British Journal of Mathematical and Statistical Psychology (2000), **53**, 275–292 *Printed in Great Britain*

Testing Thurstonian Case V ranking models using posterior predictive checks

Rung-Ching Tsai*

Department of Psychology, University of Illinois at Urbana-Champaign, USA

Grace Yao

Department of Psychology, National Taiwan University, Taiwan

This paper presents the results of a Monte Carlo study which investigates the validity of the method of posterior predictive checks (PPC) for testing the fit of a Thurstonian Case V ranking model. The PPC method is employed as an alternative to standard goodness-of-fit tests which are of limited use even when the number of items to be ranked is small. Several test quantities are formed to assess the fit of the Case V ranking model to data for various sample sizes and for two types of violations of the Case V assumptions: heterogeneous stimulus variances and rankers from different populations. The study concludes that the PPC method is useful in detecting local and global misfits of a Thurstonian Case V model, even when the ranking data are sparse.

1. Introduction

Thurstone's (1927) random utility approach has been influential in the development of ranking models (Böckenholt, 1992; Cohen & Mallows, 1983; Fligner & Verducci, 1986; 1988; Hausman & Wise, 1978; Kamakura & Srivastava, 1984; Marden, 1995; Yao, 1995). By postulating that the random utilities associated with the choice options follow a multivariate normal distribution, Thurstonian models provide a straightforward and appealing representation of ranking data. However, the use of Thurstonian models in applied research has been limited due to difficulties in estimation when a large number of objects are to be ranked. Yao and Böckenholt (1999) adopted the Gibbs sampler (Geman & Geman, 1984) to estimate the parameters of Thurstonian models for ranking data with a large number of objects. As a result, the problem of numerical intractabilities in estimation was solved. However, to ensure meaningful interpretations of the parameter estimates, it is necessary and equally important to check whether the proposed model provides an adequate representation of the data set.

Typically, the Pearson and likelihood ratio statistics are well defined for assessing model fit based on their χ^2 approximations. They are computed on the basis of the discrepancies between observed and expected frequencies of ranking patterns. However, a poor χ^2 approximation for Pearson or likelihood ratio (LR) statistics can result from sparseness

* Requests for reprints should be addressed to Rung-Ching Tsai, Department of Psychology, University of Illinois, 603 E. Daniel Street, Champaign, IL 61820, USA.

and empty cells. This suggests that a different approach is necessary to assess the fit of Thurstonian ranking models when the number of ranked objects, and consequently the number of possible ranking patterns, is large.

Yao & Böckenholt (1999) demonstrated how posterior predictive checks (PPC), a Bayesian tail-area approach (Box, 1980; Guttman, 1967; Rubin, 1984) based on the posterior distributions of the parameters, can be used to assess the fit of Thurstonian models to ranking data with a large number objects. In fact, the posterior distributions of the parameters are readily obtained as by-products of their estimation procedure. However, little is known about the validity of PPC in assessing the fit of a Thurstonian ranking model.

There are three goals in this study. Our first goal is to examine the validity of the PPC approach in fitting a Thurstonian Case V model to two types of non-Case V ranking data, namely Case III data and a mixture of Case V data for both small and large numbers of ranked objects. In addition, we compare the power of PPC and LR statistics for detecting misfit when the χ^2 approximation to the LR statistic is valid. In general, the PPC approach has the appealing feature that it does not rely on the asymptotic distribution of any pivotal quantity such as the LR statistic, so that any potential test quantity can be used to assess model fit (Gelman, Meng & Stern, 1996). The choice of a test quantity is determined by various considerations such as sample size and type of misspecification. Therefore, our second goal is to see whether certain test quantities are particularly effective at revealing overall or certain types of misfit and consequently to find the best test quantity among those under consideration. Different types of test quantity are constructed for assessing the fit of subsets of the data (local), the overall fit of the data (global), and the fit of some qualitative (axiomatic) property of the data. However, it is expected that the effectiveness of PPC will also depend on other factors, such as sample size and the type of misspecification. As a result, our third goal is to investigate the influence of three factors: type of misspecification, sample size, and type of test quantity on the performance of PPC. These issues are addressed with the help of a Monte Carlo simulation study.

This paper is structured as follows. First, Thurstonian ranking models are briefly reviewed and the PPC implementation is presented. Second, the results of several Monte Carlo simulation studies conducted to demonstrate the effectiveness of PPC in misfit detection of a Thurstonian Case V model are discussed. Comparisons between the results from PPC and LR are then presented, and the paper concludes with a discussion of the main results.

2. Thurstonian ranking models

Suppose a subject is asked to rank k objects, $O_1, O_2, \ldots, O_k$, according to some prescribed criterion. According to Thurstone (1927), the unobserved judgments of each stimulus are realizations of a random variable, and the corresponding ranking of a series of stimuli is based on the relative order of the values of the random variables on the underlying judgments. Thurstone (1927) introduced the notion of the so-called 'discriminal process' and proposed that the values assigned to each stimulus are normally distributed. For any subject or a homogeneous group of subjects, the discriminal process associated with any given stimulus can be described by a random variable v_{ij} (Bock & Jones, 1968; Böckenholt, 1990, 1992)

$$v_{ij} = \mu_i + \epsilon_{ij},$$

where μ_i is the 'affective value' which refers to the modal response judgment for object i, and ϵ_j is a random component which reflects the deviation from the modal response for person j on a particular occasion. The rank order of k stimuli is determined by the order of their corresponding random utilities $v_{ij}, i = 1, 2, \ldots, k$. For example, the probability for observing the rank order $s = (1, 3, 2, \ldots, k)$ is

$$\mathrm{P}(s) = \mathrm{P}(v_{1j} > v_{3j} > v_{2j} > \ldots > v_{kj}).$$

It is assumed that ϵ_{ij} follows a normal distribution, $\epsilon_{ij} \sim N(0, \sigma_i^2)$, for $i = 1, 2, \ldots, k$, and therefore the joint distribution of the $\boldsymbol{\epsilon}_j$ is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. As a result, the $\mathbf{v}_j$ follow a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ and covariance matrix $\boldsymbol{\Sigma}$.

Thurstone (1927) proposed the following cases:

- *Case V model*: $\boldsymbol{\Sigma}$ is a diagonal matrix with equal diagonal elements, i.e., $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. However, based on the ranking data, σ cannot be uniquely identified. Without loss of generality, σ^2 can be arbitrarily set to be 1.
- *Case III model*: $\boldsymbol{\Sigma}$ is a diagonal matrix with unequal diagonal elements.

These two cases appear to have drawn much attention throughout the literature because of their simple zero-covariance structures (Burros & Gibson, 1954; Iverson, 1987; MacKay & Chaiy, 1982) and therefore they were used in this study. Moreover, in addition to its simplicity in interpretation, the Case V model was chosen as the null model in this study because it does not require estimating additional parameters for the variance-covariance structure.

3. Posterior predictive checks

In assessing the goodness of fit of ranking models, Pearson and LR statistics are commonly used where each of the possible ranking patterns is considered as a separate cell in a multinomial distribution. However, the problem of large sparse multinomials is encountered when the number of ranked objects is large and the use of χ^2 approximations to Pearson and LR statistics is inappropriate.

The Bayesian posterior predictive checks (Gelman *et al*., 1996; Meng, 1994; Rubin, 1984) do not rely on the asymptotic distribution of any pivotal statistic, such as the LR statistic. In the PPC approach, a test quantity T can be defined as a function of both the data and unknown (nuisance) parameters (Gelman *et al*., 1996).

The PPC technique takes into account the uncertainty of the nuisance parameters by averaging over the influence of different plausible nuisance parameters under the hypothesized model. In this way, the PPC method allows us to measure the direct discrepancy between the hypothesized model and the data, whereas the maximum likelihood approach measures the discrepancy between the best-fitting model and the data.

In a Bayesian framework, a replication $\mathbf{Y}^{\mathrm{rep}}$ is defined as future observed ranking data under the hypothesized model and the same nuisance parameter which produced the current observed ranking data $\mathbf{Y}$. In our context, the hypothesized Case V model defines the variance–covariance structure $\boldsymbol{\Sigma} = \mathbf{I}$ of the ranking data, not the mean structure, $\boldsymbol{\mu}$. Therefore, $\boldsymbol{\mu}$ is regarded as a nuisance parameter of the model. However, both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are required to generate replication of ranking data. After observing $\mathbf{Y}$, it is obvious that not all possible $\boldsymbol{\mu}$

are equally likely to have generated $\mathbf{Y}$. Therefore, we characterize $\boldsymbol{\mu}$ by its posterior distribution under the hypothesized Case V model, $p(\boldsymbol{\mu}|\mathbf{Y}, \mathbf{I})$, to account for the uncertainty of $\boldsymbol{\mu}$. A replication is then defined as a single draw from $p(\mathbf{Y}^{\text{rep}}|\boldsymbol{\mu}, \mathbf{I})$, averaged over the posterior distribution of $\boldsymbol{\mu}$. In other words, the reference distribution of replications ($\mathbf{Y}^{\text{rep}}$) is (Gelman, Carlin, Stern, & Rubin, 1995):

$$p(\mathbf{Y}^{\text{rep}}|\mathbf{Y}, \mathbf{I}) = \int p(\mathbf{Y}^{\text{rep}}|\boldsymbol{\mu}, \mathbf{I}) p(\boldsymbol{\mu}|\mathbf{Y}, \mathbf{I}) \mathrm{d}\boldsymbol{\mu},$$

where $p(\boldsymbol{\mu}|\mathbf{Y}, \mathbf{I})$ is the posterior distribution of $\boldsymbol{\mu}$ under the hypothesized Case V model.

After defining a replication, we can then compute test quantities from the replicated ranking data, $\mathbf{Y}^{\text{rep}}$. If the model fits, the test quantity from the observed data, $\mathbf{Y}$, is expected to be similar to those computed from the $\mathbf{Y}^{\text{rep}}$. Consequently, an extreme test quantity or a systematic difference reveals the misfit of the hypothesized Case V model.

To evaluate how extreme the observed test quantity T is, the posterior predictive p-value (*PPP*) (Rubin, 1984; Meng, 1994; Gelman *et al.*, 1996) is defined as

$$PPP = \mathrm{P}(T(\mathbf{Y}^{\text{rep}}, \boldsymbol{\mu}) \geq T(\mathbf{Y}, \boldsymbol{\mu})|\mathbf{Y}, \mathbf{I}),$$

where $\boldsymbol{\mu} \sim p(\boldsymbol{\mu}|\mathbf{Y}, \mathbf{I})$ and T is a discrepancy measure. Note that if T is a non-directional measure (i.e., values that are too small or too large are both considered extreme), then *PPP* is defined as two-tailed. In other words, *PPP* is the proportion of the test quantities from the replicated data that is more extreme than the test quantity with respect to the observed data. An extreme *PPP* implies that the observed test quantity is unlikely under the hypothesized model, and therefore shows some evidence against the model.

Adopting the PPC method, we obtain replications under both the Case V model and the draws from the posterior distribution of the nuisance parameter $\boldsymbol{\mu}$, $p(\boldsymbol{\mu}|\mathbf{Y}, \mathbf{I})$. To minimize the influence of the prior, an informative but vague prior on $\boldsymbol{\mu}$ such that $\boldsymbol{\mu} \sim N_k(\mathbf{0}, S^2\mathbf{I})$ is employed, where S^2 is relatively large. Assuming independence between the subjects, $p(\mathbf{Y}|\boldsymbol{\mu}, \mathbf{I})$ equals $\prod_{j=1}^{n} p(\mathbf{y}_j|\boldsymbol{\mu}, \mathbf{I})$ for n randomly selected subjects, where $\mathbf{y}_j$ is the ranking vector of subject j. Therefore,

$$\begin{aligned} p(\boldsymbol{\mu}|\mathbf{Y}, \mathbf{I}) &\propto p(\boldsymbol{\mu}) p(\mathbf{Y}|\boldsymbol{\mu}, \mathbf{I}) \\ &\propto p(\boldsymbol{\mu}) \prod_{j=1}^{n} p(\mathbf{y}_j|\boldsymbol{\mu}, \mathbf{I}). \end{aligned}$$

Under the Thurstonian model, $p(\mathbf{y}_j|\boldsymbol{\mu}, \mathbf{I})$ can be determined by evaluating a $(k-1)$-variate normal distribution. However, $p(\boldsymbol{\mu}|\mathbf{Y}, \mathbf{I})$ does not have a known density form and hence the estimation of $p(\boldsymbol{\mu}|\mathbf{Y}, \mathbf{I})$ is somewhat difficult.

Fortunately, the computational difficulty in obtaining the posterior distribution of $\boldsymbol{\mu}$ can be overcome by adopting the Gibbs sampler (Yao & Böckenholt, 1999). The Gibbs sampler algorithm (Geman & Geman, 1984) simplifies the computation by reducing the simulation from a high-dimensional distribution into iterative draws from lower-dimensional fully conditional distributions. In this study, we adopt the Gibbs sampler technique to construct the posterior distribution of $\boldsymbol{\mu}$ (see Appendix). After obtaining the draws from the posterior distribution of $\boldsymbol{\mu}$ to generate replications, we can construct a posterior predictive distribution of a PPC test quantity, T, as a by-product of the replicated ranking data. In summary, the

following strategy is used to obtain a posterior p-value for assessing model misfit in a Bayesian framework:

1. Draw samples $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_m$ from the posterior distribution of $\boldsymbol{\mu}$ under the hypothesized Case V model, i.e., $p(\boldsymbol{\mu}|\mathbf{Y}, \mathbf{I})$.
2. Generate the replication ranking data $\mathbf{Y}_l^{\text{rep}}$ based on $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma} = \mathbf{I}$ for $l = 1, 2, \ldots, m$.
3. Select a test quantity T and compute both $T(\mathbf{Y}, \boldsymbol{\mu}_l)$ and $T(\mathbf{Y}_l^{\text{rep}}, \boldsymbol{\mu}_l)$ for $l = 1, 2, \ldots, m$.
4. Calculate $PPP = (1/m)\Sigma_l I_l$, where the indicator variable I_l is 1 if $T(\mathbf{Y}_l^{\text{rep}}, \boldsymbol{\mu}_l) \geq T(\mathbf{Y}, \boldsymbol{\mu}_l)$ and zero otherwise.
5. Use the obtained PPP to assess the validity of the hypothesized Case V model.

4. Simulation studies

The purposes of the study were (a) to compare the diagnostic power of the various test quantities under consideration, (b) to compare the PPC approach to the LR test with various sample sizes, and (c) to investigate how the three factors—type of misspecification, sample size, and type of test quantity—influence the efficacy of the misfit detection based on PPC.

Two simulation studies were conducted to assess the validity of the PPC method in checking the (mis)fits of the Case V model to ranking data with either four or seven objects. In order to assess the validity of the PPC method, 50 ranking data sets were generated under each factor combination, and the frequency of misfit detections was observed.

4.1. Type of misspecification

The Case V model has drawn the most attention throughout the literature because it is easy to estimate and interpret. However, because of its simplicity, the Case V model does not always provide an adequate representation of the data.

The Case III model relaxes the equal-variances assumption of Case V by allowing the variances to vary. Furthermore, in contrast to Case III and Case V where the judges are assumed to be from a homogeneous population, the *mixture model* can be formulated which assumes that the judges are from multiple (two in this study) homogeneous populations (see Böckenholt, 1993):

- Mixture model : $\mathbf{v} \sim pN_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_k(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.

In particular, the two populations chosen in this study consist of one group of individuals whose utilities follow a Case V model and the other group of individuals who simply randomly assign the ranks to the objects. Case III and mixture data were used as two types of violations of the Case V assumptions, namely heterogeneous stimulus variances and rankers from different populations. In addition, Case V data sets were also simulated as a reference to indicate how the PPPs of the test quantities behave when the true model is fitted to the data. That is, we considered the following three cells:

data	Case III(A), (B)	Mixture (A), (B)	Case V
fit	Case V	Case V	Case V

The criteria used to generate the three types of data are summarized in Table 1.

Table 1. Criteria used to generate simulated ranking data based on specific mean vectors and covariance structures for $k = 4$ and $k = 7$

Simulated data	Affective values	Covariance structure
$k = 4$		
Case III (A)	$\boldsymbol{\mu} = (-1 \; -0.33 \; 0.33 \; 1)'$	$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1, \; \sigma_4^2 = 4, \sigma_{ij} = 0 \, \forall \, i \neq j$
Case III (B)	$\boldsymbol{\mu} = (-1 \; -0.67 \; 0.67 \; 1)'$	$\sigma^2 = 4, \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 1, \sigma_{ij} = 0 \, \forall \, i \neq j$
Mixture (A)	$\boldsymbol{\mu}_1 = (-1.33 \; -0.44 \; 0.44 \; 1.33)'$	
Mixture (B)	$\boldsymbol{\mu}_1 = (-1.5 \; -0.1 \; 0.1 \; 1.5)'$	
	$\boldsymbol{\mu}_2 = (0 \; 0 \; 0 \; 0)'$	$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_4$
	$\boldsymbol{\mu} = p\boldsymbol{\mu}_1 + (1-p)\boldsymbol{\mu}_2, p = 0.7$	
Case V	$\boldsymbol{\mu} = (-1 \; -0.33 \; 0.33 \; 1)'$	$\boldsymbol{\Sigma} = \mathbf{I}_4$
$k = 7$		
Case III	$\boldsymbol{\mu} = (-1.5 \; -1 \; -0.5 \; 0 \; 0.5 \; 1 \; 1.5)'$	$\sigma_1^2 = \ldots = \sigma_6^2 = 1, \sigma_7^2 = 4, \sigma_{ij} = 0 \, \forall \, i \neq j$
	$\boldsymbol{\mu}_1 = (-1.5 \; -0.1 \; -0.5 \; 0 \; 0.5 \; 1 \; 1.5)'$	
Mixture	$\boldsymbol{\mu}_2 = (0 \; 0 \; 0 \; 0 \; 0 \; 0 \; 0)'$	$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_7$
	$\boldsymbol{\mu} = p\boldsymbol{\mu}_1 + (1-p)\boldsymbol{\mu}_2, p = 0.7$	
Case V	$\boldsymbol{\mu} = (-1.5 \; -1 \; -0.5 \; 0 \; 0.5 \; 1 \; 1.5)'$	$\boldsymbol{\Sigma} = \mathbf{I}_7$

4.2. Sample sizes

For each type of misspecification with four objects, sample sizes of 50, 100, 150 and 250 were used to evaluate the effect of sample size on misfit detection. However, when the number of ranked objects is large, all practical finite samples represent small samples relative to the total possible ranking outcomes. Therefore, a sample size of 100 was arbitrarily chosen for the case of seven objects.

4.3. Test quantities

To determine whether the replicated and the observed ranking data are likely to be from the same population, several test quantities were computed for both data sets, and their posterior predictive distributions were constructed. Moreover, the *PPP*s associated with the observed test quantities were obtained.

Although choosing a test quantity can be quite arbitrary, the goal is to use test quantities that are effective in revealing misfit in special features of the data. Three types of discrepancy measure are used in this study. Local test quantities are computed for each object or subset of the data to assess the lack of fit for subsets of objects. Global test quantities measure the overall (mis)fits of the data based on the discrepancy between the observed and expected frequencies or probabilities. Moreover, an axiomatic statistic is constructed on the basis of an important axiomatic property of the Case V model.

4.3.1. Local test quantities. Cohen & Mallows (1983) suggested partitioning the ranking data into disjoint groups, where each group is regarded as a cell to avoid the sparseness problem. In particular, they proposed comparing the observed and expected frequencies for each pair of ranked objects in order to assess the fit of the model when the number of ranked

objects is large. In their estimation of the expected frequency for each pair, the sample mean and variance of the ranks were used as estimates of the population mean (μ) and error variance (σ^2) in a Thurstonian Case V model. Since these values are informative in describing a Thurstonian model, they were both used as local test quantities, and are denoted as follows:

- Mean of the ith object's ranks for n subjects (MR),

$$MR(i) = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \text{ for object } i \ (i = 1, 2, \ldots, k).$$

- Variance of the ith object's ranks for n subjects (VR),

$$VR(i) = \frac{1}{n-1}\sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \text{ for object } i \ (i = 1, 2, \ldots, k).$$

Here $\mathbf{Y}$ are the ranking data, $\mathbf{Y}_j$ is the ranking vector of subject j, y_{ij} is the rank that subject j assigned to object i, and n is the number of subjects.

4.3.2. Global test quantities. Yao & Böckenholt (1999) computed three discrepancy measures to assess the fit of a Thurstonian model based on paired comparison, triple and quadruple rankings in a data set. The discrepancy measures are the sums over the discrepancy for each pair, triple, or quadruple. Similarly, we used the sum of squared normal deviates (Cohen & Mallows, 1983) based on pairs and triples as the global discrepancy measures:

- Discrepancy measure based on paired comparisons (PA),

$$PA(\mathbf{Y}, \boldsymbol{\mu}) = \sum_{i<j} \frac{(P_{ij} - \hat{P}_{ij})^2}{\hat{P}_{ij}(1 - \hat{P}_{ij})}, \quad \hat{P}_{ij} = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}}}\right).$$

- Discrepancy measure based on triple rankings (TRI),

$$TRI(\mathbf{Y}, \boldsymbol{\mu}) = \sum_{i<j<l} \frac{(P_{ijl} - \hat{P}_{ijl})^2}{\hat{P}_{ijl}(1 - \hat{P}_{ijl})}, \quad \hat{P}_{ijl} = \Phi_2\left(\begin{pmatrix} d_{ij} \\ d_{jl} \end{pmatrix}, \begin{pmatrix} 1 & \rho_{ijl} \\ \rho_{ijl} & 1 \end{pmatrix}\right).$$

Here P_{ij} is the probability that object i is preferred to object j in the data, $\hat{P}_{ij}$ is its expected probability, $d_{ij} = (\mu_i - \mu_j)/\left(\sqrt{\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}}\right)$ and $\rho_{ijl} = (\sigma_{ij} - \sigma_{il} - \sigma_j^2 + \sigma_{jl})/(d_{ij}d_{jl})$. In fact, since the Case V model was fitted to the data in this study, we can further set $\sigma_i^2 = 1$ and $\sigma_{jl} = 0$ for $i, j, l = 1, \ldots, k$ and $j \neq l$.

In addition, a chi-square type discrepancy measure which resembles the classical χ^2 goodness-of-fit measure was used (Gelman, Meng & Stern; 1993, 1996):

- Discrepancy measure (X^2) similar to the Pearson statistic,

$$X^2(\mathbf{Y}, \boldsymbol{\mu}) = \sum_{c=1}^{k!} \frac{(y_c - E(y_c|\boldsymbol{\mu}))^2}{\mathrm{Var}(y_c|\boldsymbol{\mu})}.$$

Here c is a possible ranking pattern.

To compute the discrepancy measure X^2, the cell frequency for each ranking pattern is required. However, as the number of ranked objects grows large, it is infeasible to list cell

frequencies for all possible patterns. Therefore, the test quantity X^2 was excluded for the data with seven objects. Instead, an object by rank position matrix was computed from each ranking data set. The object by rank matrix ($\mathbf{M}$) consists of counts $m_{(r,c)}$ representing the number of times the rth object is ranked the cth most preferred object:

- $\mathbf{M}$-based discrepancy measure (MX^2),

$$MX^2(\mathbf{Y}, \boldsymbol{\mu}) = \sum_{c=1}^{k^2} \frac{(y_c - E(y_c|\boldsymbol{\mu}))^2}{\mathrm{Var}(\mathrm{y_c}|\boldsymbol{\mu})}.$$

Here c is a cell in $\mathbf{M}$ instead of a possible ranking pattern. Note that $\mathbf{M}$ is the table of first-order summary in Diaconis (1989). Tables of higher-order summary can also be used to form chi-square discrepancy measures.

4.3.3. Axiomatic test quantity. The underlying assumptions or specific properties of the fitted model can also be used to create test quantities for assessing model fit using PPC (Hoijtink & Molenaar, 1997). For the Thurstonian Case V model with distinct affective values for the ranked objects, strong stochastic transitivity (SST), strong unimodality, and complete consensus are satisfied (Bossuyt, 1990; Critchlow, Fligner, & Verducci, 1991; Falmagne, 1985; Henery, 1981). A Case III model may only satisfy weak stochastic transitivity (WST) but not SST. A mixture of Case V models may not even satisfy WST. Violations of these properties sometimes occur due to sampling variability. However, we expect the systematic violations of SST under some Case III and mixture data to occur more frequently than the violations due to sampling variability under Case V data.

SST and WST are defined as follows: for any triple (i, j, k) in a given choice set, suppose $P_{ij} \geq \frac{1}{2}$ and $P_{jk} \geq \frac{1}{2}$. Then:

1. WST holds if $P_{ik} \geq \frac{1}{2}$;
2. SST holds if $P_{ik} \geq \max\{P_{ij}, P_{jk}\}$.

Here P_{ij} is the probability that object i is preferred to object j. We use the test quantity *VSST* (violation of strong stochastic transitivity) to evaluate the goodness of fit of the property of SST. VSST is defined as follows:

- Violation of strong stochastic transitivity,

$$VSST_{(i_1,\ldots,i_t,\ldots,i_{\binom{k}{3}})} = (d_{i_1}, \ldots, d_{i_t}, \ldots, d_{i_{\binom{k}{3}}}),$$

where $d_{i_{jt}} = 1$ if SST is violated for the triple set i_t and 0 if not. And the strict partial order with transitive indifference (Michell, 1990) imposed upon the possible patterns of VSST is defined by

$$VSST_{(i_1,i_2,\ldots,i_{\binom{k}{3}})} \geq VSST^*_{(i_1,i_2,\ldots,i_{\binom{k}{3}})} \text{ if } \mathrm{d}_{i_1} \geq \mathrm{d}^*_{i_1}, \mathrm{d}_{i_2} \geq \mathrm{d}^*_{i_2}, \ldots, \text{ and } d_{i_{\binom{k}{3}}} \geq d^*_{i_{\binom{k}{3}}}.$$

When SST is satisfied for a data set, $VSST = (0, 0, \ldots, 0)$ is the least in terms of this relation ($\succeq$). Note that this relation does not define a strict simple but a strict partial order among the response patterns. For instance, (0,0,1,1) and (0,1,0,1) are considered indifferent in this case. The *PPP* of VSST is defined as

$$PPP = P(VSST^{\mathrm{rep}} \succeq VSST^{\mathrm{obs}}).$$

For example if $VSST^{\text{obs}} = (1,1,0,0)$, then $PPP = P(VSST_i^{\text{rep}} \succeq (1,1,0,0)) = P(VSST_i^{\text{rep}} \in \{(1,1,0,0),(1,1,1,0),(1,1,0,1),(1,1,1,1)\})$.

Finally, an analysis of variance (ANOVA) of the three factors—types of misspecification (M), samples sizes (N) and various test quantities (T)—was performed on the posterior p-values, with all interaction terms included in the analysis. Since the dependent variable is the obtained p-value, a $2\sin^{-1}\sqrt{p}$ transformation was applied to reduce dependencies between the means and variances of the dependent variables.

5. Results

5.1. Simulations with $k = 4$ objects

For the various test quantities described in the previous section, we constructed posterior predictive distributions to assess the fit of the posited model (Case V). For illustration, results of most of the test quantities, for a single Case III(A) simulated data set, with sample size $n = 150$ and for $k = 4$ objects, are presented in Fig. 1. The results of $MR(i)$ and $VR(i)$ ($i = 1, 2$ and 3) are similar to those of $MR(4)$, therefore they are not presented. The posterior predictive distribution of $VSST$ is also not presented since there is no natural ordering of all the response patterns.

Figure 1 shows histograms of 100 simulations from the posterior predictive distribution of the test quantities, $T(\mathbf{Y}^{\text{rep}})$. The observed test quantities are represented as vertical lines at $T(\mathbf{Y})$ in the histograms. In addition, the scatterplots for generalized test quantities show the 100 simulations of $T(\mathbf{Y}^{\text{rep}}, \boldsymbol{\mu})$ versus $T(\mathbf{Y}, \boldsymbol{\mu})$ along with the lines representing the points where $T(\mathbf{Y}^{\text{rep}}, \boldsymbol{\mu}) = T(\mathbf{Y}, \boldsymbol{\mu})$. The PPPs are reported at the top of each plot. The PPP of $VSST$ equals 0.11.

As can be seen in Fig. 1, most of test quantities have extreme PPPs, which are clear signs of model misfit. However, no extreme PPP is obtained for $MR(4)$, PA and $VSST$. The results suggest that some, but not all, test quantities are effective in detecting misfit using PPC.

Instead of only looking at a single dataset, Fig. 2 presents the distributions of PPPs based on 50 replications for each test quantity of all the simulated data with $n = 150$.

As can be seen in Fig. 2, the distributions of $MR(4)$ and PA from misspecified data sets differ only slightly from those from correctly specified Case V data. This implies that neither $MR(4)$ nor PA is useful in detecting misfit since an extreme PPP is never obtained. On the other hand, the distributions of PPPs for the other test quantities using the misspecified data significantly differ from those for the Case V data. This suggests that these test quantities could be used as discrepancy measures. As a result, the mean of an object's rank position and the discrepancy based on paired comparisons were eliminated from all analyses discussed below.

Our analysis shows clearly that paired comparisons outcomes (PA) are not effective in assessing misfit. Instead we recommend fit assessment on the basis of triple rankings (TRI). The computation of TRI requires the standardized differences and the correlation between two pairs within each triple set. In contrast, PA only uses the standardized differences of each pair. Thus, the 'correlation' in TRI is an important factor in revealing the misfit of a Case V model.

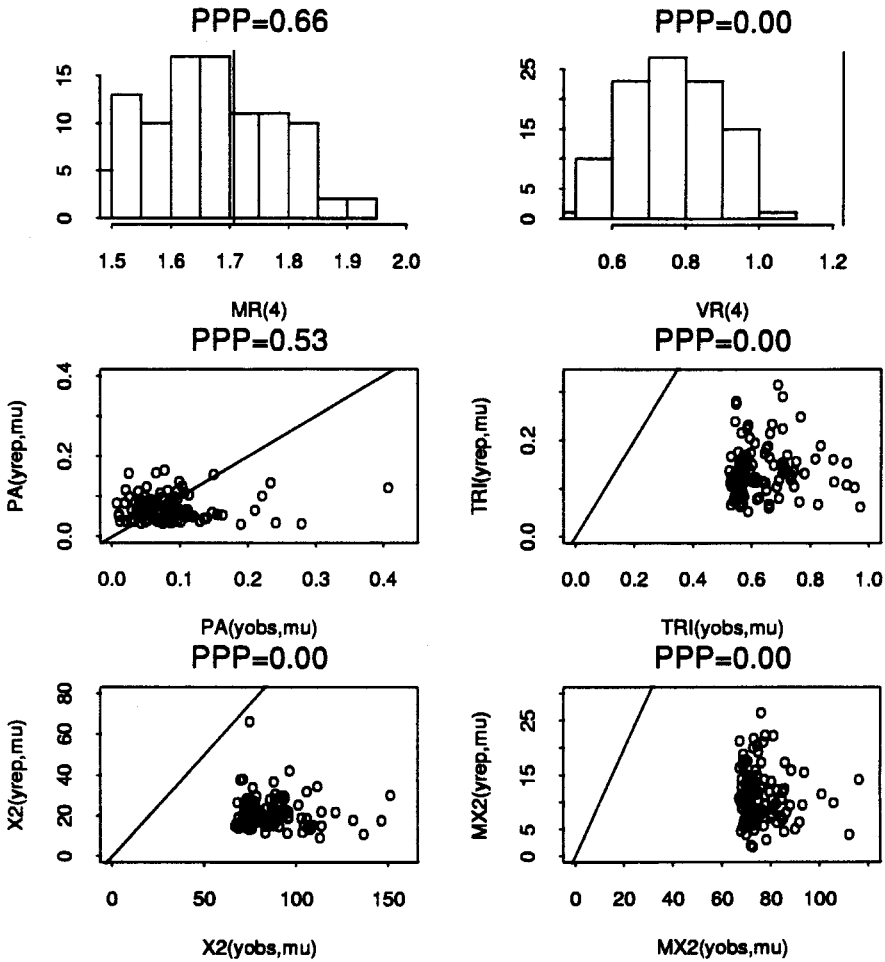

Figure 1. Posterior predictive distributions of $MR(4)$, $VR(4)$, PA, TRI, X^2 and MX^2 for a single Case III(A) data set with $k = 4$ and $n = 150$.

5.1.1. Comparison between LR and PPC. Although a cutoff p-value is seldom used as a decision rule in Bayesian modeling, in order to compare the PPC approach to the LR test it is useful to choose a single cutoff point to define extremeness in the PPC.

In contrast to the fact that in classical testing the p-values follow a uniform distribution under the true model regardless of the choice of the test statistics, the distributions of the *PPP*s under the Case V model (i.e., the correctly specified model) vary among the test quantities in Fig. 2. Thus, the extremeness of a *PPP* depends on the statistic under consideration. In fact, the histograms deviate from a uniform distribution even for sample sizes as large as 250 (not shown in the figures).

To investigate the adequacy of a cutoff *PPP*, the proportion of misrejections for Case V data was obtained when the cutoff *PPP* was used as a decision rule to reject the Case V model. Our goal was to choose a single cutoff *PPP* so that the proportions of misrejection for

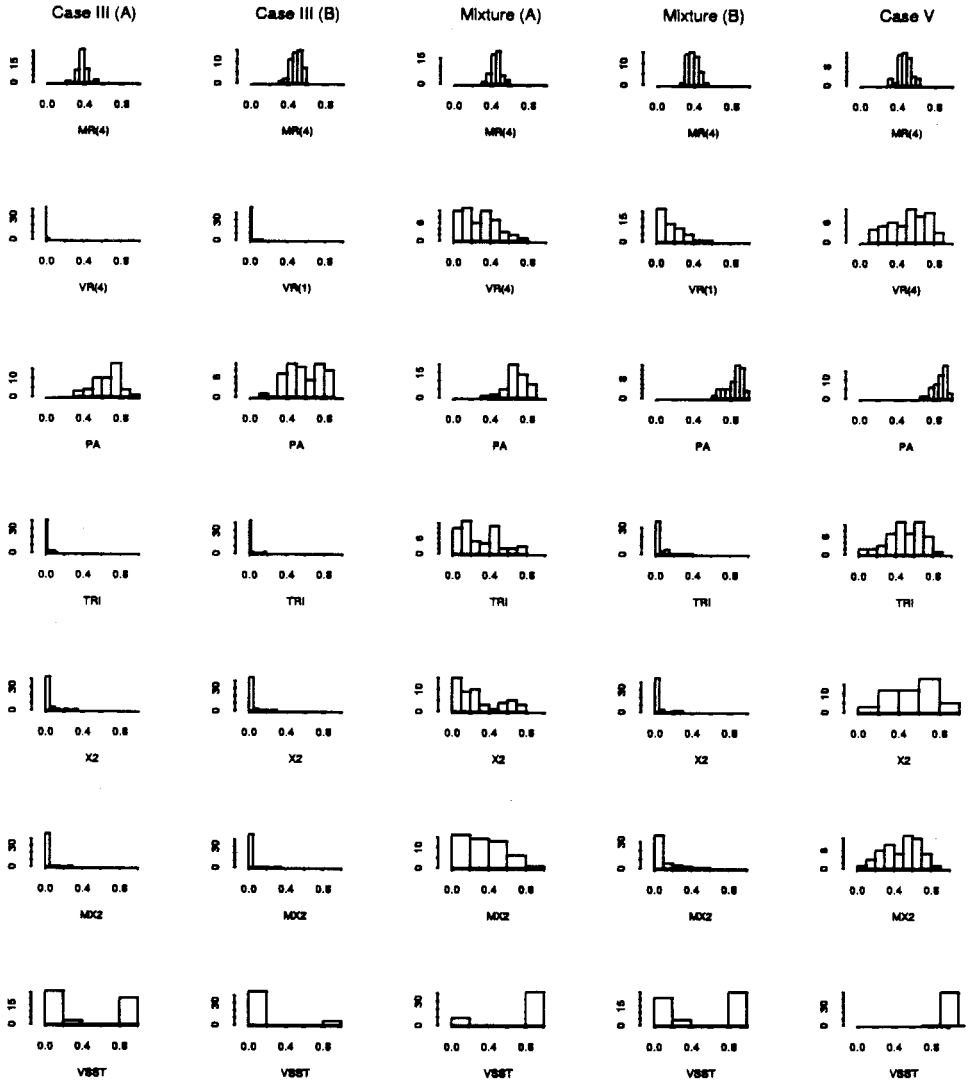

Figure 2. Distributions of posterior predictive p-values of $MR(4)$, $VR(4)$ (or $VR(1)$), PA, TRI, X^2, MX^2, and $VSST$, for 50 replications of the three types of data with $k = 4$ and $n = 150$.

all test quantities were at most 0.05. Examination of the proportions of misrejection indicated that a cutoff PPP of 0.05 was reasonable for all statistics under consideration. Consequently, the cutoff PPP of 0.05 was chosen to evaluate the extremeness of PPPs. Table 2 presents the proportions of extreme p-values from LR and PPC for the four non-Case V and the Case V data.

Surprisingly, for Case III(A) data the proportions of misfit detection for $VR(4)$ from PPC are equal to or larger than those from LR. Other test quantities also provide comparable results for PPC and LR for both Case III(A) and (B) data. For mixture (A) data, neither the test

Table 2. Proportions of extreme p-values for likelihood ratio test (LR) and posterior predictive checks (PPC) based on all the test quantities for $k = 4$ and $n = 50$, 100, 150 and 250

	LR	PPC					LR	PPC				
		$VR(4)$	TRI	X^2	MX^2	$VSST$		$VR(1)$	TRI	X^2	MX^2	$VSST$
n	Case III (A)						Case III (B)					
50	0.34	0.70	0.26	0.26	0.28	0.10	0.36	0.60	0.16	0.16	0.36	0.24
100	0.66	0.92	0.74	0.60	0.80	0.48	0.86	0.92	0.82	0.68	0.88	0.12
150	0.90	1.00	0.90	0.86	0.96	0.18	0.92	0.96	0.92	0.88	0.94	0.32
250	1.00	1.00	1.00	0.96	1.00	0.14	0.98	1.00	1.00	0.96	1.00	0.12
	Mixture (A)						Mixture (B)					
50	0.22	0.04	0.08	0.14	0.06	0.06	0.34	0.04	0.14	0.32	0.22	0.18
100	0.12	0.10	0.10	0.14	0.10	0.00	0.82	0.18	0.60	0.78	0.68	0.20
150	0.24	0.14	0.10	0.18	0.10	0.10	0.86	0.16	0.74	0.90	0.58	0.18
250	0.36	0.32	0.28	0.26	0.20	0.04	1.00	0.24	0.98	1.00	1.00	0.14
	Case V											
50	0.14	0.00	0.00	0.06	0.02	0.18						
100	0.04	0.02	0.02	0.04	0.02	0.00						
150	0.04	0.00	0.04	0.00	0.02	0.00						
250	0.04	0.00	0.02	0.06	0.04	0.02						

Note: A cutoff of 0.05 for p-values is used for both LR and PPC.

quantities nor LR have enough power to consistently detect the misfit. However, in both of the mixture (A) and (B) data, the X^2 statistic appears to have proportions of misfit detection similar to LR. These results suggest that PPC is somehow less powerful but still a useful alternative to LR in the sense that this method provides comparable results.

Not surprisingly, the effectiveness of a test quantity seems to depend on the type of misspecification. The VRs contain information about the relative magnitudes of the variances(σ^2). Consequently, $VR(4)$ and $VR(1)$ appear to be effective in detecting misfit for the Case III (A) and (B) data, respectively, but not for the mixture data. VSST does not seem to be effective in detecting misfit although its distributions under the misspecified data sets differ from the ones obtained under the Case V data. As a result, VSST was also eliminated from all analyses discussed below. However, these results are to be expected for the Case III (A) and mixture data since the particular parameter sets used to generate these data satisfy SST. In other words, their SST violations are simply due to sampling variability. In contrast, the parameter set for Case III(B) does not satisfy SST. Consequently, its distribution of PPPs for VSST differs significantly from those of other data sets. Although its proportions of extreme PPPs (using a cutoff of 0.05) do not seem satisfactory, it is expected that VSST will have more power for data with a higher degree of SST violations.

When we compare the extremeness of the PPPs for $n = 50$ and $n = 250$, the misfit detection for cases with $n = 250$ seems significantly better than for those with $n = 50$. In fact, almost no test quantities effectively reveal any kind of misfit for sample sizes as small as 50. Obviously, the effectiveness of a test quantity is sensitive to sample size. Furthermore, the sensitivity to sample size seems to differ across test quantities. Therefore, we further

investigated the effects of sample size and misspecification type on the misfit detection ability of the test quantities.

5.1.2. Effects of factors. The results of the three-factor (sample size, type of misspecification, and test quantities) ANOVA of the transformed posterior *p*-values are presented in Table 3. The data sets used in the analysis are Case III(A) and mixture (B) because they have comparable LR statistics. Due to the large number of *PPP*s from all test quantities for the 50 replications in the ANOVA, all effects appear to be statistically significant. Therefore, the strength of association measure, ω^2, was used to measure the practical significance of effects. Both the *p*-values and ω^2s for all effects are reported in Table 3.

According to the guidelines suggested by Cohen (1988), all but misspecification type (*M*) show large effects (i.e., $\omega^2 \geq 0.14$). To interpret the results in terms of the effectiveness of the chosen test quantities, there are differences between their effectiveness but no test quantity is a superior indicator of misfit for both types of misspecification. In fact, the existence of a large *MNT* effect implies that the effectiveness of a test quantity depends both on its sensitivity to sample size and on the type of misspecification.

5.2. Simulations with k = 7 objects

To investigate the behaviour of PPC in small samples, $k = 7$ objects were used. The classical χ^2 test is not useful in this case because of the large number of sparse cells. The sample size $n = 100$ was used across all types of misspecification. The posterior predictive distributions of various test quantities based on 50 replications were constructed to assess the lack of fit of the posited Case V model. The results are presented in Fig. 3 and Table 4.

Figure 3 presents the distributions of *PPP*s based on 50 replications for selected test quantities of all the simulated data with $n = 100$. As shown in Fig. 3, the distributions of *PPP*s of *VR*(7), *TRI* and MX^2 from the Case III data are very different from the Case V data. That is, they are useful in assessing model misfit in small samples. Their proportions of misfit detection are reported in Table 4. *VR*(7) and *TRI* appear more effective for

Table 3. Analysis of variance table for the three-factor study on the posterior *p*-values for $k = 4$ objects. The three factors are: type of misspecification (*M*), sample size (*N*), and test quantity (*T*); the datasets are Case III (A) and mixture (B)

Source	SS	df	MS	*F*	Pr > F	$\hat{\omega}^2$
M	25.645	1	25.645	171.97	0.0001	0.097
N	320.863	3	106.954	717.20	0.0001	0.573
T	80.412	3	26.804	179.74	0.0001	0.251
*M*N*	102.074	3	34.025	228.16	0.0001	0.299
*M*T*	346.417	3	115.472	774.32	0.0001	0.592
*N*T*	148.346	9	16.483	110.53	0.0001	0.381
*M*N*T*	177.919	9	19.769	132.56	0.0001	0.426
$n = 1600$			MSE = 0.149	$R^2 = 0.837$		

Note: $\omega^2 = 0.01$ is a small association; $\omega^2 = 0.06$ is a medium association; and $\omega^2 = 0.14$ or greater is a large association.

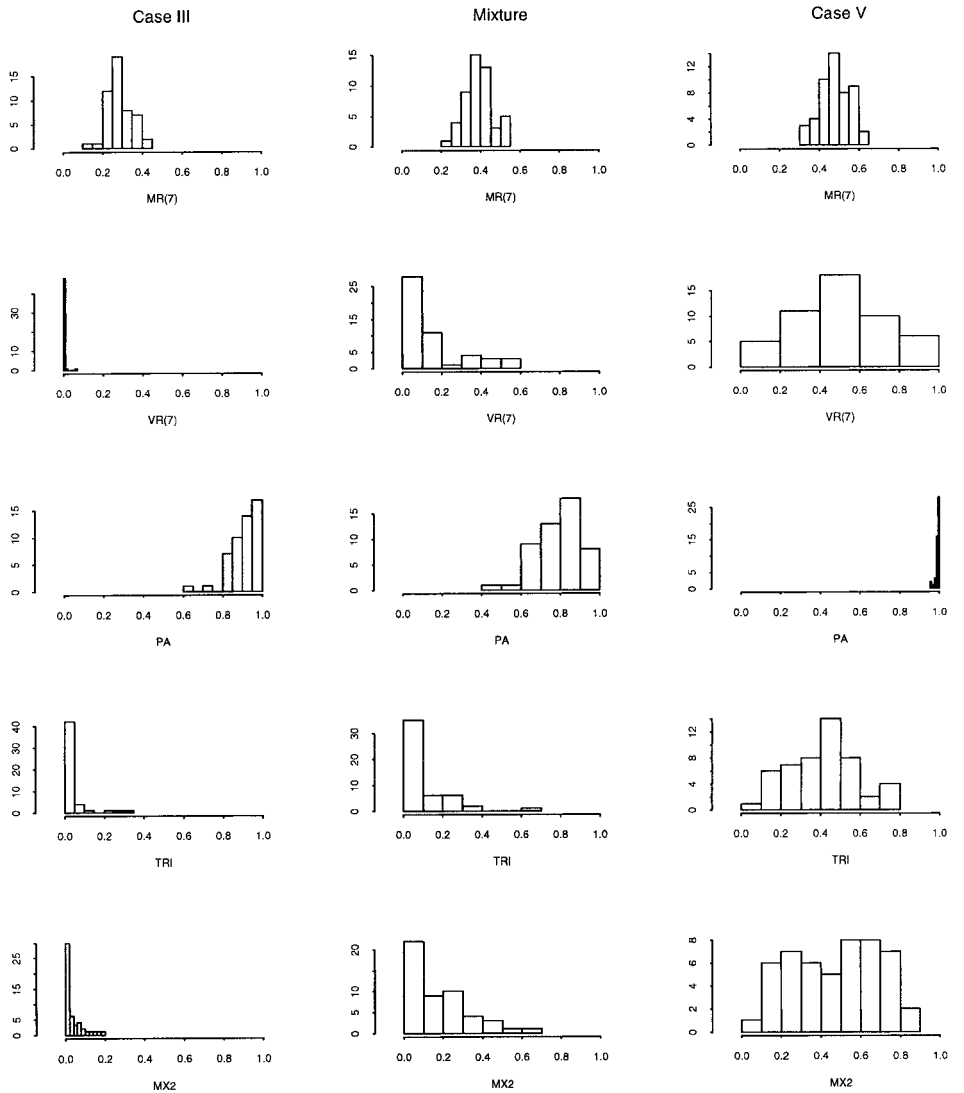

Figure 3. Distributions of posterior predictive p-values of $MR(7)$, $VR(7)$, PA, TRI, and MX^2 for 50 replications of the three types of data with $k = 7$ and $n = 100$.

detecting misfit for Case III and mixture data, respectively. MX^2 is effective for Case III data as well.

In summary, among all the test quantities, VR, TRI, and MX^2 appear most effective in detecting the misfit of the Case V model to Case III data. However, only TRI showed some power in detecting the misfit when data were generated under a mixture of Case V model. The results are similar to those obtained from the simulations with $k = 4$ objects.

Table 4. Proportions of extreme posterior predictive p-values of effective test quantities in the case of $k = 7$ and $n = 100$ (cutoff for extreme PPPs $= 0.05$).

Test Quantities	Case III	Mixture	Case V
$VR(7)$	0.98	0.34	0.00
TRI	0.88	0.58	0.00
MX^2	0.78	0.30	0.00

6. Conclusions

The goal of this paper was to evaluate the validity of the PPC approach and to search for effective test quantities in assessing misfit for Thurstonian Case V models. Two simulation studies were conducted, and the PPC technique was shown to be useful in assessing the misfit of the Thurstonian Case V model.

The validity of the PPC method was verified by comparison with the LR method for $k = 4$ objects, where the classical likelihood approach was feasible because in this case sparseness is not a problem. The results of the comparison between PPC and LR suggest that the power of the PPC method is close to the classical LR approach.

In searching for useful test quantities, the discrepancy measure based on triple rankings (TRI), and the generalized chi-square type test quantities (X^2 and MX^2), appeared effective in revealing the misfit for both Case III and mixture data. Moreover, the variance of an object's rank position (VR) is a powerful test quantity for testing misfit for Case III, but not for mixture data.

In an analysis of the influence of different test quantities, type of misspecification and sample size on misfit detection, we found that the effectiveness of the PPC method was dependent on all three factors. The presence of a three-way interaction for these three factors implied that the interaction of the effectiveness of test quantities with types of misspecification was dependent on sample size. None of the test quantities proved to be a consistently superior misfit indicator of a Case V model for both types of misspecification. Although most test quantities tend to become better at detecting misfit as the sample size increased, their sensitivity towards sample size differed.

Most importantly, PPC is valuable when the number of objects to be ranked is large and the LR approach is not feasible. In the study with a sample size of 100 with $k = 7$ objects, where over 98% of the possible ranking patterns are not observed, the MX^2 revealed the overall misfit and both the VR and TRI test quantities also suggested misfit of the Case V model for Case III data. These results demonstrate that the PPC method provides a useful alternative towards assessing the misfit of a Thurstonian ranking model, especially for sparse tables where the classical approaches are inappropriate.

Acknowledgements

The authors gratefully acknowledge the valuable comments and suggestions of Ulf Böckenholt, John Marden, Razia Azen, and Tom Smith. The authors would also like to thank the two reviewers for their constructive comments which led to significant improvements of the paper.

References

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choices.* San Francisco: Holden-Day.

Böckenholt, U. (1990). Multivariate Thurstonian models. *Psychometrika, 55,* 391–403.

Böckenholt, U. (1992). Thurstonian models for partial ranking data. *British Journal of Mathematical and Statistical Psychology, 45,* 31–49.

Böckenholt, U (1993). Estimating latent distributions in recurrent choice data. *Psychometrika, 58,* 489–509.

Bossuyt, P. (1990). *A comparison of probabilistic unfolding theories for paired comparison data.* New York: Springer-Verlag.

Box, G. E. P. (1980). Sampling and Bayes inferences in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A, 143,* 383–430.

Burros, R. H., & Gibson, W. A. (1954). A solution for Case III of the law of comparative judgment. *Psychometrika, 19,* 57–64.

Cohen, A. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, A., & Mallows, C. L. (1983). Assessing goodness of fit of ranking models to data. *The Statistician, 32,* 361–373.

Critchlow, D. E., Fligner, M. A., & Verducci, J. S. (1991). Probability models on rankings. *Journal of Mathematical Psychology, 35,* 294–318.

Diaconis, P. (1989). A generalization of spectral analysis with application to ranked data. *Annals of Statistics, 17,* 949–979.

Falmagne, J. (1985). *Elements of psychophysical theory.* New York: Oxford University Press.

Fligner, M. A., & Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society, Series B, 48,* 359–369.

Fligner, M. A., & Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical Association, 83,* 892–901.

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

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

Gelman, A., Meng, X. L., & Stern, H. S. (1993). *Bayesian model invalidation using tail area probabilities.* Unpublished manuscript. University of California, Department of Statistics, Berkeley.

Geman, D., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721–741.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, Series B, 29,* 83–100.

Hajivassiliou, V. A. (1993). Simulation estimation methods for limited dependent variable models. In G. S. Maddala, C. R. Rao, & H. D. Vinod (Eds.), *Handbook of statistics* (Vol. 11). Amsterdam: Elsevier.

Hausman, J. A. & Wise, D. A. (1978). A conditional probit model for qualitative choice: Discrete decision recognizing interdependence and heterogeneous preferences. *Econometrica, 46,* 403–426.

Heidelberger, O., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research, 31,* 1109–1144.

Henery, R. J. (1981). Permutation probabilities as models for horse races. *Journal of the Royal Statistical Society, Series B, 43,* 86–91.

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

Iverson, G. J. (1987). Thurstonian psychophysics: Case III. *Journal of Mathematical Psychology, 31,* 219–247.

Kamakura, W. A., & Srivastava, R. K. (1984). Predicting choice shares under conditions of brand interdependence. *Journal of Marketing Research, 21,* 420–434.

MacKay, D. B., & Chaiy, S. (1982). Parameter estimation for the Thurstone Case III model. *Psychometrika, 47,* 353–359.
Marden, J. I. (1995). *Analyzing and modeling rank data.* London: Chapman & Hall.
Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Erlbaum.
Meng, X. L. (1994). Posterior predictive *p*-values. *Annals of Statistics, 22,* 1142–1160.
Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12,* 1151–1172.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286.
Yao, G. (1995). *Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler.* Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.
Yao, G., & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology, 52,* 79–92.

Received 3 March 1999; revised version received 25 November 1999

Appendix: posterior distribution of $\boldsymbol{\mu}$ via Gibbs sampler

Consider n subjects randomly drawn to rank k items, where $\mathbf{Y}$ is an $(n \times k)$ ranking matrix with row vector $\mathbf{y}_j (j = 1, 2, \ldots, n)$ representing the rank outcome for participant j. The ranking outcome $\mathbf{y}_j$ is determined by the utilities $\mathbf{v}_j$ for participant j $(j = 1, 2, \ldots, n)$. For example, $\mathbf{v}_j = (0.35\ 0.24\ 0.31)$ results in the ranking outcome $\mathbf{y}_j = (1\ 3\ 2)$ for subject j. Under the Thurstonian Case V model, $\mathbf{v}_j$ is normally distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma} = \mathbf{I}$. A latent variable $\mathbf{V}$ is augmented so that we are able to sample $\boldsymbol{\mu}$ from $p(\boldsymbol{\mu}|\mathbf{V}, \mathbf{Y}, \mathbf{I})$ directly.

By using the Gibbs sampler, we can sample $(\mathbf{Y}, \boldsymbol{\mu})$ from their joint posterior distribution by iterating the following cycle of successive draws from the full conditional distributions:

1. Draw $\mathbf{v}_j$s from $p(\mathbf{v}_j|\boldsymbol{\mu}, \mathbf{y}_j, \mathbf{I})$ for $j = 1, 2, \ldots, n$.
2. Draw $\boldsymbol{\mu}$ from $p(\boldsymbol{\mu}|\mathbf{V}, \mathbf{Y}, \mathbf{I}) = p(\boldsymbol{\mu}|\mathbf{V}, \mathbf{I})$.

In step 1, to incorporate the information of $\mathbf{Y}$, a $(k-1) \times k$ contrast matrix $\mathbf{C}_j$ referring to the adjacent items in the ranking pattern is formed for each $\mathbf{y}_j, j = 1, 2, \ldots, n$. In the Thurstonian Case V model, $\mathbf{v} \sim N_k(\boldsymbol{\mu}, \mathbf{I})$, therefore $\mathbf{C}_j\mathbf{v} \sim N_{k-1}(\mathbf{C}_j\boldsymbol{\mu}, \mathbf{C}_j\mathbf{C}_j')$ follows a multivariate normal distribution. For each $\mathbf{y}_j$, one draws $\mathbf{x}_j$ from the conditional posterior of $\mathbf{C}_j\mathbf{v}|\boldsymbol{\mu}, \mathbf{y}_j$ which is a $(k-1)$-variate normal distribution truncated at $\mathbf{O}$ with mean $\mathbf{C}_j\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}_j\mathbf{C}_j'$. The rejection procedure can then be used to obtain random draws, $\mathbf{x}_j$, from the truncated multivariate normal distribution. $\mathbf{v}_j$ can then be obtained by $\mathbf{v}_j = \mathbf{C}_j^{-}\mathbf{x}_j$, where $\mathbf{C}_j^{-}$ is the generalized inverse of $\mathbf{C}_j$. But as k grows large, this method becomes inefficient since a large number of draws are rejected before one which satisfies the condition. Fortunately, a more efficient process was introduced by Hajivassiliou (1993) to generate samples from a truncated multivariate normal distribution (TMVN) by adopting the Gibbs sampler algorithm. Instead of drawing directly from a TMVN, the TMVN Gibbs sampler cycle obtains draws through the successive $(k-1)$ fully conditional truncated univariate normal distributions. The program used for this study will either find the convergent value $\mathbf{v}^{(t)}$ or use $t = 20$ replications as default. This procedure is most useful when the number of items is large.

In step 2, since

$$\begin{aligned} p(\boldsymbol{\mu}|\mathbf{V},\mathbf{Y},\mathbf{I}) &= p(\boldsymbol{\mu}|\mathbf{V},\mathbf{I}) \propto p(\boldsymbol{\mu})p(\mathbf{V}|\boldsymbol{\mu},\mathbf{I}) \\ &\propto \exp(-\tfrac{1}{2}\boldsymbol{\mu}'(S^2\mathbf{I})^{-1}\boldsymbol{\mu}) \exp\left(-\tfrac{1}{2}\sum_{j=1}^{n}(\mathbf{v}_j-\boldsymbol{\mu})'\mathbf{I}^{-1}(\mathbf{v}_j-\boldsymbol{\mu})\right) \\ &\propto \exp(-\tfrac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)'\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)) \\ &\sim N_k(\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n), \end{aligned}$$

where $p(\boldsymbol{\mu}) \sim N_k(\mathbf{0}, S^2\mathbf{I})$, $\boldsymbol{\mu}_n = (nS^2/(1+ns^2))\bar{\mathbf{Y}}$ and $\boldsymbol{\Sigma}_n = (nS^2/(1+nS^2))\mathbf{I}$. Thus, we draw $\boldsymbol{\mu}$ directly from $p(\boldsymbol{\mu}|\mathbf{V},\mathbf{Y},\mathbf{I}) \sim N_k(\boldsymbol{\mu}_n,\boldsymbol{\Sigma}_n)$.

Under weak conditions, Geman & Geman (1984) showed that as the number of iterations $t \to \infty$, the joint density of $(\mathbf{V}^{(t)}, \boldsymbol{\mu}^{(t)})$ will geometrically converge in distribution to the joint density of $(\mathbf{V}, \boldsymbol{\mu})$. Hence, after iterative drawings from the above two conditional posterior distributions for a sufficient number of times, the sequence $\{(\mathbf{V}^{(1)}, \boldsymbol{\mu}^{(1)}), (\mathbf{V}^{(2)}, \boldsymbol{\mu}^{(2)}), \ldots\}$ will converge to a single draw $(\mathbf{V}, \boldsymbol{\mu})$ from the joint posterior distribution. There are no firmly established rules to determine the number of iterations required to achieve convergence. Nevertheless, the convergence of the sequences is ensured by Heidelberger & Welch (1983) stationarity tests. In the present study, 1500 iterations were used in the simulation to reach convergence. We discarded the first 500 so-called burn-in draws and collected every 10th of the remaining 1000 draws to minimize autocorrelations between successive draws. The 100 collected draws were used to construct the posterior distribution of $\boldsymbol{\mu}$, namely $p(\boldsymbol{\mu}|\mathbf{Y},\mathbf{I})$.