ON BAYESIAN TESTING OF ADDITIVE CONJOINT MEASUREMENT AXIOMS USING SYNTHETIC LIKELIHOOD

George Karabatsos[1]

University of Illinois-Chicago

June 12, 2017

Abstract

This article introduces a Bayesian method for testing the axioms of additive conjoint measurement. The method is based on an importance sampling algorithm that performs likelihood-free, approximate Bayesian inference using a synthetic likelihood to overcome the analytical intractability of this testing problem. This new method improves upon previous methods because it provides an omnibus test of the entire hierarchy of cancellation axioms, beyond double cancellation. It does so while accounting for the posterior uncertainty that is inherent in the empirical orderings that are implied by these axioms, together. The new method is illustrated through a test of the cancellation axioms on a classic survey data set, and through the analysis of simulated data.

KEYWORDS: Axiom Testing, Conjoint Measurement, Approximate Bayesian Computation.

RUNNING HEAD: Bayesian Testing of Conjoint Measurement Axioms.

[1]Funding was provided by National Science Foundation (Grant Nos. SES-0242030 and SES-1156372).

1 Introduction

This note introduces and illustrates likelihood-free, Bayesian method for empirically testing additive conjoint measurement (ACM) axioms (Luce & Tukey, 1964). The axioms define the empirical conditions under which interval measurement scales can be constructed; and so hence, there are strong motivations for defining methods to test such axioms (e.g., Michell, 1990).

The new Bayesian method provides an omnibus test of the entire hierarchy of cancellation axioms, based on an approximate Bayesian approach to sampling the posterior distribution of model parameters. This approach makes use of a synthetic likelihood, instead of the exact model likelihood which is analytically-intractable. This is intractable because the hierarchy of cancellation axioms, together, imply a complex and highly-interdependent set of order constraints on the parameters.

Before introducing the Bayesian omnibus testing method in Section 2, and illustrating it through the analysis of survey data and simulated data, we briefly review its key related concepts in the following subsections. This includes reviews of ACM theory (§1.1); previous Bayesian approaches to testing the ACM's cancellation axioms (§1.2); the pooled-adjacent-violators algorithm (§1.3), and the original synthetic likelihood method (§1.4). Section 3 provides conclusions.

1.1 *The Theory of Additive Conjoint Measurement (ACM)*

ACM theory states that if a dependent variable is an additive function of two independent variables, then all three variables can be mapped onto a common interval scale (e.g., Domingue, 2014).

More formally, in psychometric terms, if $\mathcal{X}$ denotes a set of persons and if $\mathcal{Y}$ is a set of test items, then this function has the form $z_{ij} = f(x_i, y_j) = g(x_i) + h(y_j)$, where $z_{ij} \in \mathcal{Z}$, $x_i \in \mathcal{X}$, $y_j \in \mathcal{Y}$, with $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z} = \mathbb{R}$; and $f(x_i, y_j) \geq f(x_{i'}, y_{j'})$ when the probability of a correct response (θ_{ij}) for person x_i on item y_j exceeds the probability of a correct response ($\theta_{i'j'}$) for person $x_{i'}$ on item $y_{j'}$. A well known example of an additive function is given by $z_{ij} = f(x_i, y_j) = \log\{\theta_{ij}/(1-\theta_{ij})\} = \beta_i - \delta_j$,

the Rasch (1960) model with person ability parameters β_j and item difficulty parameters δ_i. This model can be easily extended to polytomous items.

The triple $(\mathcal{X}, \mathcal{Y}, \geq)$ is called an *additive conjoint structure* if the following axioms hold:

(1) *Single cancellation (independence)*: if $f(x_1, y_1) \geq f(x_2, y_1)$, then $f(x_1, y_2) \geq f(x_2, y_2)$ for all $y_2 \in \mathcal{Y}$, and if $f(x_1, y_1) \geq f(x_1, y_2)$, then $f(x_2, y_1) \geq f(x_2, y_2)$ for all $x_2 \in \mathcal{X}$;

(2) *Double cancellation*: if $f(x_1, y_2) \geq f(x_2, y_1)$ and $f(x_2, y_3) \geq f(x_3, y_2)$, then $f(x_1, y_3) \geq f(x_3, y_1)$ for all $x_1, x_2, x_3 \in \mathcal{X}$ and all $y_1, y_2, y_3 \in \mathcal{Y}$, in a 3×3 matrix;

(3) *Solvability*: for all $x_1 \in \mathcal{X}$, and for $y_1, y_2 \in \mathcal{Y}$, there is one $x_2 \in \mathcal{X}$, where $f(x_1, y_1) = f(x_2, y_2)$;

(4) *Archimedean Condition*: no value of a quantitative variable is infinitely larger than any other value (Michell, 1990, p.73).

These four axioms, together, define the sufficient conditions for the existence of interval scales (Luce & Tukey, 1964, Theorems VID through VIJ); and imply the existence of functions $\varphi_1 : \mathcal{X} \to \mathbb{R}$ and $\varphi_2 : \mathcal{Y} \to \mathbb{R}$ that are unique up to linear transformation, such that $f(x_1, y_1) \geq f(x_2, y_2) \Leftrightarrow \varphi_1(x_1) + \varphi_2(y_1) \geq \varphi_1(x_2) + \varphi_2(y_2)$.

The solvability and Archimedean axioms are not empirically testable. However, necessary and testable conditions for these axioms are given by the hierarchy of cancellation axioms. They include single, double, triple, quadruple cancellation, and higher-order cancellation (Scott, 1964). For example, one instance of *triple cancellation* requires that if $f(x_2, y_1) \geq f(x_1, y_2)$, $f(x_1, y_4) \geq f(x_2, y_3)$, and $f(x_3, y_4) \geq f(x_4, y_3)$, then $f(x_3, y_1) \geq f(x_4, y_2)$, for all $x_1, x_2, x_3, x_4 \in \mathcal{X}$ and all $y_1, y_2, y_3, y_4 \in \mathcal{Y}$, in a 4×4 matrix.

The hierarchy of cancellation conditions are highly-interdependent. For example, while there are 36 possible orderings of a 3×3 matrix that satisfy double cancellation, only a small subset of these orderings are logically-independent when single cancellation holds (Domingue, 2014, p. 7;

Michell, 1988; Luce & Steingrimsson, 2011; and the references therein). Also, while there are 51 orderings of a 4×4 matrix that satisfy triple cancellation, many of these orderings either contradict the single cancellation axiom, or are trivially true when double cancellation holds (Kyngdon & Richards, 2007). Such complex interdependencies have led some authors to consider cancellation tests only on a small sub-matrix of the data, instead of the entire data set (e.g., Kyngdon, 2011).

1.2 *Previous Bayesian Approaches to Testing the Cancellation Axioms*

Let $\boldsymbol{\theta} = (\theta_{ij})_{I \times J} \in [0,1]^{IJ}$ be a matrix of correct response probabilities, for I groups of persons (for $i = 1, \ldots, I$) on J dichotomous test items (for $j = 1, \ldots, J$). Each group i has a common total score on the test, a sufficient statistic for the Rasch model's ability parameter. Then, the hierarchy of cancellation (HC) axioms, up to order $\min(I-1, J-1)-1$, implies that $\boldsymbol{\theta}$ lies in a proper subset A_{HC} of $[0,1]^{IJ}$. Also, A_{HC} is a proper subset of A_{SC}, A_{DC}, A_{TC},..., corresponding to values of $\boldsymbol{\theta}$ that satisfy single cancellation (SC), double cancellation (DC), triple cancellation (TC), and so on. Further, $\boldsymbol{\theta} \in A_{DC}$ when all $\binom{I}{3}\binom{J}{3}$ of the 3×3 sub-matrices of $\boldsymbol{\theta}$ satisfy double cancellation. And, $\boldsymbol{\theta} \in A_{TC}$ when all $\binom{I}{4}\binom{J}{4}$ of the 4×4 sub-matrices of $\boldsymbol{\theta}$ satisfy triple cancellation.

Given a set of data from a sample of persons' individual responses to J items, the matrix $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_{ij} = r_{ij}/n_{ij})_{I \times J}$ gives the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$. Here r_{ij} is the number of n_{ij} persons in score group i who correctly answered item j, with $\widehat{\theta}_{ij}$ the proportion correct.

Table 1 presents a data set ($\widehat{\boldsymbol{\theta}}$) that was analyzed in a classic study of ACM and the Rasch model (Perline et al., 1979). The data were obtained from 490 released convicts who individually responded to a survey of nine dichotomous items. The survey was used to make parole decisions (Hoffman & Beck, 1974). The survey items are: (1) Grade Claimed; (2) Auto Theft; (3) Age at First Commitment; (4) Prior incarcerations; (5) Drug History; (6) Planned Living Arrangement; (7) Employment; (8) Prior Convictions; (9) Parole Revoked. Each item response was scored either

Raw	Item									Number
Score	6	1	8	7	4	9	2	3	5	of persons
1	.00	.00*	.00*	.00*	.00*	.00*	.27*	.00*	.73*	15
2	.06	.04	.04*	.19	.06*	.23*	.51*	.21*	.64*	47
3	.07	.15	.08	.39*	.18*	.33	.61	.52	.67	61
4	.18*	.24*	.12	.40*	.52	.51	.64	.68	.70*	84
5	.13*	.33	.30	.51	.73	.68	.68*	.84	.78*	82
6	.13	.28	.64	.58*	.95	.91	.77*	.97	.78*	86
7	.17*	.47*	.85*	.82	1.0*	.93	.90	.97	.90*	60
8	.17*	.85*	1.0*	.98*	1.0*	1.0*	1.0	1.0	1.0	47
9	1.0*	1.0*	1.0*	1.0*	1.0	1.0	1.0	1.0	1.0	8
									sum =	490

Table 1: The proportion of correct answers for each item in groups by raw score, for the Parole data (from Table 2 of Perline et al., 1979). An asterisk (*) indicates a proportion violating the hierarchy of cancellation axioms according to its Kullback-Leibler measure exceeding .01. This is based on the Bayesian axiom testing method introduced in Section 2.

as 1 = "presence", or 0 = "absence".

An empirical test of the HC axioms (up to order $\min(I-1, J-1)-1$), on a data set $\widehat{\boldsymbol{\theta}}$, amounts to the test of the composite null hypothesis (H_0) that $H_0 : \boldsymbol{\theta} \in A_{HC}$. Likewise, empirical tests of single, double, triple cancellation, and so on, refer to tests of composite null hypothesis $H_0 : \boldsymbol{\theta} \in A_{SC}$, $H_0 : \boldsymbol{\theta} \in A_{DC}$, $H_0 : \boldsymbol{\theta} \in A_{TC}$, ..., respectively. Perline et al. (1979) tested double cancellation on 3×3 sub-matrices of the Parole data $\widehat{\boldsymbol{\theta}}$ using multiple hypothesis tests (resp.). Arguably, the results of such tests are difficult to summarize because of the dependence between these tests, and due to Type I error rate inflation that arises from the multiple testing.

This motivated Karabatsos (2001) to propose a Bayesian beta-binomial model for testing the cancellation axioms. For the test of the null $H_0 : \boldsymbol{\theta} \in A_{HC}$ that the data $\widehat{\boldsymbol{\theta}}$ satisfy cancellation up to order $\min(I-1, J-1)-1$, this model has a posterior distribution with p.d.f. given by:

$$\pi(\boldsymbol{\theta} \,|\, \mathbf{r}, \mathbf{n}, A_{HC}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \binom{n_{ij}}{r_{ij}} \theta_{ij}^{a_{ij}+r_{ij}} (1-\theta_{ij})^{b_{ij}+n_{ij}-r_{ij}} \mathrm{be}(\theta_{ij} \,|\, a_{ij}, b_{ij}) \mathbf{1}(\boldsymbol{\theta} \in A_{HC}), \qquad (1)$$

up to a normalizing constant, where $\mathbf{1}(\cdot)$ is the (0 or 1 valued) indicator function. As shown in

(1), the model is defined by a product of IJ independent binomial likelihoods, and beta prior distributions (p.d.f.s) truncated into the parameter subspace $A_{HC} \subseteq [0,1]^{IJ}$, by $\mathbf{1}(\boldsymbol{\theta} \in A_{HC})$. The beta p.d.f. is $\text{be}(\theta \,|\, a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\theta^{a-1}(1-\theta)^{b-1}$, where Γ is the gamma function. A non-informative uniform ($a_{ij} = b_{ij} = 1$) or reference ($a_{ij} = b_{ij} = 1/2$) prior is typically chosen in practice.

Then, for the Parole data, a test of the HC axioms (up to order $\min(I-1, J-1)-1$) corresponds to a test of the null hypothesis that $\boldsymbol{\theta} = (\theta_{ij})_{9\times 9} \in A_{HC} \subseteq [0,1]^{81}$, such that $\binom{9}{3}\binom{9}{3} = 7\,056$ of the 3×3 sub-matrices satisfy double cancellation, and so on for the higher-order cancellation axioms. The subset A_{HC} is complex, because of the very large numbers of sub-matrices involved, and because the HC axioms are highly interdependent (Kyngdon & Richards, 2007). Then, the posterior distribution (1) is analytically intractable, mainly because it depends on $\mathbf{1}(\boldsymbol{\theta} \in A_{HC})$.

This led Karabatsos (2001) to propose an empirical Bayes approach to testing $H_0 : \boldsymbol{\theta} \in A_{HC}$. This involved testing $H_0 : \boldsymbol{\theta} \in \widehat{A}_{HC}$ and using $\mathbf{1}(\boldsymbol{\theta} \in \widehat{A}_{HC})$ instead of $\mathbf{1}(\boldsymbol{\theta} \in A_{HC})$ in the prior in (1). Here, $\widehat{A}_{HC}$ is the simple linear ordering that is defined by the matrix $\widehat{\boldsymbol{\theta}}_R = (\widehat{\theta}_{Rij} = \frac{\exp(\widehat{\beta}_i - \widehat{\delta}_j)}{1+\exp(\widehat{\beta}_i - \widehat{\delta}_j)})_{I\times J}$ of MLE estimates of the Rasch model parameters, so that $\widehat{A}_{HC} \subseteq A_{HC}$. Then, based on this linear ordering, inference of the empirical Bayes version $\pi(\boldsymbol{\theta} \,|\, \mathbf{r}, \mathbf{n}, \widehat{A}_{HC})$ of the posterior p.d.f. (1) can proceed through the use of standard (Metropolis or Gibbs) MCMC sampling algorithms (Gelfand, et al. 1992; Karabatsos, 2001). Using MCMC, the test of the null $H_0 : \boldsymbol{\theta} \in \widehat{A}_{HC}$ proceeds by comparing the estimated marginal posterior distribution of each individual element of $\boldsymbol{\theta}$ from the joint posterior $\pi(\boldsymbol{\theta} \,|\, \mathbf{r}, \mathbf{n}, \widehat{A}_{HC})$, with each corresponding element of the data set $\widehat{\boldsymbol{\theta}}$. This could be done by inference of marginal posterior distribution of the residuals $\theta_{ij} - \widehat{\theta}_{ij}$, for each i and j. To summarize, in order to deal with the analytical intractability of the parameter subset A_{HC}, this empirical Bayes approach makes use of the more-tractable linear-order subset $\widehat{A}_{HC}$ in order to provide a computational-tractable posterior distribution (p.d.f.), $\pi(\boldsymbol{\theta} \,|\, \mathbf{r}, \mathbf{n}, \widehat{A}_{HC})$. The trade-off is that posterior uncertainty in $\boldsymbol{\theta}$ is not fully accounted for, because $\widehat{A}_{HC}$ is fixed.

For the problem of testing single and double cancellation, that is, testing the null $H_0 : \boldsymbol{\theta} \in A_{SC} \cap A_{DC}$, Domingue (2014) adopted Karabatsos' general Bayesian beta-binomial model (1), using $\mathbf{1}(\boldsymbol{\theta} \in A_{SC} \cap A_{DC})$ instead of $\mathbf{1}(\boldsymbol{\theta} \in A_{HC})$ in the model prior. For this testing problem, he devised an ingenious and complex Metropolis MCMC sampling algorithm to perform inference of the posterior $\pi(\boldsymbol{\theta} \,|\, \mathbf{r}, \mathbf{n}, A_{SC} \cap A_{DC})$, which can search over regions of the complex subspace $A_{SC} \cap A_{DC}$ of $\boldsymbol{\theta}$ that have posterior support. For the Parole data example, it can search over the 7 056 3×3 sub-matrices. However, this MCMC method does not directly address the original problem of testing the null hypothesis, $H_0 : \boldsymbol{\theta} \in A_{HC}$. This is because testing higher-order cancellation conditions, such as triple cancellation, is a computationally-intractable problem, which would require searching over a very large number $\binom{I}{4}\binom{J}{4}$ of the total 4×4 sub-matrices of $\boldsymbol{\theta}$ (Domingue, 2014, p.16). This is 15 876 for the modest-sized Parole data. The fact that an ingenious MCMC method (Domingue, 2014) was needed to perform a full Bayesian test of cancellation up to order 2, may suggest that MCMC methods are not fully appropriate for the Bayesian testing of the complex and correct hypothesis, $H_0 : \boldsymbol{\theta} \in A_{HC}$.

The complexities mentioned in Sections 1.1-2, also suggest that focusing on individualized tests of cancellation axioms may not provide the best approach. This motivates the development of the new Bayesian omnibus test of the entire hierarchy of cancellation (HC) axioms, presented in §2. It is based on the synthetic likelihood method, and the PAVA estimator, reviewed next.

1.3 *The Synthetic Likelihood (SL) Method for Approximate Bayesian Inference*

The synthetic likelihood (SL) method, which is well-established by now (e.g., Wood, 2010; Price et al. 2017), can be used to perform approximate Bayesian inference for a model that is defined by a likelihood that is computationally and/or analytically intractable. In order to describe the general SL method, we need to set more notation. Denote $L(\mathbf{y}_n \,|\, \boldsymbol{\theta})$ as a model's likelihood probability

density (or mass) function for a data set $\mathbf{y}_n$ of sample size n, conditionally on a value of the model parameters, $\boldsymbol{\theta}$. Assume that the (exact) likelihood, $L(\mathbf{y}_n \,|\, \boldsymbol{\theta})$, is intractable, but that it is still possible to generate samples of (synthetic) data sets $\mathbf{z}_n$ from it. Let $\pi(\boldsymbol{\theta})$ be the prior p.d.f. defined on the space Θ of $\boldsymbol{\theta}$. The posterior p.d.f. is $\pi(\boldsymbol{\theta} \,|\, \mathbf{y}_n) = L(\mathbf{y}_n \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})h^{-1}(\mathbf{y}_n)$ (with $h(\mathbf{y}_n) = \int L(\mathbf{y}_n \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$). However, it cannot be easily computed due to the intractable likelihood.

The SL method addresses this intractability, as follows. Let $\mathbf{t}(\mathbf{y}_n)$ be a vector of k chosen summary statistics of the data set $\mathbf{y}_n$, which have a large-sample asymptotic multivariate normal distribution with p.d.f. $\mathrm{norm}_k(\mathbf{t} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\{-(1/2)(\mathbf{t}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{t}-\boldsymbol{\mu})\}}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}}$. The SL method approximates the posterior $\pi(\boldsymbol{\theta} \,|\, \mathbf{y}_n)$ by replacing the intractable likelihood $L(\mathbf{y}_n \,|\, \boldsymbol{\theta})$ with a tractable likelihood that is defined by (e.g., Price et al. 2017, eq. 5):

$$L^*(\mathbf{t}(\mathbf{y}_n) \,|\, \boldsymbol{\theta}) = \int \cdots \int \mathrm{norm}_k(\mathbf{t}(\mathbf{y}_n) \,|\, \widehat{\boldsymbol{\mu}}_\mathbf{t}, \widehat{\boldsymbol{\Sigma}}_\mathbf{t}) \prod_{m=1}^{N} L_\mathbf{t}(\mathbf{t}(\mathbf{z}_n^{(m)}) \,|\, \boldsymbol{\theta})\mathrm{d}\mathbf{t}(\mathbf{z}_n^{(1)}) \cdots \mathrm{d}\mathbf{t}(\mathbf{z}_n^{(N)}). \tag{2}$$

Using Monte Carlo sampling methods, an unbiased estimate of (2) is given by:

$$\mathbf{z}_n^{(s,1)}, \ldots, \mathbf{z}_n^{(s,N)} \overset{i.i.d.}{\sim} L(\mathbf{z}_n \,|\, \boldsymbol{\theta}), \text{ for } s = 1, \ldots, S, \tag{3a}$$

$$\widehat{\boldsymbol{\mu}}_\mathbf{t}^{(s)} = \tfrac{1}{S}\sum\nolimits_{s=1}^{S} \mathbf{t}(\mathbf{z}_n^{(s)}); \qquad \widehat{\boldsymbol{\Sigma}}_\mathbf{t}^{(s)} = \tfrac{1}{S-1}\sum\nolimits_{s=1}^{S} (\mathbf{t}(\mathbf{z}_n^{(s)}) - \widehat{\boldsymbol{\mu}}_\mathbf{t}^{(s)})(\mathbf{t}(\mathbf{z}_n^{(s)}) - \widehat{\boldsymbol{\mu}}_\mathbf{t}^{(s)})^\intercal; \tag{3b}$$

$$\widehat{L}^*(\mathbf{t}(\mathbf{y}_n) \,|\, \boldsymbol{\theta}) = \tfrac{1}{S}\sum\nolimits_{s=1}^{S} \mathrm{norm}_k(\mathbf{t}(\mathbf{y}_n) \,|\, \widehat{\boldsymbol{\mu}}_\mathbf{t}^{(s)}, \widehat{\boldsymbol{\Sigma}}_\mathbf{t}^{(s)}), \tag{3c}$$

where $(\widehat{\boldsymbol{\mu}}_\mathbf{t}^{(s)}, \widehat{\boldsymbol{\Sigma}}_\mathbf{t}^{(s)})$ is the MLE of the mean ($\boldsymbol{\mu}$) and covariance matrix ($\boldsymbol{\Sigma}$) of $\{\mathbf{t}(\mathbf{z}_n^{(s,1)}), \ldots, \mathbf{t}(\mathbf{z}_n^{(s,N)})\}$, from N samples of synthetic data sets $\mathbf{z}_n$ from $L(\cdot \,|\, \boldsymbol{\theta})$. Here, N is chosen to be large.

The SL method, when embedded into a Metropolis MCMC sampling algorithm (Wood, 2010), is described in Algorithm 1, below. This algorithm is run for a sufficiently-large number (S) of

iterations, so that the resulting samples $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(S)}\}$ converge to the approximate posterior p.d.f. (distribution), given by $\pi(\boldsymbol{\theta} \,|\, \mathbf{t}(\mathbf{y}_n)) \propto L^*(\mathbf{t}(\mathbf{y}_n) \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta})$, up to a normalizing constant.

The SL method can be extended to handle non-normally distributed summary statistics (Gutmann & Corander, 2016, §3.3) by replacing the normal p.d.f. $\text{norm}(\cdot \,|\, \widehat{\boldsymbol{\mu}}_{\mathbf{t}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{t}})$ in (2) with a kernel p.d.f. estimate obtained from the synthetic data (Silverman, 1986). The SL method can also be embedded into an importance sampling algorithm, an alternative to the Metropolis algorithm (Robert & Casella, 2004). More details are given in Section 2.

Algorithm 1. The General SL Metropolis MCMC Algorithm.

for $s = 1$ to S, do:

(a) Draw $\boldsymbol{\theta}^* \sim G(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(s-1)})$ from a symmetric proposal distribution, G;

(b) Draw N synthetic data sets of size n by $\mathbf{z}_n^{(s,1)}, \ldots, \mathbf{z}_n^{(s,N)} \overset{i.i.d.}{\sim} L(\mathbf{z}_n \,|\, \boldsymbol{\theta}^*)$,

then calculate the MLE $(\widehat{\boldsymbol{\mu}}_{\mathbf{t}}^{(s)}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{t}}^{(s)})$ from summary statistics, $\mathbf{t}(\mathbf{z}_n^{(s,1)}), \ldots, \mathbf{t}(\mathbf{z}_n^{(s,N)})$;

(c) Accept $\boldsymbol{\theta}^{(s)} = \boldsymbol{\theta}^*$ with probability $\min\left\{1, \dfrac{\text{norm}_k(\mathbf{t}(\mathbf{y}_n) \,|\, \widehat{\boldsymbol{\mu}}_{\mathbf{t}}^{(s)}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{t}}^{(s)})\pi(\boldsymbol{\theta}^*)}{\text{norm}_k(\mathbf{t}(\mathbf{y}_n) \,|\, \widehat{\boldsymbol{\mu}}_{\mathbf{t}}^{(s-1)}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{t}}^{(s-1)})\pi(\boldsymbol{\theta}^{(s-1)})}\right\}$.

end for

1.4 *The Pooled Adjacent Violators Algorithm (PAVA)*

Now, we review the PAVA algorithm for isotonic regression smoothing of proportion data, following Robertson and Warrack (1985). As an example, consider a 1×6 vector of proportion data (MLE), given by $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_j = r_j/n_j)_{1\times 6} = (.21, .10, .15, .15, .09, .10)$, with sample sizes $\mathbf{n} = (33, 87, 67, 83, 58, 70)$ (resp.). Also, consider a hypothesized order-constraint $\theta_1 \geq \theta_2 \geq \theta_3 \geq \theta_4 \geq \theta_5 \geq \theta_6$ on the six proportion parameters.

The PAVA estimator, denoted $\mathbf{t}(\widehat{\boldsymbol{\theta}})$, is the least-squares solution of the observed proportions, subject to that order-constraint, using weights $\mathbf{n}$. For the data example, the PAVA estimate is

$\mathbf{t}(\widehat{\boldsymbol{\theta}}) = (.21, .13, .13, .13, .09, .09)$. This was obtained over three stages (Robertson & Warrack, 1985, Fig. 2). The first stage started with the proportion data, $\mathbf{t}^{(1)}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\theta}}$, which violate the order-constraint. The second stage updates the estimate to $\mathbf{t}^{(2)}(\widehat{\boldsymbol{\theta}}) = (.21, .12, .12, .15, .09, .10)$, where the value .12 for the second and third entries is the pooled weighted-average of the order-violating pair $(.10, .15)$ in $\widehat{\boldsymbol{\theta}} = \mathbf{t}^{(1)}(\widehat{\boldsymbol{\theta}})$, which have sample sizes $(87, 67)$ (the weights). The final stage arrived at the final solution, $\mathbf{t}(\widehat{\boldsymbol{\theta}}) = \mathbf{t}^{(3)}(\widehat{\boldsymbol{\theta}})$, which satisfies the hypothesized order. Here, the value of .09 in the last two entries of $\mathbf{t}(\widehat{\boldsymbol{\theta}})$ is the weighted-average of the order-violating pair $(.09, .10)$ in $\mathbf{t}^{(2)}(\widehat{\boldsymbol{\theta}})$, which have sample sizes $(58, 70)$ (weights).

Finally, for proportion data that is already consistent with the hypothesized order-constraint, say $\widehat{\boldsymbol{\theta}} = (.21, .19, .17, .15, .13, .11)$, the PAVA estimator is given simply by $\mathbf{t}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\theta}}$.

2 The Omnibus Axiom Testing (SL) Method, with Application to the Parole Data

For testing the (correct) hypothesis $H_0 : \boldsymbol{\theta} \in A_{HC}$, the HC axioms up to order $\min(I-1, J-1)-1$, we present a novel SL method to approximate Bayesian inference. This method is based on a truncated exact model likelihood (not truncated prior), given by:

$$L(\mathbf{r} \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \tbinom{n_{ij}}{r_{ij}} \theta_{ij}^{r_{ij}} (1-\theta_{ij})^{n_{ij}-r_{ij}} \mathbf{1}(\boldsymbol{\theta} \in A_{HC}), \tag{4}$$

and the corresponding posterior p.d.f. is given by:

$$\pi(\boldsymbol{\theta} \,|\, \mathbf{r}, \mathbf{n}, A_{Hc}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \tbinom{n_{ij}}{r_{ij}} \theta_{ij}^{r_{ij}} (1-\theta_{ij})^{n_{ij}-r_{ij}} \mathbf{1}(\boldsymbol{\theta} \in A_{HC}) \mathrm{be}(\theta_{ij} \,|\, a_{ij}, b_{ij}). \tag{5}$$

This SL method adopts the view that the likelihood (4) (and posterior, (5)) is intractable due to the complexity of the order-constrained subspace A_{HC}, especially when I and J are large.

The novel SL method is based on summary statistics ($\mathbf{t}(\widehat{\boldsymbol{\theta}})$) that are defined by a PAVA-

smoothed version of the proportion data matrix (MLE), $\widehat{\boldsymbol{\theta}}$. Specifically, PAVA is employed to transform the elements of $\widehat{\boldsymbol{\theta}}$ to a matrix $\mathbf{t}(\widehat{\boldsymbol{\theta}}) = (\mathbf{t}_{ij}(\widehat{\theta}_{ij}))_{I\times J}$ which has cell values that are monotone (non-decreasing) ordered by the corresponding cell values of the Rasch model matrix of MLEs, $\widehat{\boldsymbol{\theta}}_{Rij} = (\widehat{\theta}_{Rij} = \frac{\exp(\widehat{\beta}_i - \widehat{\delta}_j)}{1+\exp(\widehat{\beta}_i - \widehat{\delta}_j)})_{I\times J}$. Then, the PAVA estimate $\mathbf{t}(\widehat{\boldsymbol{\theta}})$ lies in the subset A_{HC} of $[0,1]^{IJ}$ satisfying the entire hierarchy of cancellation axioms, at least approximately. The PAVA smoothing is applied to the observed data, and also applied to samples of the synthetic data sets. This is done in the spirit of the MONANOVA algorithm (Kruskal, 1964), which uses PAVA to transform a data set $\widehat{\boldsymbol{\theta}}$ into data that conforms to a two-way additive ANOVA model. Here, we use the (additive) Rasch model instead of ANOVA.

Using the PAVA estimator $\mathbf{t}(\widehat{\boldsymbol{\theta}})$, the novel SL method approximates the posterior (5) by:

$$\pi(\boldsymbol{\theta} \,|\, \mathbf{t}(\widehat{\boldsymbol{\theta}})) \propto L^*(\mathbf{t}(\widehat{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{6}$$

This posterior (6) is based on the approximate likelihood (L^*), defined by:

$$L^*(\mathbf{t}(\widehat{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta}) = \int \cdots \int \prod_{i=1}^{I} \prod_{j=1}^{J} \tfrac{1}{N h_{ij}} \sum_{m=1}^{N} \mathcal{K}\left(\tfrac{\mathbf{t}_{ij}(\widehat{\theta}_{ij}) - \mathbf{t}_{ij}(\widehat{\theta}_{ij}^{(m)})}{h_{ij}}\right) \prod_{m=1}^{N} L_{\mathbf{t}}(\mathbf{t}(\widehat{\theta}_{ij}^{(m)}) \,|\, \boldsymbol{\theta}) \mathrm{d}\mathbf{t}(\widehat{\theta}_{ij}^{(1)}) \cdots \mathrm{d}\mathbf{t}(\widehat{\theta}_{ij}^{(N)}), \tag{7}$$

where $\mathcal{K}$ is a smooth and symmetric kernel density (p.d.f.) function with mode 0, and h_{ij} is the kernel bandwidth. It is assumed that the kernel is given by, with $\mathcal{K}(\cdot) = \frac{\exp(-(\cdot)^2/2)}{\sqrt{2\pi}}$, the p.d.f. of the standard Normal$(0,1)$ distribution; along with the automatic bandwidth given by $h_{ij} = 1.06\widehat{\sigma}_{\mathbf{t},ij} N^{-1/5}$, according to the normal reference rule. Here, $\widehat{\sigma}_{\mathbf{t},ij}$ is the standard deviation of the statistics $\{\mathbf{t}_{ij}(\widehat{\theta}_{ij}^{(m)})\}_{m=1}^{N}$ over the N synthetic data sets. So, whereas the standard SL method (§1.3) employs an approximate likelihood (2) that assumes normally-distributed summary statistics, the new SL method employs an approximate likelihood (7) that is inferred by kernel density estimation. Thus, the new method does not pre-suppose normality or any other specific distributional form.

This is important because PAVA estimates ($\mathbf{t}(\widehat{\boldsymbol{\theta}})$) are typically not asymptotically normal (Li, 2008, Ch.3). Further, by definition of approximate likelihood (7), the approximate posterior (6) assigns more weight to values of $\boldsymbol{\theta}$ that satisfy the order constraints of A_{HC}, as in the exact posterior (5).

Importance sampling (IS) can be used to infer any function $g(\boldsymbol{\theta})$ of interest of the approximate posterior (6), while using the prior $\pi(\boldsymbol{\theta})$ as the approximating density (p.d.f.), having c.d.f. denoted by $\Pi(\boldsymbol{\theta})$. In this IS approach (F. Leisen, October 13, 2016, personal communication; Zhu et al. 2016), the inference of the integral $\int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \,|\, \mathbf{t}(\widehat{\boldsymbol{\theta}}))\mathrm{d}\boldsymbol{\theta}$ can be re-written as the inference of:

$$\frac{\int g(\boldsymbol{\theta}) L^*(\mathbf{t}(\widehat{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta}) \Pi(\mathrm{d}\boldsymbol{\theta})}{\int L^*(\mathbf{t}(\widehat{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta}) \Pi(\mathrm{d}\boldsymbol{\theta})}, \tag{8}$$

so that an IS estimator of (8) can be obtained by:

$$\frac{\frac{1}{S}\sum_{s=1}^{S} g(\boldsymbol{\theta}^{(s)}) L^*(\mathbf{t}(\widehat{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta}^{(s)})}{\frac{1}{S}\sum_{s=1}^{S} L^*(\mathbf{t}(\widehat{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta}^{(s)})}. \tag{9}$$

Then, the likelihood L^* in (9) is obtained by the following Monte sampling procedure:

$$\boldsymbol{\theta}^{(s)} \overset{iid}{\sim} \pi(\boldsymbol{\theta}), \text{ for } s = 1, \ldots, S, \tag{10a}$$

$$r_{ij}^{(s,m)} \overset{ind}{\sim} \text{Binomial}(n_{ij}, \theta_{ij}^{(s)}), \text{ for } i = 1, \ldots, I,\ j = 1, \ldots, J, \text{ and } m = 1, \ldots, N, \tag{10b}$$

$$\widehat{L}_{ij}^*(\mathbf{t}_{ij}(\widehat{\theta}_{ij}) \,|\, \boldsymbol{\theta}^{(s)}) = \frac{1}{N h_{ij}^{(s)}} \sum_{m=1}^{N} \mathcal{K}\left(\frac{\mathbf{t}_{ij}(\widehat{\theta}_{ij}) - \mathbf{t}_{ij}(\widehat{\theta}_{ij}^{(s,m)})}{h_{ij}}\right), \text{ (with } \widehat{\theta}_{ij} = r_{ij}^{(s,m)}/n_{ij}) \tag{10c}$$

$$\widehat{L}^*(\mathbf{t}(\widehat{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta}^{(s)}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\widehat{L}_{ij}^*(\mathbf{t}_{ij}(\widehat{\theta}_{ij}) \,|\, \boldsymbol{\theta}^{(s)})\right)_{I \times J}. \tag{10d}$$

Algorithm 2, based on this IS approach and novel SL method, is shown below. It provides the basis for the Bayesian omnibus test of the hierarchy of cancellation conditions (i.e., hypothesis $H_0 : \boldsymbol{\theta} \in A_{HC}$) on a set of proportion data $\widehat{\boldsymbol{\theta}}$. (The labeling of three substeps ((a), (b), (c)) in Algorithm 2 is used to facilitate comparison with Algorithm 1). This algorithm produces sample

output, $(\boldsymbol{\theta}^{(s)}, \boldsymbol{\omega}^{(s)})_{s=1}^{S}$, of the parameters $\boldsymbol{\theta}$ and the corresponding importance sampling weights $(\boldsymbol{\omega}^{(s)})$. Also, it gives a basis for performing this omnibus test while accounting for posterior uncertainty in the subspace $\widehat{A}_{HC}$, as now it can vary over the N synthetic data sets, per sampling iteration; and can vary across sampling iterations $s = 1, \ldots, S$.

Algorithm 2. The SL Importance Sampling Algorithm for testing HC.

for $s = 1$ to S do

(a) Sample from prior, $\boldsymbol{\theta}^{(s)} \sim \prod_{i=1}^{I} \prod_{j=1}^{J} \text{be}(\theta \,|\, a_{ij}, b_{ij})$;

($b1$) Sample $\mathbf{r}^{(s,m)} = (r_{ij}^{(s,m)})_{I\times J} \overset{i.i.d.}{\sim} \prod_{i=1}^{I} \prod_{j=1}^{J} \text{Binomial}(n_{ij}, \theta_{ij}^{(s)})$, for $m = 1, \ldots, N$.

($b2$) Find the Rasch model MLE $\widehat{\boldsymbol{\theta}}_R^{(s,m)} = (\widehat{\theta}_{Rij}^{(s,m)} = \frac{\exp(\widehat{\beta}_i^{(s,m)} - \widehat{\delta}_j^{(s,m)})}{1+\exp(\widehat{\beta}_i^{(s,m)} - \widehat{\delta}_j^{(s,m)})})_{I\times J}$,

and obtain the PAVA estimate $\mathbf{t}(\widehat{\boldsymbol{\theta}}^{(s,m)} = (r_{ij}^{(s,m)}/n_{ij})_{I\times J}) = (\mathbf{t}_{ij}(\widehat{\theta}_{ij}^{(s,m)}))_{I\times J}$, for $m = 1, \ldots, N$

(using weights $\mathbf{n} = (n_{ij})_{I\times J}$), where the cell values of $\mathbf{t}(\widehat{\boldsymbol{\theta}}^{(s,m)})$ are monotonically

(non-decreasing) ordered by the corresponding cell values of the Rasch MLE matrix

$\widehat{\boldsymbol{\theta}}_R^{(s,m)} = \left(\widehat{\theta}_{Rij}^{(s,m)} = \frac{\exp(\widehat{\beta}_i^{(s,m)} - \widehat{\delta}_j^{(s,m)})}{1+\exp(\widehat{\beta}_i^{(s,m)} - \widehat{\delta}_j^{(s,m)})}\right)_{I\times J}$, to define $\widehat{A}_{HC}^{(s,m)}$, where $\mathbf{t}(\widehat{\boldsymbol{\theta}}^{(s,m)}) \in \widehat{A}_{HC}^{(s,m)}$.

(c) Set the importance sampling weights for $\boldsymbol{\theta}^*$ as $\boldsymbol{\omega}^{(s)} = (\omega_{ij}^{(s)} \equiv \widehat{L}_{ij}^{*(s)}(\mathbf{t}_{ij}(\widehat{\theta}_{ij}) \,|\, \boldsymbol{\theta}^{(s)}))_{I\times J}$,

where $\mathbf{t}(\widehat{\boldsymbol{\theta}}) = (\mathbf{t}_{ij}(\widehat{\theta}_{ij}))_{I\times J}$ is the PAVA estimate (using weights, $\mathbf{n}$), such that

the cell values of $\mathbf{t}(\widehat{\boldsymbol{\theta}})$ are monotonically (non-decreasing) ordered by the

corresponding cell values of the Rasch MLE matrix $\widehat{\boldsymbol{\theta}}_R = (\widehat{\theta}_{Rij} = \frac{\exp(\widehat{\beta}_i - \widehat{\delta}_j)}{1+\exp(\widehat{\beta}_i - \widehat{\delta}_j)})_{I\times J}$

obtained from the original data $(\widehat{\boldsymbol{\theta}})$. This defines $\widehat{A}_{HC}$, where $\mathbf{t}(\widehat{\boldsymbol{\theta}}) \in \widehat{A}_{HC}$.

end for

A basic quantity of interest is given by the (approximate) posterior mean $\overline{\boldsymbol{\theta}} = (\overline{\theta}_{ij})_{I\times J}$, with $\overline{\boldsymbol{\theta}} = \int \boldsymbol{\theta} \pi(\boldsymbol{\theta} \,|\, \mathbf{t}(\widehat{\boldsymbol{\theta}})) \mathrm{d}\boldsymbol{\theta}$, subject to the order constraints defined by A_{HC}. Of course, $\overline{\theta}_{ij}$ gives the (approximate) posterior predictive probability of a correct response for test score group i on item

j, under those constraints. So, when analyzing data $\widehat{\boldsymbol{\theta}}$, the algorithm is run for a sufficiently-large number of iterations (S), until the sample output converges to samples to the approximate posterior distribution (6). The estimate of the (approximate) marginal posterior mean $\overline{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, subject to the order constraints of A_{HC}, is given by $\overline{\boldsymbol{\theta}}^{(S)} = (\overline{\theta}_{ij}^{(S)} = \sum_{s=1}^{S} \theta_{ij}^{(s)} \Big/ \sum_{s=1}^{S} \omega_{ij}^{(s)})_{I\times J}$. Effective sample size (ESS) statistics, $\text{ESS}_{ij}^{(S)} = 1/\sum_{s=1}^{S}\{\omega_{ij}^{(s)}/\sum_{s=1}^{S}\omega_{ij}^{(s)}\}^2$, can be used to evaluate the convergence of $\overline{\boldsymbol{\theta}}^{(S)}$ to its true marginal posterior mean, $\overline{\boldsymbol{\theta}}$. $\text{ESS}_{ij}^{(S)}$ ranges between 1 (very poor outcome) to S (perfect outcome; where $(\boldsymbol{\theta}^{(s)})_{s=1}^{S}$ are i.i.d.) (Liu, 2001).

Let $\overline{\theta}_{0ij} = \frac{\frac{1}{2}+r_{ij}}{1+n_{ij}}$ be the posterior mean estimate of θ_{ij} under a non-informative reference be$(\theta_{ij} \mid \frac{1}{2}, \frac{1}{2})$ prior (Bernardo, 1979) and no order constraints on $\boldsymbol{\theta}$, for each cell ij. The hypothesis $H_0 : \boldsymbol{\theta} \in A_{HC}$ can be tested by using the Kullback-Leibler (KL) divergence statistic:

$$\mathcal{D}_{ij}^{(S)} = \overline{\theta}_{0ij} \log\left(\frac{\overline{\theta}_{0ij}}{\overline{\theta}_{ij}^{(S)}}\right) + (1-\overline{\theta}_{0ij}) \log\left(\frac{1-\overline{\theta}_{0ij}}{1-\overline{\theta}_{ij}^{(S)}}\right) \geq 0. \tag{11}$$

Then, the hypothesis $H_0 : \boldsymbol{\theta} \in A_{HC}$ is rejected whenever $\mathcal{D}_{ij}$ is large for one or more cells ij in an $I \times J$ matrix, leading to the conclusion that the given data $\widehat{\boldsymbol{\theta}}$ violate the HC axioms. This hypothesis testing procedure is based on a general method that can be used to evaluate whether or not a simple Bayesian model (here, given by H_0) can provide a reasonable approximation to a more flexible model for the data (McCulloch, 1989) (here, this model has estimator $\overline{\boldsymbol{\theta}}_0$).

There are at least three advantages for using the KL divergence ($\mathcal{D}$) for hypothesis testing (see also, Karabatsos, 2006). First, it can be interpreted as a measure of information loss when approximating the true estimated probabilities ($\overline{\boldsymbol{\theta}}_0$) by H_0. Further, a global KL measure across the IJ cells can be easily computed by $\mathcal{D}_{\cdot\cdot}^{(S)} = \sum_{i=1}^{I}\sum_{j=1}^{J}\mathcal{D}_{ij}^{(S)}$. Second, the KL divergence can be easily calibrated. For example, the KL value $\mathcal{D}_{ij} = .02$ amounts to using a coin that has probability .6 of heads in one flip (i.e., $(1 + [1 - \exp\{-2(.02)\}]^{1/2})/2 = .6$) in place of a fair coin that yields

a probability of heads .5. Similarly, the KL value $\mathcal{D}_{ij} = .14$ amounts to using a coin that has probability .75 of heads instead of a fair coin, and so on (McCulloch, 1989, Table 1). A zero divergence ($\mathcal{D}_{ij} = 0$) corresponds to comparing one fair coin with another.

Third, there are large-sample asymptotic justifications for using the KL divergence ($\mathcal{D}$) method for hypothesis testing, along with Algorithm 2. Specifically, it is easy to show that if $\boldsymbol{\theta}_0$ is the true (population) value of the parameter $\boldsymbol{\theta}$, then as $n_{ij}, N, S \to \infty$, $\overline{\boldsymbol{\theta}}_0 \to \boldsymbol{\theta}_0$, and further when H_0 is true such that $\boldsymbol{\theta}_0 \in A_{HC}$, the approximate posterior density $\pi(\boldsymbol{\theta} \,|\, \mathbf{t}(\widehat{\boldsymbol{\theta}}))$ (in 6) converges to a point mass at $\boldsymbol{\theta}_0$, so that $\mathcal{D}_{ij}^{(S)} \to 0$ for all IJ cells ij. These three advantages are highlighted in the following simulation study.

2.1 *Simulation Study*

First, the axiom testing method was evaluated through its application to simulated data sets of dimension 4×3, involving four ability levels and three test items. Each simulated data analysis reported in this subsection assumed independent reference beta priors, chosen with hyper-prior parameters $a_{ij} = b_{ij} = 1/2$, and was based on running $S = 30\,000$ sampling iterations of Algorithm 2, using $N = 100$. This produced a median $\text{ESS}_{ij}^{(S)}$ value of 2807 over all simulated data sets.

First, data sets were simulated under the Rasch model, which satisfy the hypothesis H_0 : $\boldsymbol{\theta}_R \in A_{HC}$ of the hierarchy of cancellation conditions. Data sets were also simulated under 2-parameter logistic IRT model, $\boldsymbol{\theta}_{\text{2PL}} = \left(\theta_{ij} = \frac{\exp\{\alpha_j(\beta_i - \delta_j)\}}{1+\exp\{\alpha_j(\beta_i - \delta_j)\}}\right)_{4\times 3}$, with item discrimination (slope, α_j) parameters chosen to define a probability matrix that violated the ACM axioms (i.e., $\boldsymbol{\theta}_{\text{2PL}} \notin A_{HC}$). Table 2 presents the simulation study design, including the values of the person ability, item difficulty and discrimination parameters that were used to simulate data, along with the corresponding 4×3 matrix of predicted correct item response probabilities, for the Rasch model and for the 2PL model. Multiple data sets were simulated for sample sizes $n_{ij} = 10, 20, 30, 65, 100,$

per cell ij, respectively, using the matrix of probabilities of the Rasch and of the 2PL model (i.e., $\boldsymbol{\theta}_R$ and $\boldsymbol{\theta}_{2\text{PL}}$, resp.). The underlined probabilities in Table 2 shows where the 2PL model's matrix of probabilities ($\boldsymbol{\theta}_{2\text{PL}}$) violate the single cancellation axiom.

Table 3 shows the results of the analyses of the simulated 4×3 data sets. They are based on analyzing each simulated data set using the Bayesian method for testing $H_0 : \boldsymbol{\theta} \in A_{HC}$.

		Rasch model			2PL model		
	Item:	1	2	3	1	2	3
	Difficulty:	-1.5	0	1.5	-1.5	0	1.5
	Discrimination:	1	1	1	1	2	2.3
Ability (θ)	-1.5	.50	.18	.05	.50	.43	.01
	$-.5$	.73	.38	.12	.73	.48	.07
	.5	.88	.62	.27	.88	.52	.41
	1.5	.95	.82	.50	.95	.57	.88

Table 2: The design of the simulation study.

		Rasch model					2PL model				
		cell sample size, n_{ij}					cell sample size, n_{ij}				
Ability	Item	10	20	30	65	100	10	20	30	65	100
-1.5	1	.06	.00	.00	.00	.00	.21	.00	.00	.00	.00
$-.5$	1	.02	.00	.01	.00	.00	.00	.01	.00	.00	.01
.5	1	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00
1.5	1	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
-1.5	2	.00	.02	.00	.00	.00	.09	.00	.00	.00	.00
$-.5$	2	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01
.5	2	.00	.00	.00	.00	.00	.04	.01	.00	.05	.03
1.5	2	.00	.00	.00	.00	.00	.00	.01	.01	.01	.00
-1.5	3	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
$-.5$	3	.04	.02	.00	.00	.00	.00	.00	.00	.00	.00
.5	3	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
1.5	3	.02	.00	.00	.00	.00	.05	.03	.02	.08	.11
global KL=		.15	.05	.02	.02	.01	.40	.07	.05	.15	.16

Table 3: The Kullback-Leibler divergence results of the simulation study (in 2 significant digits).

The results of Table 3 show that as the sample size increases for all 12 cells, the KL divergence measures correctly converge towards zero for Rasch model simulated data, while the KL measures tend to increasingly exceed beyond .01 for 2PL model data simulated. The latter is true especially for the cell corresponding to ability $\theta = 1.5$ and item 3, which violates the single cancellation axiom,

as mentioned earlier. In conclusion, the simulation study seems to support the use of KL critical values $\mathcal{D}_{ij} > .01$ for rejection the hypothesis $H_0 : \boldsymbol{\theta} \in A_{HC}$ in practice. Further, the KL measures behaved as predicted according to the large sample (n) theory mentioned earlier.

Other cell-based statistics (not shown) for testing H_0 were considered to analyze these simulated data. They include the posterior mean of standardized residual fit statistic of cell count data, and a test of whether the data $\widehat{\theta}_{ij}$ is contained in a 50% (or 95%) posterior credible interval of θ_{ij} under H_0. However, they seemed to reject H_0 too frequently and infrequently, respectfully.

Next, additional data sets were simulated under the Rasch and 2PL model (resp.). This time, a 9×9 design was used along with the same cell sample sizes as in the Parole data (Table 1). For each model, the data were simulated according to person ability (β) and item difficulty parameters (δ) specified by nine equally-spaced points on the interval $[-1.5, 1.5]$, respectively. The Rasch model's item discrimination parameters were set to 1. The 2PL model's item discrimination parameters were specified by nine equally-spaced points on the interval $[.1, 5]$, in order to ensure that the 2PL produced a 9×9 matrix of probabilities that violated the HC axioms.

According to the Bayesian axiom testing method, the 2PL simulated data and the Rasch model simulated data yielded global KL measures of $\mathcal{D}_{\cdot\cdot}^{(S)} = 4.7$ and $\mathcal{D}_{\cdot\cdot}^{(S)} = 2.3$, respectfully. For the 2PL simulated data, 37 of the total 81 cells in the 9×9 matrix obtained a $\mathcal{D}_{ij}^{(S)}$ measure that exceeded .01. Across the 81 cells, the median, third quartile, 90%ile, and maximum KL measures were .01, .04, .16, and .44, respectively.

2.2 *Parole Data Revisited*

The ACM hypothesis $H_0 : \boldsymbol{\theta} \in A_{HC}$ was tested on the Parole data (Table 1), by running Algorithm 2 for $S = 30\,000$ sampling iterations, using $N = 100$, and using independent reference beta priors with hyper-prior parameters $a_{ij} = b_{ij} = 1/2$. This sampling run obtained ESS values with a median

$\text{ESS}_{ij}^{(S)}$ value of 914 across the $IJ = 81$ cells. A longer sampling run would not appear to change the basic conclusions about the following results of this hypothesis test.

The asterisks in Table 1 indicate the Parole data proportions that violate the hierarchy of cancellation axioms. They are indicated by KL measures ($\mathcal{D}_{ij}^{(S)}$) exceeding .01. Across the 81 cells, the median, third quartile, 90%ile, and maximum KL measures were .02, .06, .16, and .48, respectively. The global KL measure was $\mathcal{D}_{\cdot\cdot}^{(S)} = 5.5$. We conclude that the Parole data violate the hierarchy of cancellation axioms. This casts doubt about whether an ACM representation exists for these data.

3 Conclusions

Psychometricians often like to claim that cognitive abilities are continuous quantities that are measured on an interval or ratio scale. According to ACM theory, the existence of such a scale requires that the hierarchy of cancellation axioms hold. This makes it important to devise probabilistic tests of these axioms. One challenge in constructing such a test is that this hierarchy implies a set of highly-interdependent order-constraints on model parameters (e.g., Domingue, 2014), which serve as the basis for axiom testing.

As one possible way to overcome this challenge, this article introduced an omnibus test of the entire hierarchy of cancellation axioms. This is based on a novel synthetic likelihood approach to approximate Bayesian inference. This axiom testing approach was illustrated through the analysis of the Parole data, and simulated data. It was straightforward to run the sampling algorithm. It required writing only 45 lines of MATLAB code, including 6 lines to set up the Parole data analysis. The MATLAB code, the Parole data, and all of the simulated data sets and detailed results of this article, are provided as Supplementary Material. The code provides a simpler interpretation of Algorithm 2.

Finally, while this article focused on testing ACM axioms, the Bayesian testing method can also be extended to test other ordered hypotheses. This can be achieved by making simple adjustments to the PAVA algorithm in the code.

4 Acknowledgements

The author is grateful for the detailed comments and suggestions by two anonymous reviewers and the Editor. They have helped improve the presentation of this article.

References

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

Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, *79*, 1-19.

Gelfand, A., Smith, A., & 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.

Gutmann, M., & Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, *17*, 1-47.

Hoffman, P., & Beck, J. (1974). Parole decision-making: A salient factor score. *Journal of Criminal Justice*, *2*, 195-206.

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. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology*, *50*, 123-148.

Kruskal, W. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*, 115-129.

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology*, *64*, 478-497.

Kyngdon, A., & Richards, B. (2007). Attitudes, order and quantity: Deterministic and direct probabilistic tests of unidimensional unfolding. *Journal of Applied Measurement*, *8*, 1-34.

Li, Z. (2008). *Some Problems in Statistical Inference Under Order Restrictions*. Unpublished doctoral dissertation, University of Michigan.

Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Luce, R., & Steingrimsson, R. (2011). Theory and tests of the conjoint commutativity axiom for additive conjoint measurement. *Journal of Mathematical Psychology*, *55*, 379-385.

Luce, R., & Tukey, J. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1-27.

McCulloch, R. (1989). Local model influence. *Journal of the American Statistical Association*, *84*, 473-478.

Michell, J. (1988). Some problems in testing the double cancellation condition in conjoint measurement. *Journal of Mathematical Psychology*, *32*, 466-473.

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*. New York: Psychology Press.

Perline, R., Wright, B., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237-255.

Price, L., Drovandi, C., Lee, A., & Nott, D. (2017, in press). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, *na*, na-na.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Robert, C., & Casella, G. (2004). *Monte Carlo Statistical Methods (2nd Ed.).* New York: Springer.

Robertson, T., & Warrack, G. (1985). An application of order restricted inference methodology to a problem in psychiatry. *Psychometrika*, *50*, 421-427.

Scott, D. (1964). Measurement models and linear inequalities. *Journal of Mathematical Psychology*, *1*, 233-247.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Boca Raton, Florida: Chapman and Hall.

Wood, S. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, *466*, 1102-1104.

Zhu, W., Marin, J., & Leisen, F. (2016). A bootstrap likelihood approach to Bayesian computation. *Australian and New Zealand Journal of Statistics*, *58*, 227-244.