \in (0, 1); k = 1, \ldots, M\}$, and the GCM1, employs the restriction $\{g_i = .5, \delta_k = d \in (0, 1); i = 1, \ldots, N, k = 1, \ldots, M\}$. In the latter two models, it is useful to set $d = .5$, because then (7) yields $D_{ik} = \theta_i$, convenient for the interpretation of ability.

As a matter of simplification, (6a) and (6b) are combined to yield a GCM expression of the probability of a correct response to an item, p_{ik}, as follows:

$$p_{ik} = \Pr[X_{ik} = Z_k | Z_k, \theta_i, g_i, \delta_k] = D_{ik}^{Z_k} + g_i(1 - D_{ik})(2Z_k - 1), \tag{8}$$

where in (8), D_{ik} is a function of θ_i and δ_k through (7). In (8), p_{ik} is expressed in terms of the answer key parameters and the other parameters of the model. The form emphasizes that the GCM directly uses $\mathbf{X}$, instead of $\mathbf{Y}$, in estimation of the model parameters.

MCMC Estimation Theory for the GCM

Bayes Inference of the GCM3

The paper proposes Bayesian methods to estimate the various GCM submodels by employing an approach using MCMC. It is natural to incorporate Bayes theorem into the MCMC estimation framework, though MCMC can also benefit non-Bayesian statistical inference (Geyer, 1996). This paper adapts a Bayesian perspective for GCM estimation, because it can handle both uninformative and informative priors, either which may be important in a particular GCM application. Also MCMC allows one to employ modern methods of Bayesian model assessment and model selection, described later.

The GCMs joint posterior distribution for $\mathbf{Z}$ and $\Omega = \{\theta, g, \delta\}$ is

$$p(\mathbf{Z}, \Omega | \mathbf{X}) = \frac{L(\mathbf{X}|\mathbf{Z}, \Omega)\pi(\mathbf{Z}, \Omega)}{\int L(\mathbf{X}|\mathbf{Z}, \Omega)\pi(\mathbf{Z}, \Omega)d(\mathbf{Z}, \Omega)}, \tag{9}$$

where $L(\mathbf{X}|\mathbf{Z}, \Omega)$ refers to the GCM likelihood, $\pi(\mathbf{Z}, \Omega)$ is the joint prior distribution, and the normalizing constant in the denominator is the marginal distribution of $\mathbf{X}$.

In (9), under Axioms 1 and 2, the GCMs likelihood is given by:

$$L(\mathbf{X}|\mathbf{Z}, \Omega) = \prod_{i=1}^{N}\prod_{k=1}^{M} p_{ik}^{X_{ik}Z_k + (1-X_{ik})(1-Z_k)}(1 - p_{ik})^{X_{ik}(1-Z_k) + (1-X_{ik})Z_k}, \tag{10}$$

where p_{ik} is given by (8). The exponent on p_{ik} in (10) is 1 for a "correct" response, and 0 otherwise. Similarly, the exponent on $(1 - p_{ik})$ indicates an incorrect response. These exponents enable simultaneous estimation of the GCM without requiring a priori knowledge of $\mathbf{Y}$, the performance profile.

The current study specifies the prior distribution of the answer key by independent item priors, with marginal distribution $\pi(Z_k) = \phi_k^{Z_k}(1 - \phi_k)^{1-Z_k}$, where $\phi_k = \Pr[Z_k = 1], k = 1, \ldots, M$. The current study sets $\phi_k = \frac{1}{2}$ for all items, which yields a noninformative prior for the answer key $\mathbf{Z}$. The rationale is not to suppose any prior knowledge about the correct answers,

but to let the data "speak for themselves". For example, an anthropologist might give a set of questions in an exotic language to a person who will analyze the data with the GCM. The analyst is not thought to peruse (or even read or know) the content of the questions, so the Bernoulli prior probability $\frac{1}{2}$ represents this view. Of course, informative priors on $\mathbf{Z}$ can be useful when, for instance, prior information is available on the respondents' beliefs.

This study also employs independent priors for the parameters in Ω, though dependence might be useful when, for example, respondents with more knowledge (θ_i) tend to have larger guessing biases (g_i). Assuming independence, the beta distribution may be specified as the marginal for each GCM parameter $\gamma \in \Omega$, as it is the natural conjugate prior for a random variable in (0, 1) (e.g., Carlin & Louis, 1998). The beta density has a flexible form (Johnson & Kotz, 1970) given by:

$$\pi(\gamma) = \frac{\gamma^{a-1}(1-\gamma)^{b-1}}{B(a,b)} \tag{11}$$

The denominator $B(a, b)$ is the so-called beta function, and $a, b > 0$ govern the shape of the beta distribution, which can vary over each element in Ω.

This study most often employs noninformative (uniform, $a = b = 1$) beta priors on all the respondent ability, respondent bias, and item difficulty parameters. Of course, these noninformative priors on the model parameters result in a prior on the H_{ik} and F_{ik} which satisfies neither independence nor marginal uniformity, because in (6) and (7) they are nonlinearly related to the GCM3 parameters. From a psychological perspective, guessing bias, the g_i, abilities, the θ_i, and difficulties, the δ_k, can be viewed as substantively separable. They are so because independent experimental manipulations, that are typical of studies in signal detection (Swets, 1996), could effect one parameter without having an effect on the others (e.g., the base rate of "yes" items, more education, or difficult item wording). Psychologically, by all accounts, parameter classes like H_{ik} and F_{ik} depend on the same underlying processes (guessing and ability), and this might make one more skeptical about postulating independent uniform priors for them. The current authors believe that independent noninformative beta priors are a reasonable choice, given the substantive considerations mentioned, and especially when there is no prior information available on the latent parameters. However, for pragmatic reasons, this study explores informative priors in a couple of GCM applications.

The GCM MCMC Algorithm

The posterior distribution in (9) has a complicated form, and it seems impossible to either derive closed-form estimators, or even to employ direct simulation methods to generate independent samples from $p(\mathbf{Z}, \Omega|\mathbf{X})$. An alternative approach to model estimation involves implementing Markov Chain Monte Carlo (MCMC) to generate T dependent samples $\{\{\mathbf{Z}, \Omega\}^{(1)}, \{\mathbf{Z}, \Omega\}^{(2)}, \ldots, \{\mathbf{Z}, \Omega\}^{(t)}, \ldots, \{\mathbf{Z}, \Omega\}^{(T)}\}$ from the posterior distribution, with each iterate $\{\mathbf{Z}, \Omega\}^{(t)}$ randomly generated as a function of the previous iterate $\{\mathbf{Z}, \Omega\}^{(t-1)}$. Comprehensive treatments of MCMC inference are found in many recent texts (e.g., Carlin & Louis, 1998; Gilks et al., 1996).

This section develops a general MCMC algorithm for estimating the posterior distribution of the GCM. Under certain conditions such as irreducibility and geometric ergodicity (see Tierney, 1994, for more details), if the number of iterations T is large enough, then the set of observations from the chain $\{\{\mathbf{Z}, \Omega\}^{(t)}; t = 1, \ldots, T\}$ leads to an approximate sample from $p(\mathbf{Z}, \Omega|\mathbf{X})$.

In a single sampling iteration t, the MCMC estimation algorithm for the GCM3 entails four steps, each containing two sub-steps (see Patz & Junker, 1999, for a related algorithm). Let $\theta_{-i} = \{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_N\}$, with g_{-i}, δ_{-k}, and $\mathbf{Z}_{-k}$ similarly defined. Also, the respondent and item response vectors are given by $\mathbf{X}_i = (X_{i1}, \ldots, X_{iM})$ and $\mathbf{X}_k = (X_{1k}), \ldots, X_{Nk})$, respectively. The following MCMC algorithm is introduced, which adds an "answer key sampler" for Step 1.

Step 1. Draw $\mathbf{Z}^{(t)} \sim p(\mathbf{Z}|\Omega^{(t-1)}, \mathbf{X})$, where $p_{ik}^{(t)} = \Pr[X_{ik} = Z_k|\{Z_k, \theta_i, g_i, \delta_k\}^{(t-1)}]$.
(a) Independently draw $r_k \sim \text{Unif}[0, 1]$ for each $1 \leq k \leq M$.
(b) *Decide:*

$$Z_k^{(t)} = \begin{cases} 1 & \text{iff } r_k \leq \left[1 + \prod_{i=1}^{N} \left(\frac{p_{ik}^{(t)}}{1-p_{ik}^{(t)}}\right)^{1-X_{ik}} \left(\frac{1-p_{ik}^{(t)}}{p_{ik}^{(t)}}\right)^{X_{ik}} \left(\frac{1-\phi_k}{\phi_k}\right)\right]^{-1} \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Step 2. Attempt to draw $\theta^{(t)} \sim p(\theta|\mathbf{Z}, g, \delta, \mathbf{X})$, where $\text{rest2} = \{\mathbf{Z}^{(t)}, \{\theta_{-i}, g, \delta\}^{(t-1)}, \mathbf{X}\}$.
(a) *Candidate generation:* Independently draw $\theta_i^c \sim \text{Unif}(0, 1)$ and $r_i \sim \text{Unif}[0, 1]$ for each $1 \leq i \leq N$.
(b) *Decide:*

$$\theta_i^{(t)} = \begin{cases} \theta_i^c & \text{iff } r_i \leq \dfrac{L(\mathbf{X}_i|\theta_i^c, \text{rest2})\pi(\theta_i^c)}{L(\mathbf{X}_i|\theta_i^{(t-1)}, \text{rest2})\pi(\theta_i^{(t-1)})} \\ \theta_i^{(t-1)} & \text{otherwise.} \end{cases} \tag{13}$$

Step 3. Attempt to draw $g^{(t)} \sim p(g|\mathbf{Z}, \theta, \delta, \mathbf{X})$, where $\text{rest3} = \{\{\mathbf{Z}, \theta\}^{(t)}, \{g_{-i}, \delta\}^{(t-1)}, \mathbf{X}\}$.
(a) *Candidate generation:* Independently draw $g_i^c \sim \text{Unif}(0, 1)$ and $r_i \sim \text{Unif}[0, 1]$ for each $1 \leq i \leq N$.
(b) *Decide:*

$$g_i^{(t)} = \begin{cases} g_i^c & \text{iff } r_i \leq \dfrac{L(\mathbf{X}_i|g_i^c, \text{rest3})\pi(g_i^c)}{L(\mathbf{X}_i|g_i^{(t-1)}, \text{rest3})\pi(g_i^{(t-1)})} \\ g_i^{(t-1)} & \text{otherwise.} \end{cases} \tag{14}$$

Step 4. Attempt to draw $\delta^{(t)} \sim p(\delta|\mathbf{Z}, \theta, g, \mathbf{X})$, where $\text{rest4} = \{\{\mathbf{Z}, \theta, g\}^{(t)}, \delta_{-k}^{(t-1)}, \mathbf{X}\}$.
(a) *Candidate generation:* Independently draw $\delta_k^c \sim \text{Unif}(0, 1)$ and $r_k \sim \text{Unif}[0, 1]$ over $M - 1$ items (for an item k, set $\delta_k = \frac{1}{2}$ over $t = 0, \ldots T$).
(b) *Decide:*

$$\delta_k^{(t)} = \begin{cases} \delta_k^c & \text{iff } r_k \leq \dfrac{L(\mathbf{X}_k|\delta_k^c, \text{rest4})\pi(\delta_k^c)}{L(\mathbf{X}_k|\delta_k^{(t-1)}, \text{rest4})\pi(\delta_k^{(t-1)})} \\ \delta_k^{(t-1)} & \text{otherwise.} \end{cases} \tag{15}$$

The equations in substep (b) of the four MCMC steps use full-conditional posterior distributions that characterize the Gibbs sampler (e.g., Gelfand & Smith, 1990). Steps 2, 3, and 4 employ the Metropolis-Hastings algorithm (e.g., Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), where a step generates a candidate parameter value in sub-step (a), and sub-step (b) decides whether to accept it. The decision is based on a posterior odds-ratio that compares the posterior probability given the parameter candidate with the posterior probability given the parameter value at iteration $t-1$. Clearly, special cases of the MCMC algorithm handle the submodels of the GCM3. GCMδ needs only MCMC Steps 1 through 3, GCMg only Steps 1, 2, and 4, and GCM1 only requires Steps 1 and 2.

Some observations should be made about the MCMC algorithm:

1. Step 1 is a data augmentation procedure (e.g., Tanner, 1996) which uses Gibbs sampling to obtain probable "latent data" $\mathbf{Z}^{(t)}$, where the bracketed equation in (12) is item k's posterior key probability at iteration t. The equation implies that, for any item k, the GCM compares the set of p_{ik} (correct-response probabilities) among the group individuals who respond $X_{ik} = 1$, against the set of p_{ik} among the group of individuals who respond $X_{ik} = 0$. Since the p_{ik} may

differ over respondents, a minority of respondents can "outweigh" the majority in determining the posterior mode estimate $\hat{Z}_k$.

2. Substep (a) of Metropolis–Hastings Steps 2, 3, and 4 has a convenient feature. For each parameter θ_i, δ_k, and g_i, a candidate can simply be drawn from the uniform (0, 1) distribution, and furthermore, uniform candidate distributions result in an irreducible MCMC chain. Such a chain can reach any region of the domain of the GCMs true posterior in a finite number of iterations.
3. As mentioned, the odds-ratio in substep (b) of Metropolis Steps 2, 3, and 4 is a posterior odds-ratio, simplified to the form (assuming Beta priors):

$$\frac{L(\mathbf{X}|\gamma^c, \text{rest})(\gamma^c)^{a-1}(1-\gamma^c)^{b-1}}{L(\mathbf{X}|\gamma^{(t-1)}, \text{rest})(\gamma^{(t-1)})^{a-1}(1-\gamma^{(t-1)})^{b-1}}, \tag{16}$$

where the posterior marginal density and $B(a, b)$ constants cancel out of the numerator and denominator.

The posterior distribution $p(\mathbf{Z}, \Omega|\mathbf{X})$ obtained from the generated T MCMC samples can be summarized in a number of ways. This study chooses a discrete interpretation of the answer key estimate, with the posterior mode $\hat{\mathbf{Z}}$, having posterior key probability $\{\Pr[Z_k = \hat{Z}_k]; 1 \leq k \leq M\}$. Also, the study interprets the point estimate of Ω in standard fashion with the posterior mean $\overline{\Omega} = \{\overline{\theta}, \overline{g}, \overline{\delta}\}$, though the mode or median are other options. Furthermore, the posterior variance of Ω may be detailed by 95% Bayesian intervals, bracketed by .025 and .975 posterior quantiles.

In the GCM3 and GCM2g, the median (and average) over the posterior means $\overline{\theta} = \{\theta_1, \ldots, \theta_N\}$, and that of $\overline{\delta} = \{\overline{\delta}_1, \ldots, \overline{\delta}_M\}$, depends on which item difficulty parameter is fixed before MCMC estimation. It is customary in IRT inference to center θ and δ estimates around some useful point. Accordingly, the current study centers the posteriors $\overline{\theta}$ and $\overline{\delta}$ to .5, relative to an item with median (posterior mean) difficulty δ_{med} among the M items. By extending equation 4 of Crowther, Batchelder, and Hu (1995) to the current context, any point estimate $\eta \in \{\overline{\theta}_1, \ldots, \overline{\theta}_N, \overline{\delta}_1, \ldots, \overline{\delta}_M\}$ is centered around δ_{med} by:

$$\eta^* = \eta\left(1 + \frac{.5 - \delta_{\text{med}}}{\delta_{\text{med}} - .5\delta_{\text{med}}}\right) \Big/ \left[1 + \eta\left(\frac{.5 - \delta_{\text{med}}}{\delta_{\text{med}} - .5\delta_{\text{med}}}\right)\right] \tag{17}$$

Define $\vec{D}_{ik}$ as the result, after plugging in any pair of posterior means $\{\overline{\theta}_i, \overline{\delta}_k\}$ into (7). Then each of the posterior means $\{\overline{\theta}_i, \overline{\delta}_k\}$, through (17), are centered to $\{\overline{\theta}_i^*, \overline{\delta}_k^*\}$. Plugging these centered values into (7) yields $\vec{D}_{ik}^*$ such that invariance holds, namely $\vec{D}_{ik}^* = \vec{D}_{ik}$. The posterior quantiles of θ and δ are centered in the same manner.

MCMC Convergence Analysis

Of course, the degree to which the posterior estimates are useful depends on how well the MCMC sample converges to the target joint posterior, $p(\mathbf{Z}, \Omega|\mathbf{X})$. Cowles and Carlin (1996) review the many methods available to assess MCMC convergence, which help the analyst choose an appropriate number of MCMC iterations T. In general, before interpreting a model's posterior, it is good practice to discard the samples obtained from the first B "burn-in" iterations $\{\{\mathbf{Z}, \Omega\}^{(t)}; t = 1, \ldots, B\}$, as they may depend on possibly arbitrary starting values $\{\mathbf{Z}, \Omega\}^{(0)}$. It is also important that the observed sequence of the chain is mixed adequately, that is, it does not have too much dependence. This is checked by autocorrelation analysis on each of the GCM parameters over the T iterations. In fact, Geyer (1992) suggests that the burn-in B should be greater than the lag needed to achieve negligible autocorrelations.

It is not the scope of the present study to enter into details on convergence analysis. However, a later section devotes to a simulation study that evaluates how well the MCMC algorithm reproduces data-generating GCM parameters.

GCM Model Evaluation and Selection with MCMC

It is possible to detect important systematic differences between the statistical model and observed data, by generating replicated data sets from the posterior predictive distribution, and then examining differences between the observed and replicated data sets. For any GCM, the posterior predictive distribution is as follows:

$$p(\mathbf{X}^{\text{rep}}|\mathbf{X}) = \int p(\mathbf{X}^{\text{rep}}|\mathbf{Z}, \Omega)p(\mathbf{Z}, \Omega|\mathbf{X})d(\mathbf{Z}, \Omega), \tag{18}$$

where (18) is the density of "future" data replications $\mathbf{X}^{\text{rep}}$, which are conditioned on model parameters $\{\mathbf{Z}, \Omega\}$, and the data $\mathbf{X}$ used to estimate them. Thus, $\mathbf{X}^{\text{rep}}$ can be viewed as "probable data" under the GCMs posterior. The predictive distribution (18) is simulated over the T MCMC iterations, simply as a byproduct of the four-step algorithm described earlier. After the 4th step of iteration t, independently over all $i = 1, \ldots, N$ and $k = 1, \ldots, M$, $Y_{ik}^{\text{rep}(t)} = 1$ is generated with probability $p_{ik}^{(t)}$, otherwise $Y_{ik}^{\text{rep}(t)} = 0$, yielding:

$$X_{ik}^{\text{rep}(t)} = Y_{ik}^{\text{rep}(t)} Z_k^{(t)} + (1 - Y_{ik}^{\text{rep}(t)})(1 - Z_k^{(t)}). \tag{19}$$

As in Gelman, Meng, and Stern (1996), posterior predictive p-values are obtainable from (18) over the $T - B$ MCMC (nonburn-in) samples to evaluate the fit of a GCM to data. They can be obtained to assess the data-fit of any particular important aspect of the model. For instance, the posterior p-value of observed response X_{ik} is estimated by:

$$\Pr(X_{ik} = X_{ik}^{\text{rep}}) = (T - B)^{-1} \sum_{t=B+1}^{T} \left[X_{ik} X_{ik}^{\text{rep}(t)} + (1 - X_{ik})(1 - X_{ik}^{\text{rep}(t)})\right], \tag{20}$$

where a low p-value indicates that observation X_{ik} is unlikely to have occurred under the model. Now consider the Bernoulli deviance function, stated in terms of the GCM:

$$D(\mathbf{X}; \mathbf{Z}, \Omega) = -2 \sum_{i=1}^{N} \sum_{k=1}^{M} \ell n\left[p_{ik}^{X_{ik}Z_k + (1-X_{ik})(1-Z_k)} (1 - p_{ik})^{X_{ik}(1-Z_k) + (1-X_{ik})Z_k}\right], \tag{21}$$

where, as in (8), p_{ik} is conditional on $\{\mathbf{Z}, \Omega\}$ (for examples of deviance functions, see McCullaugh & Nelder, 1983). Then the global posterior p-value of a GCM is estimated by

$$\Pr[D(\mathbf{X}^{\text{rep}}; \mathbf{Z}, \Omega) \geq D(\mathbf{X}; \mathbf{Z}, \Omega)]$$
$$= (T - B)^{-1} \sum_{t=B+1}^{T} \begin{cases} 1 & \text{iff } D(\mathbf{X}^{\text{rep}}; \{\mathbf{Z}, \Omega\}^{(t)}) \geq D(\mathbf{X}; \{\mathbf{Z}, \Omega\}^{(t)}) \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

As shown, the p-value in (22) is based on the MCMC-estimated reference distribution of $D(\mathbf{X}^{\text{rep}}; \mathbf{Z}, \Omega)$, where a low p-value implies that data $\mathbf{X}$ is unlikely to have occurred under the model. The method can be simplified to examine the fit of a GCM to any subset of $\mathbf{X}$. For example, the posterior predictive p-value of a respondent's vector $\mathbf{X}_i$, obtained by modifying (21) and (22) to evaluate $D(\mathbf{X}_i^{\text{rep}}; \mathbf{Z}, \Omega)$ and $D(\mathbf{X}_i; \mathbf{Z}, \Omega)$, checks whether respondent i's responses are consistent with the single culture assumption of Axiom 1. The posterior predictive p-value of an item's response vector $\mathbf{X}_k$, obtained in the same manner, provides an item-specific evaluation of fit.

Within the context of discrete-data regression, Gelman et al. (1999) note in passing that the posterior predictive inference can be employed to evaluate a model's assumption of conditional independence. Yen's Q, known as a (frequentist) statistic that tests the zero correlation null hypothesis of an IRT model's conditional independence assumption (see Chen & Thissen, 1997), is adapted into the current posterior predictive framework of the GCM. The posterior p-value of the Q statistic is estimated by:

$$\Pr[Q(\mathbf{Y}_i^{\text{rep}}, \mathbf{Y}_j^{\text{rep}}; \mathbf{Z}, \Omega) \geq Q(\mathbf{Y}_i, \mathbf{Y}_j; \mathbf{Z}, \Omega)] = (T-B)^{-1} \sum_{t=B+1}^{T} \begin{cases} 1 & \text{iff } \left|Q_{ij}^{\text{rep}(t)}\right| \geq \left|Q_{ij}^{(t)}\right| \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

The posterior p-value estimator in (23) depends on the MCMC-estimated reference distribution of $Q_{ij}^{\text{rep}(t)}$, which correlates

$$\text{res}(\mathbf{Y}_i^{\text{rep}(t)}, \mathbf{p}_i^{(t)}) = (Y_{i1}^{\text{rep}(t)} - p_{i1}^{(t)}, \ldots, Y_{iM}^{\text{rep}(t)} - p_{iM}^{(t)})$$

between respondents i and j on replicate data. The statistic $Q_{ij}^{(t)}$ pertains to the observed data, as it correlates $\text{res}(\mathbf{Y}_i^{(t)}, \mathbf{p}_i^{(t)})$ between i and j, where $Y_{ik}^{(t)} = X_{ik}Z_k^{(t)} + (1 - X_{ik})(1 - Z_k^{(t)})$. A low p-value supports a violation of independence in Axiom 2, for respondents i and j.

Comparing the fit between different GCM versions also provides an evaluation of the item inhomogeneity assumption of Axiom 3, and the assumption of neutral guessing bias among the N respondents. The evaluation can be achieved with model selection methods that aim to identify the best fitting model among a competing set of models, that is, the one that has the best tradeoff between data underfit and overfit. Bayesian model selection is offered by the Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, and van der Linde, 2002). For a given GCM, the DIC yields:

$$\text{DIC(GCM)} = D(\mathbf{X}; \hat{\mathbf{Z}}, \overline{\Omega}) + 2(\overline{D(\mathbf{X}; \mathbf{Z}, \Omega)} - D(\mathbf{X}; \hat{\mathbf{Z}}, \overline{\Omega})), \tag{24}$$

where $D(\mathbf{X}; \hat{\mathbf{Z}}, \overline{\Omega})$ is the GCM deviance (21) evaluated at the posterior mode $\hat{\mathbf{Z}}$ and posterior mean $\overline{\Omega}$, and $\overline{D(\mathbf{X}; \mathbf{Z}, \Omega)}$ is the posterior mean of (21) over the MCMC chain, $\overline{D(\mathbf{Z}, \Omega)} = (T-B)^{-1} \sum_{B+1}^{T} D(\mathbf{X}; \{\mathbf{Z}, \Omega\}^{(t)})$. The difference component in (24), $\overline{D(\mathbf{X}; \mathbf{Z}, \Omega)} - D(\mathbf{X}; \hat{\mathbf{Z}}, \overline{\Omega})$, penalizes model complexity. The DIC penalty approximates Ye's (1998) generalized formulation of degress of freedom (Spiegelhalter et al. 2002), which measures the sensitivity of the model parameters to changes in the observed data. The model with the lowest DIC is identified as best fitting, where models within 2 to 3 DIC units of the "best" deserve consideration (Spiegelhalter et al. 2002).

According to Spiegelhalter et al. (2002), one practical advantage of DIC model selection is that it does not posit that one of the competing models is "true" (unlike the Bayes factor method; see Bernardo & Smith, 1994). Furthermore, it can be applied to arbitrarily complex models, which may either have informative priors, correlated parameters, or where the number of parameters increase with the number of observations (as is the case with the GCM). On the other hand, the current authors agree that model selection analysis should not lead to an "automatic" choice of a model (see Gelman & Rubin, 1995, 1999), something that Spiegelhalter et al. recommend for DIC applications. All the GCM comparisons performed in this study therefore use DIC and posterior predictive model assessment in combination.

GCM Computer Program

A program was written using the S-PLUS (Insightful Corporation, 1995) statistical software package to perform all the MCMC estimations of the GCM described in this study. The code can be obtained directly from the corresponding author.

GCM Analysis on Real Data

This section applies the GCM on two data sets to compare the data-fit performance between the different model versions, and to demonstrate technical aspects of GCM answer key inference. This section bases each model analysis on 5,500 MCMC samples, discarding the first 500 as burn-in (justified by Geyer's, 1992, criterion).

General Information Test

The first data set contains responses of 40 University of California-Irvine psychology undergraduates to 40 True/False trivia items of the General Information Test (GIT, Nelson & Narens, 1980). This data set benchmarked the GCM1 in previous studies (Batchelder & Romney, 1988). This section analyzes the GIT data with the GCM1, GCM2g, GCMδ, and GCM3, using independent uninformative priors on $(\mathbf{Z}, \Omega)$.

The DIC model selection analysis in Table 1a concludes GCM2g as best fitting. The GCM3 was second best, but it has the highest DIC penalty. Also, item inhomogeneity models GCM2g and the GCM3 fit much better than item homogeneity models GCM1 and GCM2δ. This is interesting because previous analysis of the GIT data set assumed item homogeneity. Furthermore, the item difficulty parameters, compared to the bias parameters, better capture the variance of the GIT response data, as shown in comparing DIC values between the GCM2g and GCM2δ.

The posterior predictive model assessments of Table 1b show that, at most, only 10 responses (out of 1600 = 40*40) misfit the four models. Also, all respondents fit all four models, supporting single culture Axiom 1. Furthermore, the posterior predictive assessments add further support for item inhomogeneity Axiom 3. The GCM1 and GCM2δ have about 25% misfitting items, while the item inhomogeneity models GCM2g and GCM3 have at most 1 misfitting item. Finally, Table 1b shows that the GCM2g and GCM3 are generally consistent with conditional independence Axiom 2.

In the GCM2g, the average of posterior means $\overline{\theta}$ was .46 with variance .04, and the average over $\overline{\delta}$ was .49 with variance .06. The model's estimated answer key $\hat{\mathbf{Z}}$, with exception to items

TABLE 1.
Evaluations of GCM fit to the general information test data

	Model selection analysis with the deviance information criterion			
Model	$D(\mathbf{X}; \hat{\mathbf{Z}}, \overline{\Omega})$	$\overline{D(\mathbf{X}; \mathbf{Z}, \Omega)}$	Penalty	DIC
GCM2g	1590.2	1656.5	66.3	1722.8
GCM3	1531.1	1629.2	98.0	1727.2
GCM1	1732.7	1768.1	35.4	1803.5
GCM2δ	1663.1	1733.8	70.7	1804.4

Posterior predictive fit analysis of the GCM.

	Number of posterior predictive p-values < .10				
Model	Misfitting Responses	Misfitting Respondents	Misfitting Items	Global p-value	% of Respondent Pairs Violating CI ($p < .10$)
GCM2g	9	0	0	.30	3.3
GCM3	10	0	1	.04	3.1
GCM1	0	0	10	.29	14.0
GCM2δ	5	0	9	.05	6.3

Note: CI refers to Conditional Independence

15 and 30 (GIT items listed in Romney, Weller, & Batchelder, 1986), corresponds to the actual correct answers on the GIT. This implies that the respondents' cultural beliefs generally matched the objectively true answers for the items, though in many applications of GCM this is not the case (Romney, Weller, & Batchelder, 1986). Over all items, the model posteriors $\Pr[Z_k = \hat{Z}_k]$ averaged .96 with variance .01.

Items 15 and 30 in the GCM2g do have the highest estimated difficulty, $\{\bar{\delta}_{15} = .87, \bar{\delta}_{30} = .91\}$. Item 15's estimate $\hat{Z}_{15} = 1$ is common to the four models, and the posterior p-value concludes that the item fits in three of them (criterion $\geq .10$). Item 30 fits the GCM2g and GCM3 (p-values .25 and .20 respectively), which infer $\hat{Z}_{30} = 1$. But the same item misfits the GCM1 and GCM2δ (p-value = .00 for both models), which infer $\hat{Z}_{30} = 0$. The estimate $\hat{Z}_{30} = 1$ thus seems more reasonable. On closer inspection, with respect to the correct-probability vector $(\tilde{p}_{1,15}, \ldots \tilde{p}_{N,15})$ obtained by plugging $\hat{\mathbf{Z}}$ and $\bar{\Omega}$ into the GCM2g, the 22 majority respondents responding 0 = "False" had an average correct-response probability of .58, and the 18 minority respondents who responded 1 = "True" averaged .60. Thus, the minority's slight superior ability overturned the majority, and the posterior probability $\Pr[\hat{Z}_{30} = 1] = .57$ reflects these slight differences.

To conclude the analysis of the GIT data set, the DIC model selection analysis and posterior predictive model assessments support that item difficulty parameters can certainly improve the fit of a GCM. These parameters also can help increase the precision of answer key estimates, which may be important in "close" answer key decisions, as shown with GIT item 30, where majority rule was "reversed" to decide $\hat{Z}_{30} = 1$.

Medical Student Communication Performance

The second data set involves a medical student of a major Louisiana university, who was evaluated on his ability to communicate effectively to a live actor who simulated a 5-minute medical scenario. The scenario required the student to effectively deliver the news of a patient's death to a family member, played by the actor. Student performance was judged by the actor (A) simulating the scenario, a medical school professor (P) serving as an independent observer, and the student (S) evaluating his own performance. The judges evaluated the student's performance through 1 = "Yes", 0 = "No" responses on a 33-item test, where each item pertains to a desirable patient communication behavior.

In addition to the four GCM versions with uninformative priors, the second data example considers two other GCMs with informative priors. First, the model GCM3p15 is a GCM3 with $a = b = 15$ beta priors on each and every item difficulty parameter (i.e., prior mean of .50, and variance .008). Tightening the variance of the beta distribution adds information to the item difficulty parameters, since each only has three responses from the data ($N = 3$). Second, the model GCM-ORD makes the following assumptions: (1) the restriction $\theta_P > \theta_A > \theta_S$ specifying the professor to have highest judge ability and the student the least; (2) only the student can have nonneutral guessing bias, $\{g_P = g_A = .5, 0 < g_S < 1\}$, as he may be over- or under-optimistic in his own performance evaluation; and (3) item homogeneity. From a Bayesian perspective, the order constraints $0 \leq \theta_S < \theta_P < \theta_A \leq 1$ constitute an informative prior:

$$\pi(\theta) = \begin{cases} 6 & \text{iff } \theta_P > \theta_A > \theta_S \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

where (25) is a probability density function on $(0, 1)^{N=3}$. As in Gelfand, Smith, and Lee (1992), such a complicated prior is routinely accommodated for GCM-ORD posterior estimation by modifying Step 2 of the MCMC algorithm to sample each of the θ_i one-by-one. For example, that step generates candidate θ_A^c from the "updated" candidate distribution $\text{Unif}(\theta_P^{(t)}, \theta_S^{(t-1)})$. The reader is also referred to Karabatsos (2001) for related MCMC applications of order-restricting priors, used for the estimation of the isotonic ordinal probabilistic model of IRT (Scheiblechner, 1995).

TABLE 2.
Comparison of GCM models by DIC and the posterior predictive p-value

Model	$D(\mathbf{X}; \hat{\mathbf{Z}}, \overline{\Omega})$	$\overline{D(\mathbf{X}; \mathbf{Z}, \Omega)}$	Penalty	DIC	Posterior p-value
GCM2δ	79.4	88.4	9.0	97.4	.49
GCM3p15	78.5	88.5	10.0	98.4	.46
GCM3	71.7	85.9	14.2	100.1	.42
GCM-ORD	90.1	95.2	5.1	100.3	.51
GCM1	89.5	99.3	9.8	109.0	.50
GCM2g	82.2	99.6	17.3	116.9	.38

Table 2 shows the DIC and posterior predictive fit for the six GCM versions. It concludes the GCM2δ as best fitting, with three other close competitors within 3 DIC units. The highest DIC penalties of GCM3 and GCM2g imply that the item difficulty parameters cause model overfit, which is not surprising with $N = 3$. However, the $a = b = 15$ item difficulty priors did improve model fit, as GCM3p15 and GCM2δ have similar DIC. Finally, in all six models, the posterior predictive assessments of all responses, items, respondents, and of conditional independence, yielded p-values greater than .10.

Tables 3 and 4 give detailed results of the GCM2δ analysis. Table 3 lists the judge posteriors

$$\{\overline{\theta}_{\mathrm{P}} = .44, \overline{g}_{\mathrm{P}} = .36\}, \{\overline{\theta}_{\mathrm{A}} = .64, \overline{g}_{\mathrm{A}} = .69\}, \quad \text{and} \quad \{\overline{\theta}_{\mathrm{S}} = .43, \overline{g}_{\mathrm{S}} = .79\},$$

which suggest that the student was overly optimistic about his own performance, while the professor was the most severe judge. For each of the 33 items, Table 4 gives the key estimate, and the item response pattern $\mathbf{X}_k = (X_{\mathrm{P}k}, X_{\mathrm{A}k}, X_{\mathrm{S}k})$ observed over the 3 judges. The student's performance $\hat{\mathbf{Z}}$ is the same as what is obtained from observing the majority rule of the item responses. The posterior probabilities $\Pr[Z_k = \hat{Z}_k]$, over items, averaged .88 with variance .02, which is remarkable for $N = 3$.

The GCM-ORD concludes $\overline{\theta}_{\mathrm{P}} = .69$, $\overline{\theta}_{\mathrm{A}} = .47$, and $\{\overline{\theta}_{\mathrm{S}} = .24, \overline{g}_{\mathrm{S}} = .83\}$, and infers an answer key $\hat{\mathbf{Z}}$ different from the GCM2δ on nine items. Denote an item's responses by $\mathbf{X}_k = (X_{\mathrm{P}k}, X_{\mathrm{A}k}, X_{\mathrm{S}k})$. Eight items had pattern $\mathbf{X}_k = (0, 1, 1)$, and one had $\mathbf{X}_k = (1, 0, 0)$. Conclusions $\hat{Z}_k = 0$ and $\hat{Z}_k = 1$ are equally likely for all these items, where the posterior key probability $\Pr[Z_k = \hat{Z}_k]$ was around .50 for them. Thus, a single item response of the professor, postulated to have the highest medical expertise and no bias, had about as much weight as the responses of the two student judges.

The second data set illustrated that, for very small N, the GCM can obtain high answer key posterior probabilities. Respondent parameters, like the item difficulty parameters, may also be instrumental in reversing majority rule for the answer key estimate $\hat{\mathbf{Z}}$. Furthermore, informative priors are useful for improving a GCMs fit under small N (or M) conditions. They may also address substantive aspects of answer key inference, as shown with the GCM-ORD, which represents only one simple example.

TABLE 3.
GCM2δ judge posteriors, involving the medical student data set

	Posterior Distribution of the Judges						
	Posterior Ability θ_i			Posterior Bias g_i			Posterior predictive
Judge	Mean	2.5%	97.5%	Mean	2.5%	97.5%	p-value fit
Professor	.44	.05	.87	.36	.07	.66	.39
Actor	.64	.12	.96	.69	.48	.97	.57
Student	.43	.05	.90	.79	.21	.96	.39

TABLE 4.
GCM2δ item posteriors, involving the medical student data set

Item Vector	Key Posteriors $\hat{Z}_k$	$\Pr[Z_k = \hat{Z}_k]$	Posterior p-value fit	Item
111	1	.99	.95	1. Fully introduced his/herself.
011	1	.88	.52	2. Fully identified the listener.
111	1	.99	.95	3. Addressed listener by proper surname.
111	1	1.0	.96	4. Quickly established rapport.
111	1	.99	.95	5. Sat down when communicating.
111	1	.99	.95	6. Initially assessed listener's understanding
111	1	.99	.96	7. Used open-ended questions.
111	1	.99	.96	8. Did not interrupt speaker.
011	1	.88	.50	9. Used good nonverbal communication.
111	1	.99	.96	10. Recounted events in sequential order.
111	1	.99	.96	11. Explained the diagnosis to the listener.
110	1	.72	.24	12. Used understandable language.
010	0	.60	.25	13. Showed integrated medical knowledge.
111	1	.99	.96	14. Provided eye contact when listening.
111	1	.99	.96	15. Provided eye contact when speaking.
011	1	.88	.52	16. Understood what the listener felt.
011	1	.87	.51	17. Made listener feel at ease.
001	0	.80	.55	18. Interested in the listener's concerns.
101	1	.56	.21	19. Addressed emotional needs.
010	0	.60	.27	20. Talked about listener's support system.
011	1	.88	.52	21. Offered to help inform loved ones.
100	0	.70	.35	22. Offered to contact other family.
001	0	.82	.57	23. Addressed listener's need for comfort.
001	0	.82	.56	24. Acknowledged the listener's distress.
111	1	.99	.95	25. Ensured access during times of need.
001	0	.81	.56	26. Summarized at the end of encounter.
000	0	.92	.72	27. Assessed listener's understanding.
000	0	.92	.72	28. Asked: "Anything we didn't cover?"
111	1	.99	.95	29. Treated the listener with respect.
111	1	.99	.95	30. Did not use rehearsed communication.
011	1	.87	.51	31. Conveyed confidence.
011	1	.86	.51	32. Behaved as the listener's advocate.
011	1	.88	.52	33. Informed listener about social services.

Note: *Item vector* refers to the item response vector $\mathbf{X}_i = (X_{Pi}, X_{Ai}, X_{Si})$.

GCM Simulation Study

The current section generates response profile data, and evaluates the ability of each GCM model to recover the generating parameters and answer keys through MCMC estimation. To reflect data sets more typical of CCT research, four types of small respondent sample sizes are considered, $N = 3, 7, 15$, and 25. These sample sizes were examined on short, medium, and long tests, $M = 15, 31$, and 61 items, respectively. The GCM3 is the focus of the 12 (= 4*3) simulations, as it enables a simultaneous study of all GCM parameters.

For the $N = 3$ simulation conditions, the generating respondent ability vector was $\theta = (.25, .50, .75)$. For $N = 7$, the generating vector was $\theta = (.02, .18, .34, .50, .66, .82, .98)$, hence, the ability distribution is spaced by increments of .16 around the .5 median. Similarly, the $N = 15$ and $N = 25$ conditions used .069 and .04 ability increments, respectively. For the generating bias parameters, $N = 3$ conditions used $g = (.50, .75, .25)$, and for each of

the $N = 7$, 15, and 25 conditions, g values were independently drawn from the (0, 1) uniform distribution.

On the short test conditions, item 8 was assigned median difficulty of .5, and the remaining 14 items increased (decreased) in .069 difficulty increments around the median. The medium test conditions assigned item 16 with median difficulty and used .03 increments, while the long test conditions used item 31 and .016 increments, respectively. To facilitate comparisons between the generating and estimated item difficulties, the median generating item was fixed to .5 difficulty in MCMC estimation, and the resulting difficulty posteriors were uncentered. Finally, for each of the 12 simulations, M independent draws, from the discrete uniform distribution in {0, 1}, were made to obtain the generating answer key.

The twelve simulation conditions provide strict evaluations of the GCM3. For example, the DIC model selection analysis in Table 1 showed that, by far, the GCM3 had the highest overfit (DIC penalty) on a $N = 40$ data set. Furthermore, each of the twelve simulations will base posterior estimates on only 3,500 MCMC samples (first 500 samples discarded as burn-in), which possibly render the conditions stricter.

Table 5 gives the results of the simulation study, presenting the degree to which the GCM3 estimates recover the generating parameters $\Omega = \{\theta, g, \delta\}$. This degree is measured by "Avg. |Dis|", the absolute distance between a parameter's estimated posterior means and its corresponding generated parameter, averaged over all simulated respondents (or items). The standard deviation of the set of absolute distances is labeled "s.d. |Dis|". Over all twelve simulations, for the ability, bias and item difficulty parameters, the average discrepancy ranged from .04 to .19.

TABLE 5.
GCM3 simulation study: discrepancies between generating parameters with the corresponding posterior means and intervals

	|Dis|		NotPI		|Dis|		NotPI		|Dis|		NotPI	
	Avg.	s.d.	95	99	Avg.	s.d.	95	99	Avg.	s.d.	95	99
	$N = 3, M = 15$				$N = 3, M = 31$				$N = 3, M = 61$			
θ	.15	.12	0	0	.04	.02	0	0	.12	.11	0	0
g	.24	.13	0	0	.16	.03	0	0	.11	.04	0	0
δ	.25	.13	0	0	.19	.14	0	0	.25	.15	1	0
	$N = 7, M = 15$				$N = 7, M = 31$				$N = 7, M = 61$			
θ	.12	.06	1	0	.15	.14	1	1	.09	.08	0	0
g	.11	.10	0	0	.15	.15	0	0	.07	.06	0	0
δ	.24	.12	0	0	.19	.10	0	0	.18	.13	1	0
	$N = 15, M = 15$				$N = 15, M = 31$				$N = 15, M = 61$			
θ	.19	.14	0	0	.11	.09	0	0	.06	.04	0	0
g	.17	.10	0	0	.09	.07	0	0	.07	.07	0	0
δ	.14	.10	0	0	.13	.11	0	0	.13	.09	0	0
	$N = 25, M = 15$				$N = 25, M = 31$				$N = 25, M = 61$			
θ	.15	.11	1	0	.10	.10	1	0	.07	.05	0	0
g	.09	.08	0	0	.11	.11	0	0	.10	.12	2	0
δ	.10	.06	0	0	.11	.09	0	0	.10	.08	1	0

Note: |Discr| Avg. is the average absolute discrepancy between parameter posterior means and their respective generating parameters. |Discr| s.d. is the standard deviation of these discrepancies. NotPI 95 refers to the number of generating parameters not contained in their corresponding 95% posterior intervals (quantiles: .025, .975). NotPI 99 refers to 99% posterior intervals (quantiles: .01, .99).

The ability and bias parameter discrepancies generally decreased with M, and the discrepancies of the item difficulty decreased with N, as expected.

For each of the twelve simulated data sets, Table 5 also displays results on the number of generating Ω parameters falling outside of their corresponding estimated posterior interval. In Table 5, for example, the "NotPI 95" refers to the 95% interval on posterior quantiles .025 and .975, and "NotPI 99" pertains to the 99% interval on quantiles .01 and .99. The Table shows that over all these data sets, only nine of the 728 GCM3 generating Ω parameters fell outside the 95% interval, and only one outside the 99%. The only inconsistent 99% posterior interval, pertaining to an ability parameter, had a .61 value for the .99 posterior quantile, while the generating ability was .66. This interval may have included the generating value if more MCMC samples were utilized (e.g., 10,000) to obtain better convergence to the posterior.

With respect to convergence, unfortunately, the simulations also showed that the MCMC autocorrelations of the ability and bias parameters increased with M, and the autocorrelations of the item difficulty parameters increased with N. For example, in the $N = 25$ and $M = 25$ condition, over all θ and g parameters, the lag needed to yield nonsignificant autocorrelations ranged between 4 and 45, with 80 the most extreme. For $N = 25$ and $M = 61$ simulation, the range of that lag was 10 and 80, with the three most extreme parameters ranging from 113 to 184. Similarly, in the $N = 3$ and $M = 61$ condition, over all δ parameters, that lag was at worst 13. In contrast, the $N = 25$ and $M = 61$ condition had a lag range of 6 and 41, and three difficulty parameters had the most extreme lag, ranging from 66 and 225. All these autocorrelation results occurred because the uniform candidate densities, employed in MCMC Steps 2 through 4, cause the full conditionals of θ, δ, and g to become more concentrated on a part of the (0, 1) interval as NM increases. Increasing the concentration decreases the candidate acceptance rates, which in turn increases the autocorrelations. But despite this autocorrelation issue, a relatively small number of MCMC iterations, under small N conditions, did seem to yield consistency between the generating parameters and posterior intervals.

Table 6 presents the GCM3 ability to recover the generating answer keys. Table 6a presents the number of matches between the generating item keys, and the key posterior modes, for all

TABLE 6.
GCM3 simulation study: Agreement between generating answer keys and their posterior estimates

(a) Number of agreements between generating item answer keys and the corresponding posterior mode item keys.

	$M = 15$	31	61
$N = 3$	15	27	54
7	15	30	57
15	14	31	61
25	14	31	60

(a) Average estimated posterior probability of generating answer keys.

	Item difficulties generated from interval:									
	(0, .1]	(.1, .2]	(.2, .3]	(.3, .4]	(.4, .5]	(.5, .6]	(.6, .7]	(.7, .8]	(.8, .9]	[.9,1)
$N = 3$	.97	.88	.91	.89	.94	.87	.77	.80	.58	.70
7	1	.99	.98	.96	.97	.95	.95	.95	.84	.95
15	1	1	1	1	1	1	1	.96	.96	.90
25	1	1	1	1	1	1	1	.99	.97	.84
	Number of items generated from difficulty interval over the 12 simulations:									
	11	10	11	10	13	10	10	12	10	10

twelve simulations. The key recovery rate ranges from 87% to 100%, where the rates increase with N. Of the 17 items with posterior $\hat{Z}_k$ mismatching the generating item key in the simulations, twelve had a generating item difficulty in the range [82, .98], and five had range [.55, .72]. This result is expected, because item difficulty correlates with the posterior key probability. The estimate $\overline{\delta}_k \approx 1$ implies that all N responses are governed by guessing, and hence provide little information about $\hat{Z}_k$. When $\overline{\delta}_k \approx 0$, all N respondents know that item, and there is very high certainty about $\hat{Z}_k$. This correlation is supported in Table 6b, which presents the estimated posterior probabilities of the generating item keys, as a function of item difficulty, and sample size. These posterior probabilities are averaged over items with generating difficulties in the corresponding range, over all the twelve simulation conditions. The posterior key probabilities of the GCM3 did generally have high recovery rates of the generating item keys, even for very small N.

Conclusion

Markov Chain Monte Carlo inference enables routine analysis of General Condorcet Models with answer key, respondent ability, respondent guessing bias, and item difficulty parameters. The two real-data applications support that the item difficulty and guessing bias parameters can usefully capture a significant amount of variance in test response data, and can be important for answer key estimation. Finally, it was shown that the GCM accurately recovers the answer key for a small number of respondents. Future research will adapt the MCMC framework developed in this study to estimate GCMs formulated for multiple choice and ranking test formats, as well as for "multi-culture" GCMs able to estimate more than one answer key (Batchelder & Romney, 1989, 2000).

References

Aarts, E., & Kours, T.J. (1989). Simulated Annealing and Boltzman machines: *Stochastic approach to combinatorial optimization and neural computing*. New York, NY: John Wiley & Sons.

Baker, F.B. (1992). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Batchelder, W.H., Kumbasar, E., & Boyd, J.P. (1997). Consensus analysis of three-way social network data. *Journal of Mathematical Sociology, 22*, 29–58.

Batchelder, W.H., & Romney, A.K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (pp. 103–112). Greenwich, CT: JAI Press.

Batchelder, W.H., & Romney, A.K. (1988). Test theory without an answer key. *Psychometrika, 53*, 71–92.

Batchelder, W.H., & Romney, A.K. (1989). New results in test theory without an answer key. In E.E. Roskam (Ed.), *Mathematical psychology in progress*. Berlin, Germany: Springer-Verlag.

Batchelder, W.H., & Romney, A.K. (2000). *Extending cultural consensus theory to comparisons among cultures*. Institute of the Mathematical Behavioral Sciences (Tech. Rep. 00–017). Irvine, CA: University of California, Irvine.

Bernardo, J.M., & Smith, A.F.M. (1994). *Bayesian theory*. Chichester, England: John Wiley & Sons.

Carlin, B.P., & Louis, T.A. (1998). *Bayes and empirical Bayes methods for data analysis* (first reprint). Boca Raton, FL: Chapman & Hall/CRC.

Chen, W.H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

Clogg, C.C. (1981). New developments in latent structure analysis. In D.M. Jackson & E.F. Borgatta (Eds.), *Factor analysis and measurement in sociological research* (pp. 215–246). Beverly Hills, CA: Sage Publications.

Cowles, M.K., & Carlin, B.P. (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*, 883–904.

Crowther, C.S., Batchelder, W.H., & Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logic model of perception. *Psychological Review, 102*, 396–408.

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 (with discussion). *Statistica Sinica, 6*, 733–807.

Gelman, A., & Rubin, D.B. (1995). Avoiding model selection in Bayesian social research. In Peter V. Marsden (Ed.), *Sociological Methodology* (pp. 165–173). Cambridge, MA: Blackwell Publishing.

Gelman, A., & Rubin, D.B. (1999). Evaluating and using statistical methods in the social sciences. *Sociological Methods and Research, 27*, 407–410.

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo (with discussion). *Statistical Science, 7*, 473–483.
Geyer, C.J. (1996). Estimation and optimization of functions. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 241–255). Boca Raton, FL: Chapman & Hall/CRC.
Gilks, W.R., Richardson, S., & Spiegelhalter, D.J. (Eds.). (1996). *Markov Chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.
Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics*. New York, NY: John Wiley & Sons.
Grofman, B., & Owen, G. (Eds.). (1986). *Information pooling and group decision making*. Greenwich, CT: JAI Press.
Hastings, W.K. (1970). Monte Carlo methods using Markov Chains and their applications. *Biometrika, 57*, 99–109.
Insightful Corporation. (1995). *S-PLUS documentation*. Seattle, WA: Author. (Formerly Statistical Sciences, Inc.)
Johnson, N.L., & Kotz, S. (1970). *Continuous univariate distributions, Vol. 2*. Boston, MA: Houghton-Mifflin.
Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2*, 389–423.
Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent structure analysis*. New York, NY: Houghton Mifflin.
Lord, F. (1983). Small *N* justifies the Rasch model. In D.J. Weiss (Ed.), *New horizons in latent trait test theory and computerized adaptive testing* (pp. 51–61). New York, NY: Academic Press.
Macmillan, N.A., & Creelman, C.D. (1991). *Detection theory: A user's guide*. New York, NY: Cambridge University Press.
McCullaugh, P., & Nelder, J.A. (1983). *Generalized linear models*. London, U.K.: Chapman and Hall.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of chemical physics, 21*, 1087–1091.
Nelson, T.O., & Narens, L. (1980). Norms of 300 general information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior, 19*, 338–368.
Patz, R.J., & Junker, B.W. (1999). A straightforward approach to Markov Chain Monte Carlo Methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146–178.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research.
Romney, A.K., & Batchelder, W.H. (1999). Cultural consensus theory. In R.A. Wilson & F.C. Keil (Eds.), *The MIT enclyclopedia of the cognitive sciences* (pp. 208–209). Cambridge, MA: The MIT Press.
Romney, A.K., Weller, S.C., & Batchelder, W.H. (1986). Culture as consensus: A theory of culture and respondent accuracy. *American Anthropologist, 88*, 313–338.
Roskam, E.E., & Jansen, P.G.W. (1984). A new derivation of the Rasch model. In E. Degreef & J. Van Buggenhaut (Eds.), *Trends in mathematical psychology* (pp. 293–307). North-Holland: Elsevier Science Publishers.
Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60*, 281–304.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (in press). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*.
Swets, J.A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers (scientific psychology series)*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Tanner, M.A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (3rd ed.). New York, NY: Springer.
Tierney, L. (1994). Exploring posterior distributions with Markov chains (with discussion). *Annals of Statistics, 22*, 1701–1762.
Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association, 93*, 120–132.

Manuscript received 18 MAR 2001
Final version received 9 APR 2002