Research Report

Model Diagnostics for Bayesian Networks

Sandip Sinharay

Research & Development

April 2004
RR-04-17

Model Diagnostics for Bayesian Networks

Sandip Sinharay

ETS, Princeton, NJ

April 2004

Abstract

Assessing fit of psychometric models has always been an issue of enormous interest, but there exists no unanimously agreed upon item fit diagnostic for the models. Bayesian networks, frequently used in educational assessments (see, for example, Mislevy, Almond, Yan, & Steinberg, 2001) primarily for learning about students' knowledge and skills, are no exception. This paper employs the *posterior predictive model checking* method (Guttman, 1967; Rubin, 1984), a popular Bayesian model checking tool, to assess fit of simple Bayesian networks. A number of aspects of model fit, those of usual interest to practitioners, are assessed in this paper using various diagnostic tools. The first diagnostic used is direct data display—a visual comparison of the observed data set and a number of the posterior predictive data sets (that are predicted by the model). The second aspect examined here is item fit. Examinees are grouped into a number of equivalence classes, based on the generated values of their skill variables, and the observed and expected proportion correct scores on an item for the classes are combined to provide a χ^2-type and a G^2-type test statistic for each item. Another (similar) set of χ^2-type and G^2-type test statistic is obtained by grouping the examinees by their raw scores and then comparing their observed and expected proportion correct scores on an item. This paper also suggests how to obtain posterior predictive p-values, natural candidate p-values from a Bayesian viewpoint, for the χ^2-type and G^2-type test statistics. The paper further examines the association among the items, especially if the model can explain the odds ratios corresponding to the responses to pairs of items. Finally, in an effort to examine the issue of differential item functioning (DIF), this paper suggests a version of the Mantel-Haenszel statistic (Holland, 1985), which uses "matched groups" based on equivalence classes, as a discrepancy measure with posterior predictive model checking. Limited simulation studies and a real data application examine the effectiveness of the suggested model diagnostics.

Key words: Discrepancy measure, Mantel-Haenszel statistic, p-values

Acknowledgements

A number of ideas in this paper originated as a result of the author's discussions with Shelby Haberman. The author thanks Russell Almond, Robert Mislevy, and Hal Stern for their invaluable advice. The author gratefully acknowledges the help of Rochelle Stern and Kim Fryer with proofreading.

Table of Contents

1 Introduction

Bayesian inference networks (BIN/BN; Pearl, 1988) are frequently used in educational assessments (see, for example, Mislevy, Almond, Yan, & Steinberg, 2001 and the references there), especially in the context of evidence-centered design (ECD; Almond & Mislevy, 1999), for learning about students' knowledge and skills, given information about their performance in an assessment.

In a typical application of a BN in the context of psychometrics, task analysis of the items suggest that solving each item in an assessment requires at least a certain level of one or more of a number of prespecified skills. One then assumes that each examinee has a discrete proficiency variable for each skill, stating the skill-level of the examinee. Finally, a BN models the response of an examinee to an item by considering whether the examinee possesses the levels of different skills required to solve the item. For example, a simple BN might assume that the skill variables are binary (i.e., one either has the skill or doesn't), that the examinees who have all the required skills for item j have a certain probability π_{1j} of solving the item, while those who do not have the required skills have probability $\pi_{0j} < \pi_{1j}$ of solving the item. Finally, a BN specifies prior distributions over the skill variables, usually with the help of a graphical model (Pearl, 1988), and the item parameters.

Model checking for BNs is not a straightforward task. The possible number of response patterns is huge for even moderately large number of items. The presence of latent and multivariate proficiency variables makes it even worse; the traditional model checking techniques do not apply here, except for the trivial case of a test with few items. As an outcome, the use of statistical diagnostic tools in this context is notably lacking (Williamson et al., 2000).

Recently, Yan, Mislevy, and Almond (2003) and Sinharay, Almond, and Yan (2004) suggested a number of model checking tools for BNs, including item fit plots and an item-fit test statistic. However, the null distribution of the item fit test statistic suggested in the latter paper is not well-established. Further, there is little work (in the context of BNs) regarding checking if a model can predict the association among the items successfully. Another area of interest is differential item functioning (DIF), which is of major concern to

any test administrators. However, there is no published example of any work regarding DIF analysis for BNs. In summary, there is a significant scope of further research in all these areas.

The posterior predictive model checking (PPMC) method (Guttman, 1967; Rubin, 1984) is a popular Bayesian model checking tool because of its simplicity, strong theoretical basis, and obvious intuitive appeal. The method primarily consists of comparing the observed data with *replicated data*–those predicted by the model–using a number of *discrepancy measures*, including a discrepancy measure like a classical test statistic that measures the difference between an aspect of the observed data set and a replicated data set. Practically, a number of replicated data sets are generated from the predictive distribution of replicated data conditional on the observed data (the *posterior predictive distribution*). Any systematic differences between the discrepancy measures for the observed data set and those for the replicated data sets indicate potential failure of the model to explain the data. Graphical display is the most natural and easily comprehensible way to examine the difference. Another powerful tool is the *posterior predictive p-value*, the Bayesian counterpart of the classical p-value. Sinharay and Johnson (2003) apply the technique to assess the fit of common dichotomous item response theory (IRT) models (the 1-, 2-, and 3PL models).

First, this paper suggests a direct data display as a quick overall check of model fit. All the responses in the observed data set are plotted in a graph, along with those in a number of posterior predictive data sets. The display, although very simple, may reveal interesting facts regarding model fit.

This article then uses the PPMC method to examine item fit for BNs. First, the posterior predictive p-values corresponding to the point biserial correlations are examined to check if the model can predict those quantities adequately. As a next step, this paper examines two sets of item fit measures. As in Sinharay et al. (2004), the equivalence class membership (which is determined by the combination of skills) of the examinees is the basis of forming the examinee groups in one set of item fit measures. Even though the memberships are unknown (latent) quantities, the Markov chain Monte Carlo (MCMC) algorithm used for fitting the model produces values of the latent abilities in each

iteration, which determine the equivalence class memberships and allow the grouping. The observed and expected values of proportion correct of the examinees are computed for each combination of item and equivalence class. This paper then suggests using two discrepancy measures, a χ^2-type measure and a G^2-type measure, to compare the observed and expected values for each item. The posterior predictive p-values (PPP-value) corresponding to the discrepancy measures quantify the information regarding fit for each item. Even though grouping of the examinees is based on the unknown equivalence class membership, the PPMC method takes care of the unknown nature of them by properly integrating them out to provide a natural Bayesian p-value. Examinee groups with respect to raw scores form the basis of another set of χ^2-type measure and G^2-type measure.

Sinharay and Johnson (2003) apply the PPMC method to find that the odds ratios corresponding to the responses to pairs of items to be powerful discrepancy measures in detecting misfit of the simple IRT models. This paper uses the PPP-values for the odds ratios to check whether the BNs can explain the association among the items adequately.

For assessing DIF, this paper uses the Mantel-Haenszel test suggested by Holland (1985). As developed by Holland (1985), the test uses "matching groups" based on raw scores of examinees; however, the same type of groups may not be homogeneous in the context of BNs. Therefore, this paper applies the PPMC method using a discrepancy based on a version of the Mantel-Haenszel test that uses the equivalence class memberships (rather than raw scores) to form the "matching groups."

The Bayesian networks deal with latent classes, and hence there is a possibility of nonidentifiability or weak identifiability of some parameters. Weak identifiability may lead to nonapplicability of standard asymptotic theory, problems with MCMC convergence, and worse, misleading inferences. Therefore, a diagnostic for model identifiability, although not directly a model checking tool, is a necessary tool for assessing the performance of a Bayesian network. This paper uses plots of prior vs. posterior distributions of the model parameters to assess their identifiability.

Section 2 introduces the Bayesian network models. Section 3 introduces the mixed-number subtraction example (Tatsuoka, 1990) that will form the basis of a number of analyses later. Section 4 provides a brief discussion of the PPMC method and the

posterior predictive p-value. Section 5 introduces the discrepancy measures to be used for assessing different aspects of model fit. Section 6 applies the direct data display to the mixed-number data and a simulated data. Section 7 applies the suggested item fit measures to the mixed-number subtraction data set and a simulated data set. Section 8 discusses results of examining association among item pairs. Section 9 discusses results of the DIF analysis. Section 10 assessess identifiability of the parameters. The conclusions of this study are discussed in Section 11.

2 Bayesian Networks

The idea of a Bayesian inference network (also known as a Bayesian network or Bayes net; Pearl, 1988) comes primarily from the theory of graphical models, where it is defined as an *acyclic-directed graph* in which each node represents a random variable, or uncertain quantity, which can take two or more possible values. To a statistician, BNs are statistical models that describe a number of observations from a process with the help of a number of latent (or unobserved) categorical variables. In the context of educational assessments, BNs are appropriate when the task analysis of an assessment suggests that the satisfactorily completion of any task requires an examinee to have at least a certain level of a number of skills. The unobserved (or latent) variables, representing the level of the skills in the examinees, affect the probability distribution of the observed responses in a BN. Note that despite the name, BNs do not necessarily imply a commitment to Bayesian methods; they are called so because they use Bayes' rule for inference. Almond and Mislevy (1999) argue how BNs may be useful in psychometrics.

Consider an assessment with I examinees and J items. Suppose X_{ij} denotes the response of the i-th examinee to the j-th item. In general, X_{ij} can be a vector-valued quantity; that is, a single "task" could consist of multiple "items." It can contain polytomous outcomes as well. In our example, each item produces a single dichotomous outcome, which is 1 if the response is correct and 0 if it is incorrect. Denote the ability/skill variable vector for the i-th examinee to be $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots \theta_{ip})'$, θ_{ik} denoting the skill level of Examinee i for Skill k, and the item parameter vector for the j-th item to be $\boldsymbol{\pi}_j = (\pi_{j1}, \pi_{j2}, \ldots \pi_{jm})'$. A standard assumption is that of local independence, that is,

independence of the responses conditional on the skills.

The next step is to assume that $P(X_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\pi}_j) = f(\boldsymbol{\theta}_i, \boldsymbol{\pi}_j)$ for a suitable function $f(\boldsymbol{\theta}_i, \boldsymbol{\pi}_j)$.

As an example of a simple BN, assume that solving the j-th item needs a number of skills, define $\boldsymbol{\theta}_i$ to consist of the indicators of attainment of the skills, let $\boldsymbol{\pi}_j = (\pi_{j0}, \pi_{j1})'$, and choose

$$f(\boldsymbol{\theta}_i, \boldsymbol{\pi}_j) = \begin{cases} \pi_{j1} \text{ if the examinee } i \text{ has the skills required to solve the item } j, \\ \pi_{j0} \text{ otherwise}. \end{cases}$$

In order to perform a Bayesian analysis of a BN, one needs a prior (population) distribution $p(\boldsymbol{\theta}|\boldsymbol{\lambda})$ on the ability variables $\boldsymbol{\theta}$. Mislevy (1995) shows that forming a graphical model (Pearl, 1988; Lauritzen & Spiegelhalter, 1988) over the skill variables provides a convenient way to specify this distribution.

3 The Mixed-number Subtraction Example

Increasingly, users of educational assessments want more than a single summary statistic out of an assessment. They would like to see a profile of the state of acquisition of a variety of knowledge, skills and proficiencies for each learner. One technique for *profile scoring* is the use of a Bayesian network. The analysis starts with a cognitive analysis of a number of tasks in a domain to determine the "attributes," which are important for solving different kinds of problems. The experts then produce a Q-matrix, an incidence matrix showing for each item in an assessment in which attributes are required to solve that item. To illustrate the method, we introduce what will be a running example used through the paper, one regarding a test of mixed-number subtraction.

This example is grounded in a cognitive analysis of middle school students' solutions of mixed-number subtraction problems. The focus here is on the 325 students learning to use the following method (Klein, Birnbaum, Standiford, & Tatsuoka, 1981):

Method B: Separate mixed numbers into whole number and fractional parts; subtract as two subproblems, borrowing one from minuend whole number if necessary; and then simplify and reduce if necessary.

The cognitive analysis mapped out a flowchart for applying Method B to a universe of fraction subtraction problems. A number of key procedures appear, a subset of which is required to solve a given problem according to its structure. To simplify the model, we eliminate the items for which the fractions do not have a common denominator (leaving us with 15 items). The procedures are as follows:

- Skill 1: Basic fraction subtraction.
- Skill 2: Simplify/reduce fraction or mixed number.
- Skill 3: Separate whole number from fraction.
- Skill 4: Borrow one from the whole number in a given mixed number.
- Skill 5: Convert a whole number to a fraction.

Furthermore, the cognitive analysis identified Skill 3 as a prerequisite of Skill 4, that is, no students have Skill 4 without also having Skill 3. Thus, there are only 24 possible combinations of the five skills that a given student can possess.

Table 1 lists 15 items from a data set collected by Tatsuoka (1990) that will be studied here, characterized by the skills they require. The column marked "Skills Required" represents the Q-matrix. The table also shows the proportion-correct scores for each item.

Sinharay et al. (2004) discuss that Items 4 and 5 may not require any mixed-number subtraction skills; for example, an examinee may solve Item 4 just by noticing that any number minus the same number results in zero.

A number of features of this data set can be identified by studying Table 1. First, note that many rows of the Q-matrix are identical, corresponding to a group of items that require the same set of skills to solve. Following the terminology of evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2003) we call the patterns corresponding to the rows *evidence models*.

Second, note that certain patterns of skills will be indistinguishable on the basis of the results of this test. For example, because every item requires Skill 1, the 12 profiles (i.e., skill patterns) that lack Skill 1 are indistinguishable on the basis of this data. Similar logic

Table 1.

Skill Requirements for the Mixed-number Subtraction Problems

Item	Text of	Skills required					Evidence	Proportion
no.	the item	1	2	3	4	5	model	correct
2	$\frac{6}{7}$ - $\frac{4}{7}$	x					1	0.79
4	$\frac{3}{4}$ - $\frac{3}{4}$	x					1	0.70
8	$\frac{11}{8}$ - $\frac{1}{8}$	x	x				2	0.71
9	$3\frac{4}{5}$ - $3\frac{2}{5}$	x		x			3	0.75
11	$4\frac{5}{7}$ - $1\frac{4}{7}$	x		x			3	0.74
5	$3\frac{7}{8}$ - 2	x		x			3	0.69
1	$3\frac{1}{2}$ - $2\frac{3}{2}$	x		x	x		4	0.37
7	$4\frac{1}{3}$ - $2\frac{4}{3}$	x		x	x		4	0.37
12	$7\frac{3}{5}$ - $\frac{4}{5}$	x		x	x		4	0.34
15	$4\frac{1}{3}$ - $1\frac{5}{3}$	x		x	x		4	0.31
13	$4\frac{1}{10}$ - $2\frac{8}{10}$	x		x	x		4	0.41
10	2 - $\frac{1}{3}$	x		x	x	x	5	0.38
3	3 - $2\frac{1}{5}$	x		x	x	x	5	0.33
14	7 - $1\frac{4}{3}$	x		x	x	x	5	0.26
6	$4\frac{4}{12}$ - $2\frac{7}{12}$	x	x	x	x		6	0.31

reveals that there are only nine identifiable *equivalence classes* of student profiles. This property of the test design will manifest itself later in the analysis. Table 2 describes the classes by relating them to the evidence models.

Often, distinctions among members of the same equivalence class are instructionally irrelevant. For example, students judged to be in Equivalence Class 1 would all be assigned remedial work in basic subtraction, so no further distinction is necessary.

Mislevy (1995) analyzed the data using Bayesian networks and we will use that model here. Let $\boldsymbol{\theta}_i = \{\theta_{i1}, \ldots, \theta_{i5}\}$ denote the attribute/skill vector for Examinee i. In the context of ECD, where the BNs have found most use, θ_{ik}s are called *proficiency variables*. These

Table 2.

Description of Each Equivalence Class With Evidence Models Items That Should Be Solved by Examinees of an Equivalence Class

Equivalence	Class	EM					
class	description	1	2	3	4	5	6
1	No Skill 1						
2	Only Skill 1	x					
3	Skills 1 & 3	x		x			
4	Skills 1, 3, & 4	x		x	x		
5	Skills 1, 3, 4, & 5	x		x	x	x	
6	Skills 1 & 2	x	x				
7	Skills 1, 2, & 3	x	x	x			
8	Skills 1, 2, 3, & 4	x	x	x	x		x
9	All Skills	x	x	x	x	x	x

are latent variables describing knowledges, skills, and abilities of the examinees we wish to draw inferences about. (Note that these are sometimes referred to as *person parameters* in the IRT literature. However, we use the term *variable* to emphasize its person specific nature.) The distribution of these variables, $P(\boldsymbol{\theta}_i)$, is known as the *proficiency model.* In the mixed-number subtraction example, the proficiency model consists of the distribution of five binary variables related to the presence or absence of the five skills.

3.1 Mixed-number Subtraction Link Models

The "link models" in ECD terminology state the likelihood distribution of the data given the model parameters and latent variables. The model implicit in Table 1 is a conjunctive skills model; that is, a participant needs to have mastered all of the skills shown in the appropriate row in order to solve the problem. If the participant has mastered all of the skills necessary to solve a particular item, one says that the student has mastered the item. In general, students will not behave according to the ideal model; we will get false

positive and false negative results.

The link models described below follow that intuition. The 2LC model uses two parameters per link model, the true positive and false positive probabilities, in

$$P(X_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\pi}_j) = \begin{cases} \pi_{j1} \text{ if examinee } i \text{ mastered all the skills needed to solve item } j \\ \pi_{j0} \text{ otherwise.} \end{cases} \quad (1)$$

Suppose the j-th item uses the Evidence Model $s, s = 1, 2, \ldots 6$. Although s is determined by the item, this notation does not reflect that. Let $\delta_{i(s)}$ be the 0/1 indicator denoting whether the examinee i has mastered the skills needed for tasks using Evidence Model s or not. Note that the $\delta_{i(s)}$s for any examinee are completely determined by the values of $\theta_1, \theta_2, \ldots \theta_5$ for that examinee. The likelihood of the response of the i-th examinee to the j-th item is then expressed as

$$X_{ij}|\pi_{j\delta_{i(s)}}, \boldsymbol{\theta}_i \sim \text{Bernoulli}(\pi_{j\delta_{i(s)}}). \quad (2)$$

The model relies on the assumption of local independence, that is, given the proficiency $\boldsymbol{\theta}_i$, the responses of an examinee to the different items are assumed independent. The probability π_{j1} represents a "true-positive" probability for the item, this is, it is the probability of getting the item right for students who have mastered all of the required skills. The probability π_{j0} represents a "false-positive" probability; it is the probability of getting the item right for students who have yet to master at least one of the required skills. The probabilities π_{j0} and π_{j1} are allowed to differ over j (i.e., from item to item). However, we use the same independent priors for all items:

$$\begin{aligned} \pi_{j0} &\sim Beta(3.5, 23.5), \quad (3) \\ \pi_{j1} &\sim Beta(23.5, 3.5)\cdot \end{aligned}$$

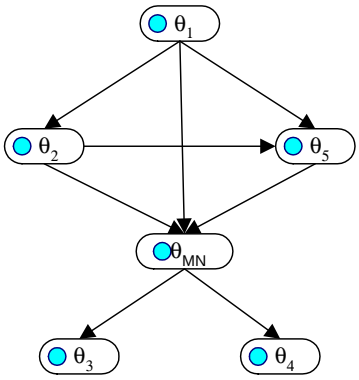

***Figure* 1. The graphical representation of the student model for 2LC model.**

This model is similar to the *deterministic inputs, noisy "and"* gate (DINA) model of Junker and Sijtsma (2001), *noisy inputs deterministic and* gate (NIDA) model of Maris (1999), the higher order latent trait models (de la Torre & Douglas, in press), and the fusion model of Hartz, Roussos, and Stout (2002).

3.2 Mixed-number Subtraction Proficiency Model

In the ECD framework, the proficiency model states the population distribution on the proficiency variables and the prior distributions on the model parameters. Here, the prior (population) distribution $P(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is expressed as a discrete Bayesian network or graphical model (Pearl, 1988; Lauritzen & Spiegelhalter, 1988).

Prior analyses revealed that Skill 3 is a prerequisite to Skill 4. A three-level auxiliary variable θ_{WN} incorporates this constraint. Level 0 of θ_{WN} corresponds to the participants who have mastered neither skill; Level 1 represents participants who have mastered Skill 3 but not Skill 4; Level 2 represents participants who mastered both skills. Figure 1 shows the dependence relationships among the skill variables provided by the expert analysis (primarily correlations, but Skill 1 is usually acquired before any of the others so all of the remaining skills are given conditional distributions given Skill 1). It corresponds to the factorization

$$p(\boldsymbol{\theta}|\boldsymbol{\lambda}) = p(\theta_3|\theta_{WN},\boldsymbol{\lambda})p(\theta_4|\theta_{WN},\boldsymbol{\lambda})p(\theta_{WN}|\theta_1,\theta_2,\theta_5,\boldsymbol{\lambda})p(\theta_5|\theta_1,\theta_2,\boldsymbol{\lambda})p(\theta_2|\theta_1,\boldsymbol{\lambda})p(\theta_1,\boldsymbol{\lambda}) \ .$$

The parameters $\boldsymbol{\lambda}$ are defined as follows:

$$
\begin{aligned}
\lambda_1 &= P(\theta_1 = 1) \ . \\
\lambda_{2,m} &= P(\theta_2 = 1 | \theta_1 = m) \qquad \text{for } m = 0, 1 \ . \\
\lambda_{5,m} &= P(\theta_5 = 1 | \theta_1 + \theta_2 = m) \qquad \text{for } m = 0, 1, 2 \ . \\
\lambda_{WN,m,n} &= P(\theta_{WN} = n | \theta_1 + \theta_2 + \theta_5 = m) \qquad \text{for } m = 0, 1, 2, 3 \text{ and } n = 0, 1, 2 \ .
\end{aligned}
$$

Finally, one requires prior distributions $P(\boldsymbol{\lambda})$. We assume that λ_1, $\boldsymbol{\lambda}_2$, $\boldsymbol{\lambda}_5$, and $\boldsymbol{\lambda}_{WN}$ are *a priori* independent.

The natural conjugate priors for the components of $\boldsymbol{\lambda}$ are either Beta or Dirichlet distributions. The hyper-parameters are chosen, as in Mislevy et al. (2001), so that they sum to 27 (relatively strong numbers given the sample size of 325). With such a complex latent structure, strong priors such as the ones here are necessary to prevent problems with identifiability. These must be supported by relatively expensive elicitation from the experts. Prior probabilities give a chance of 87% of acquiring a skill when the previous skills are mastered and 13% of acquiring the same skill when the previous skills are not mastered. They are given below:

$$\lambda_1 \sim \text{Beta}(23.5, 3.5)$$

$$\lambda_{2,0} \sim \text{Beta}(3.5, 23.5); \quad \lambda_{2,1} \sim \text{Beta}(23.5, 3.5)$$

$$\lambda_{5,0} \sim \text{Beta}(3.5, 23.5); \quad \lambda_{5,1} \sim \text{Beta}(13.5, 13.5); \quad \lambda_{5,2} \sim \text{Beta}(23.5, 3.5)$$

$$\boldsymbol{\lambda}_{WN,0,\cdot} = (\boldsymbol{\lambda}_{WN,0,0}, \boldsymbol{\lambda}_{WN,0,1}, \boldsymbol{\lambda}_{WN,0,2}) \sim \text{Dirichlet}(15, 7, 5)$$

$$\boldsymbol{\lambda}_{WN,1,\cdot} \sim \text{Dirichlet}(11, 9, 7); \quad \boldsymbol{\lambda}_{WN,2,\cdot} \sim \text{Dirichlet}(7, 9, 11); \quad \boldsymbol{\lambda}_{WN,3,\cdot} \sim \text{Dirichlet}(5, 7, 15)\cdot$$

Haertel and Wiley (1993) note that whenever the proficiency model consists of binary skills, it implicitly induces a number of latent classes. In this example, there are 24 values of $\boldsymbol{\theta}$, which have nonzero prior probability. The graphical model $p(\boldsymbol{\theta}|\boldsymbol{\lambda})$, described above, is a compact and structured way of representing the prior probability over those latent

classes. Although this distribution is overall 24 possible latent classes, only nine of them are identifiable from the data (Table 2).

3.3 Fitting the 2LC Model to the Mixed-number Subtraction Data

As in Mislevy et al. (2001), the BUGS software (Spiegelhalter, Thomas, Best, & Gilks, 1995) is used to fit an MCMC algorithm to the data set. For details of the model fitting, see, for example, Sinharay et al. (2004). Five chains of size 2,000 each are used, after a burn-in of 1,000 each, resulting in a final posterior sample size of 10,000 discussed in Section 3.3 (hence there are 10,000 replicated data sets, which provide an estimate of a PPP-value).

4 Posterior Predictive Model Checking Techniques

Guttman (1967) suggested the idea behind the PPMC method. Let $p(\boldsymbol{y}|\boldsymbol{\omega})$ denote the likelihood distribution for a statistical model applied to data (examinee responses in this context) $\boldsymbol{y}$, where $\boldsymbol{\omega}$ denotes all the parameters in the model. Let $p(\boldsymbol{\omega})$ be the prior distribution on the parameters. Then the posterior distribution of $\boldsymbol{\omega}$ is $p(\boldsymbol{\omega}|\boldsymbol{y}) \equiv \frac{p(\boldsymbol{y}|\boldsymbol{\omega})p(\boldsymbol{\omega})}{\int_{\boldsymbol{\omega}} p(\boldsymbol{y}|\boldsymbol{\omega})p(\boldsymbol{\omega})d\boldsymbol{\omega}}$. Let $\boldsymbol{y}^{rep}$ denote replicate data that one might observe if the process that generated the data $\boldsymbol{y}$ is replicated with the same value of $\boldsymbol{\omega}$ that generated the observed data.

The PPMC method suggests checking a model using the *posterior predictive distribution* (or the predictive distribution of replicated data conditional on the observed data),

$$p(\boldsymbol{y}^{rep}|\boldsymbol{y}) = \int p(\boldsymbol{y}^{rep}|\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{y})d\boldsymbol{\omega}, \tag{4}$$

as a reference distribution for the observed data $\boldsymbol{y}$.

The next step in the PPMC method is to compare the observed data $\boldsymbol{y}$ to its reference distribution (4). In practice, *test quantities* or *discrepancy measures* $D(\boldsymbol{y}, \boldsymbol{\omega})$ are defined (Gelman, Meng, & Stern, 1996), and the posterior distribution of $D(\boldsymbol{y}, \boldsymbol{\omega})$ is compared to the posterior predictive distribution of $D(\boldsymbol{y}^{rep}, \boldsymbol{\omega})$, with any significant difference between them indicating a model failure. One may use $D(\boldsymbol{y}, \boldsymbol{\omega}) = D(\boldsymbol{y})$, a discrepancy measure depending on the data only, if appropriate.

A popular summary of the comparison is the tail-area probability or *posterior predictive p-value* (PPP-value), or *Bayesian p-value*:

$$\begin{aligned} p_b &= P\left(D(\boldsymbol{y}^{rep}, \boldsymbol{\omega}) \geq D(\boldsymbol{y}, \boldsymbol{\omega}) | \boldsymbol{y}\right) \\ &= \int\int I_{[D(\boldsymbol{y}^{rep}, \boldsymbol{\omega}) \geq D(\boldsymbol{y}, \boldsymbol{\omega})]} p(\boldsymbol{y}^{rep}|\boldsymbol{\omega}) p(\boldsymbol{\omega}|\boldsymbol{y}) d\boldsymbol{y}^{rep} d\boldsymbol{\omega}, \end{aligned} \tag{5}$$

where $I_{[A]}$ denotes the indicator function for the event A.

Because of the difficulty in dealing with (4) or (5) analytically for all but simple problems, Rubin (1984) suggests simulating replicate data sets from the posterior predictive distribution in practical applications of the PPMC method. One draws N simulations $\boldsymbol{\omega}^1$, $\boldsymbol{\omega}^2$,...,$\boldsymbol{\omega}^N$ from the posterior distribution $p(\boldsymbol{\omega}|\boldsymbol{y})$ of $\boldsymbol{\omega}$, and draws $\boldsymbol{y}^{rep,n}$ from the likelihood distribution $p(\boldsymbol{y}|\boldsymbol{\omega}^n)$, $n = 1, 2, \ldots N$. The process results in N draws from the joint posterior distribution $p(\boldsymbol{y}^{rep}, \boldsymbol{\omega}|\boldsymbol{y})$, and, equivalently, from $p(\boldsymbol{y}^{rep}|\boldsymbol{y})$. The expression (5) suggests that the posterior predictive p-value is also the expectation of $I_{[D(\boldsymbol{y}^{rep}, \boldsymbol{\omega}) \geq D(\boldsymbol{y}, \boldsymbol{\omega})]}$, where the expectation is with respect to the joint posterior distribution $p(\boldsymbol{y}^{rep}, \boldsymbol{\omega}|\boldsymbol{y})$. As an immediate consequence, the proportion of the N replications for which $D(\boldsymbol{y}^{rep,n}, \boldsymbol{\omega}^n)$ exceeds $D(\boldsymbol{y}, \boldsymbol{\omega}^n)$ provides an estimate of the PPP-value. Extreme posterior predictive p-values (close to 0, or 1, or both, depending on the nature of the discrepancy measure) indicate model misfit.

Gelman, Meng, and Stern (1996) suggest that the preferable way to perform PPMC is to compare the realized discrepancies $D(\boldsymbol{y}, \boldsymbol{\omega}^n)$ and the replicated/predicted discrepancies $D(\boldsymbol{y}^{rep,n}, \boldsymbol{\omega}^n)$ and by plotting the pairs $\{D(\boldsymbol{y}, \boldsymbol{\omega}^n), D(\boldsymbol{y}^{rep,n}, \boldsymbol{\omega}^n)\}, n = 1, 2, \ldots N$, in a scatter-plot.

Empirical and theoretical studies so far suggest that PPP-values generally have reasonable long-run frequentist properties (Gelman et al., 1996). The PPMC method combines well with the MCMC algorithms (Gelman, Carlin, Stern, & Rubin, 1995).

Robins, van der Vaart, and Ventura (2000) show that the PPP-values are conservative (i.e., often fails to detect model misfit), even asymptotically, for some choices of discrepancy measure, for example, when the discrepancy measure is not centered. However, a conservative statistical test with reasonable power is better than tests with unknown

properties or poor Type I error rates, more so for item fit, because items are costly and it may be to the advantage of the administrators to fail to reject the hypothesis for a borderline item and thus retain the item for future use. Gelman et al. (1996) comment that the PPMC method is useful if one thinks of the current model as a plausible ending point with modifications to be made only if substantial lack of fit is found and item fit is an application exactly as referred to by them; test administrators would like to discard an item only if the model fails substantially for the item.

5 The Suggested Model Diagnostics

5.1 Direct Data Display

Suitable display showing the observed data and a few replicated data sets may be a powerful tool to provide a rough idea about model fit (e.g., Gelman et al., 2003). For small data sets like the mixed-number subtraction data, it is possible to plot the whole data. For large data sets, however, plotting all data points is prohibitive—but plots for a sample (for example, few examinees each from low-scoring, high-scoring, and middle groups) may reveal interesting patterns of differences between the observed and replicated data.

5.2 Item Fit Analysis

This paper first examines if the fitted model can predict a simple and natural quantity, the point biserial correlations of the test items. The tool used is the posterior predictive p-values corresponding to the point biserial correlations. Sinharay and Johnson (2003) apply a similar approach to the simple IRT models.

5.2.1 Item Fit Measures Based on Equivalence Class Membership

Here, the paper examines item fit measures based on equivalence class memberships of the examinees. We use ideas of Yan et al. (2003) to group the examinees according to the equivalence classes because the equivalence class membership of an examinee, although latent, is a natural quantity in this context. The natural groups are then the nine equivalence classes defined in Table 2. Equivalence Class 1 represents no skills and Equivalence Class 9 represents all skills. The classes in between are roughly ordered in

order of increasing skills; however, the ordering is only partial.

The first discrepancy measure examined is the proportion of examinees $\tilde{O}_{jk}$ in equivalence class $k, k = 1, 2, \ldots K = 9$ who answer item j correctly, $j = 1, 2, \ldots J$. The quantity $\tilde{O}_{jk}$ is not a truly observed quantity (reflected in the notation), but depends on the parameter values of the model (that is why they are referred to as discrepancy measures here, as in Gelman et al., 1996). They become known if one knows the equivalence class membership of the examinees. For a replicated data set, there is one such proportion, denoted $\tilde{O}^{rep}_{jk}$. For each combination of item j and equivalence class k, comparison of the values of $\tilde{O}_{jk}$ and $\tilde{O}^{rep}_{jk}$ over the iterations of the MCMC provides an idea regarding the fit of the item. This paper computes the posterior predictive p-value (PPP-value) for each combination of item and equivalence class. Even though $\tilde{O}_{jk}$s depend on the latent equivalence class membership and are unknown, the PPMC method integrates (sums in this context) out the unknown quantities with respect to their posterior distribution to provide natural Bayesian p-values. These p-values provide useful information about the fit of the items, and may suggest possible improvements of the model. For example, significant p-values for most items for Equivalence Class 1 (examinees with no Skill 1) will mean that the model is not flexible enough to explain examinees in that class and that one should add a component to the model to address that issue (this happens in the mixed-number example later).

The above mentioned item fit p-values provide useful feedback about item fit, but it will be useful to summarize the fit information for each item into one plot or one number, preferably a p-value. To achieve that, this paper uses the two test statistics (χ^2-type and G^2-type), somewhat similar to those in Sinharay (2003) as discrepancy measures.

Consider for this discussion that the values of the $\boldsymbol{\theta}_i$s and $\boldsymbol{\pi}_j$s are known, which means that the equivalence class memberships of all examinees are known (which is true for each iteration of the MCMC algorithm). For each item, one can then compute $\tilde{O}_{jk}, k = 1, 2, \ldots K$ and $\tilde{O}^{rep}_{jk}$s, $k, k = 1, 2, \ldots K$, which are described above. Let $N_k, k = 1, 2, \ldots K$ denote the number of examinees in Equivalence Class k. Given the $\boldsymbol{\theta}_i$s and $\boldsymbol{\pi}_j$s, the expected probability of an examinee in Equivalence Class k answering Item j correctly, E_{jk}, is the suitable $\pi_{j\delta i(s)}$. For example, for examinees in Equivalence Class 1 (one without

skill 1), $E_{j1} = \pi_{j0} \forall j$, because these examinees do not have the necessary skills to solve any item. On the other hand, for examinees in Equivalence Class 2 (with Skill 1 only), $E_{j2} = \pi_{j1} I_{[j \in \{2,4\}]} + \pi_{j0} I_{[j \notin \{2,4\}]}$, because these examinees have necessary skills to solve Items 2 and 4 only. Let us denote the set of all item parameters and ability variables, that is, the set of all $\boldsymbol{\theta}_i$s and $\boldsymbol{\pi}_j$s, as $\boldsymbol{\omega}$.

The χ^2-type measure for item j, denoted henceforth as $D_j^{eq,\chi}(\boldsymbol{y}, \boldsymbol{\omega})$, is given by

$$D_j^{eq,\chi}(\boldsymbol{y}, \boldsymbol{\omega}) = \sum_{k=1}^{K} N_k \frac{(\tilde{O}_{jk} - E_{jk})^2}{E_{jk}(N_k - E_{jk})}. \tag{6}$$

The G^2-type measure, denoted henceforth as $D_j^{eq,G}(\boldsymbol{y}, \boldsymbol{\omega})$, is given by

$$D_j^{eq,G}(\boldsymbol{y}, \boldsymbol{\omega}) = 2 \sum_{k=1}^{K} \left[\tilde{O}_{jk} \log \left(\frac{\tilde{O}_{jk}}{E_{jk}} \right) + (N_k - \tilde{O}_{jk}) \log \left(\frac{N_k - E_{jk}}{N_k - E_{ik}} \right) \right]. \tag{7}$$

Each of these two statistics summarizes the fit information for an item, with large values indicating poor fit. A comparison between the posterior distribution of $D_j^{eq,\chi}(\boldsymbol{y}, \boldsymbol{\omega})$ [or $D_j^{eq,G}(\boldsymbol{y}, \boldsymbol{\omega})$] and the posterior predictive distribution of $D_j^{eq,\chi}(\boldsymbol{y}^{rep}, \boldsymbol{\omega})$ [$D_j^{eq,G}(\boldsymbol{y}^{rep}, \boldsymbol{\omega})$] provides a summary regarding the fit of item i. The comparison can be done using a graphical plot (e.g., Figure 5). Another convenient summary is the posterior predictive p-value (PPP-value) for the discrepancy measure, which can be estimated using the process described in Section 4. Here, a PPP-value very close to 0 indicates a problem (indicating that the variability in the data set appears unusually large compared to that predicted by the model).

Adding the discrepancy measures over the items, an overall discrepancy measure for assessing the fit of the model can be obtained as:

$$D_{overall}^{eq,\chi}(\boldsymbol{y}, \boldsymbol{\omega}) = \sum_j D_j^{eq,\chi}(\boldsymbol{y}, \boldsymbol{\omega}). \tag{8}$$

The posterior predictive p-value corresponding to this measure will provide an idea about the overall fit of the model to the data set.

The computation of these PPP-values does not need any approximation (e.g., a rough χ^2 approximation, as in Sinharay et al., 2004), and hence results in natural Bayesian p-values summarizing the fit of the items. The cells (item-group combination) with small frequencies (especially for low or high raw scores) is not a problem with the PPMC approach because this does not need a χ^2 assumption. However, for more stability of the discrepancy measures and the PPP-values, equivalence classes with too few examinees can be pooled. Examining the distribution of the values of $D_j^{eq,\chi}(\boldsymbol{y}^{rep,m}, \boldsymbol{\omega}_m), m = 1, 2, \ldots M$ is a way to check for stability. These quantities should look like draws from a χ^2-distribution with $(K - d - 1)$ d.f., where K is the number of equivalence classes for the problem and d is the adjustment in the d.f. due to pooling, if there are sufficient number of examinees for each class—departure from the χ^2 distribution will point to the instability of the discrepancy measure. It is possible to examine the mean, variance and histogram of the replicated discrepancies to ensure the stability of the measure.

5.2.2 Item Fit Measures Based on Raw Scores

The latent quantities (equivalent class memberships), on which the above item fit measure is based on, may themselves be estimated with error, especially for a small data set, and hence may lead to a test with poor error rates. Therefore, this part examines item fit using examinee groups based on observed raw scores (as in Orlando & Thissen, 2000). Even though the raw scores do not have a clear interpretation in the context of the multidimensional 2LC model, they are natural quantities for any assessment with dichotomous items. The approach used here to assess item fit is very similar to that in Sinharay (2003).

The first discrepancy examined is the proportion of examinees in raw score group $k, k = 1, 2, \ldots (J - 1)$ who answer item j correctly, $j = 1, 2, \ldots J$. Denote the observed proportion of examinees in group k answering item j correctly as O_{jk}. For each replicated data set, there is one (replicated) proportion correct for the j-th item and k-th group, denoted O_{jk}^{rep}. For each item j, comparison of the values of O_{jk} and O_{jk}^{rep}s,

$k, k = 1, 2, \ldots (J-1)$, provides an idea regarding the fit of the item ($O_{j0} = 0$ and $O_{jJ} = 1$; so they are not included). One way to make the comparison is to compute the posterior predictive p-value (PPP-value) for each item-group combination. This paper uses a graphical approach and suggests item fit plots, in the same spirit as Hambleton and Swaminathan (1985) to make the comparison. The approach will be described in detail with the data examples later.

The item fit plots promise to provide useful feedback about item fit, but it will be useful to summarize the fit information for each item into one number, preferably a p-value. To achieve that, this paper uses the two test statistics (χ^2-type and G^2-type) similar to those suggested by Orlando and Thissen (2000) as discrepancy measures. Because the measures depend on $E_{jk} = E(O_{jk}|\boldsymbol{\pi}, \boldsymbol{\lambda})$, the next paragraph describes computation of E_{jk}.

Computation of E_{jk}. The computation of E_{jk}, the expected proportion of examinees in group k answering item j correctly, is not trivial. Orlando and Thissen (2000) use the recursive approach of Lord and Wingersky (1984) to compute the expectations in the context of IRT models and this paper uses a slight variation of that. Suppose $S_k(\boldsymbol{\theta})$ denote the probability of obtaining a raw score of k in an J-item test by an examinee with proficiency variable $\boldsymbol{\theta}$. For convenience, the dependence of $S_k(\boldsymbol{\theta})$ on $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ is not reflected in the notation. Let $T_j(\boldsymbol{\theta})$ be the probability of success on Item j for proficiency $\boldsymbol{\theta}$ (remember that for the 2LC model, this will be either π_{j1} or π_{j0}, depending on $\boldsymbol{\theta}$ and the skill requirements of the item). Further, denote $S_k^{*j}(\boldsymbol{\theta})$ to be the probability of a raw score k at proficiency $\boldsymbol{\theta}$ on items $1, 2, \ldots j-1, j+1, \ldots J$ (i.e., omitting Item j from the set of all items). Then,

$$E_{jk} = \frac{\sum_{\boldsymbol{\theta}} T_j(\boldsymbol{\theta}) S_{k-1}^{*j}(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\sum_{\boldsymbol{\theta}} S_k(\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\lambda})}, \tag{9}$$

where the summation is over all possible values of $\boldsymbol{\theta}$ (which is 24 for this problem) and $p(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is the prior distribution on $\boldsymbol{\theta}$. Note that for an IRT model, the summation in (9) is replaced by an integration. For computing $S_k^{*j}(\boldsymbol{\theta})$ and $S_k(\boldsymbol{\theta})$, this paper uses the recursive approach of Lord and Wingersky (1984) and Orlando and Thissen (2000). The algorithm

uses the terms $T_j(\boldsymbol{\theta})$s. Note that E_{jk} does not depend on the latent proficiency variables $\boldsymbol{\theta}_i$s.

The suggested χ^2-type measure, denoted henceforth as $D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$, is given by

$$D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega}) = \sum_{k=1}^{J-1} N_k \frac{(O_{jk}-E_{jk})^2}{E_{jk}(1-E_{jk})}. \tag{10}$$

The G^2-type measure, denoted henceforth as $D_j^{raw,G}(\boldsymbol{y},\boldsymbol{\omega})$, is given by

$$D_j^{raw,G}(\boldsymbol{y},\boldsymbol{\omega}) = 2\sum_{k=1}^{J-1} N_k \left[O_{jk}\log\left(\frac{O_{jk}}{E_{jk}}\right) + (1-O_{jk})\log\left(\frac{1-O_{jk}}{1-E_{jk}}\right)\right]. \tag{11}$$

Each of the above two statistics summarizes the fit information for an item, with large values indicating poor fit. One can use $\sum_j D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ to provide an idea regarding the overall fit of the model. As discussed for the equivalence class-based measures earlier, a researcher can use graphical plots or PPP-values for $D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ [or $D_j^{raw,G}(\boldsymbol{y},\boldsymbol{\omega})$] to summarize item fit, and these measures can be monitored for stability and examinee groups can be pooled, if required.

5.3 Measure of Association Among the Items

Molenaar (1983), van den Wollenberg (1982), Reiser (1996), etc. show that the second order marginals are useful quantities to examine in the context of item response theory models in order to assess if the model can predict the association among the items properly. Sinharay and Johnson (2003), applying the PPMC method, find that the odds ratios corresponding to the responses of the examinees to pairs of items to be powerful discrepancy measures in detecting misfit of the simple IRT models. This paper examines the PPP-values for the odds ratios in the context of BNs.

Consider the item pair consisting of items i and j in a test. Denote $n_{kk'}$ to be the number of individuals obtaining a score k on item i and k' on item j, $k,k' = 0,1$. We examine the odds ratio

$$OR_{ij} = \frac{n_{11}n_{00}}{n_{10}n_{01}}.$$

The quantity on the right-hand side in the above definition is a sample odds ratio (see, for example, Agresti, 2002, p. 45) corresponding to the population odds ratio:

$$\frac{P\ (\text{item } i \text{ correct} \mid \text{item } j \text{ correct})\ /P\ (\text{item } i \text{ wrong} \mid \text{item } j \text{ correct})}{P\ (\text{item } i \text{ correct} \mid \text{item } j \text{ wrong})\ /P\ (\text{item } i \text{ wrong} \mid \text{item } j \text{ wrong})}.$$

This work investigates the performance of the sample odds ratio (referred to as the *odds ratio* hereafter) as a discrepancy measure with PPMC method. Odds ratio is a measure of association—therefore, examining it should help a researcher to detect if the fitted model can adequately explain the association among the test items. The BNs are multidimensional and attempt to capture the associations among items; if the model fails, these plots might provide some insight regarding any associations that the model failed to capture.

5.4 Differential Item Functioning

Holland (1985) suggests the Mantel-Haenszel test statistic for studying DIF in the context of item response data. The test has the strength that it is nonparametric and applies to virtually any test data.

Suppose one is interested in examining if a given item j shows DIF over a "focal group" F and a "reference group" R. In the Mantel-Haenszel test, the examinees are divided into K matching groups. In applications of the test, the raw scores of the examinees are typically used to form the groups. For an item, the data from the k-th matched group of reference and focal group members are arranged as a 2×2 table, as shown in Table 3.

The Mantel-Haenszel test statistic for an item is then given by

$$\chi^2_{MH} = \frac{(|\sum_k A_k - \sum_k \mu_k| - 0.5)^2}{\sum_k \sigma_k^2}, \tag{12}$$

Table 3.

Table for Mantel-Haenszel Statistic

	Right on an item	Wrong on an item	Total
Reference	A_k	B_k	n_{Rk}
Focal	C_k	D_k	n_{Fk}
Total	R_k	W_k	n_{+k}

where $\mu_k = n_{Rk}R_k/n_{+k}$, and $\sigma_k^2 = (n_{Rk}n_{Fk}R_kW_k)/(n_{+k}^2(n_{+k}-1))$· This paper applies the above test statistic to determine DIF for BNs.

However, the test employs matched groups based on raw scores of examinees, which are not the most natural quantities in the context of BNs. No DIF in an application of a BN means that success probabilities of the reference group and the focal group are the same conditional on the skill variables, and raw scores do not necessarily contain all information about the skills. Therefore a test incorporating that fact may perform better than the Mantel-Haenszel test.

One simple approach then is to fit the model separately to the two groups (focal and reference) and compare the values of estimates of the $\boldsymbol{\pi}_j$s for the two groups using, for example, a normal or a χ^2-test (as in Lord, 1980, in the context of testing for DIF in IRT models). The approach may be costly in a setting like here, where an MCMC algorithm is used to fit the model, especially if an investigator is interested in a number of different focal groups and/or reference groups. Further, the approach may be problematic if one or more of the two groups is small—the estimates obtained by fitting a BN to a small data set may not be stable. This line of work is not pursued any more in this paper.

This paper examines a discrepancy measure based on the Mantel-Haenszel test statistic, forming matched groups with respect to their latent skills. Matching with respect to skills (or, equivalence classes) is a natural concept in this context; however, the problem in doing so is that the skills are unknown. Fortunately, the draws from the MCMC and the PPMC method provides us a way to get around that problem.

Suppose we know the skill variables for each examinee and hence know their equivalence class memberships (this will be true for each iteration of MCMC, the generated value acting as the known values). For each equivalence class, we can compute a table as in Table 3 and can compute a discrepancy as in (12). Let us denote the measure for Item j as $D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega})$. It is straightforward to obtain an estimate of the PPP-value corresponding to $D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega})$, by computing the proportion of times $D_j^{MH}(\boldsymbol{y}^{rep,m}, \boldsymbol{\omega}_m)$ exceeds $D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega}_m)$, where $\boldsymbol{\omega}_m, m = 1, 2, \ldots M$, is a posterior sample and $\boldsymbol{y}^{rep,m}, m = 1, 2, \ldots M$, are the corresponding posterior predictive data sets.

5.5 Advantages and Disadvantages of the Suggested Plots and P-values

Based on the above discussion, the advantages of the suggested Bayesian p-values (and plots) are that

1. they provide natural probability statements from a Bayesian point of view, and
2. computing the p-values requires neither complicated theoretical derivations nor extensive simulation studies.

One disadvantage of these techniques is the conservativeness of the PPMC methods. However, as argued earlier, a conservative test is better than a test with unknown properties, and the conservativeness of the PPMC technique may be a boon here. Another disadvantage is that these techniques are based on the MCMC algorithm and hence are computation-intensive. A standard practice by the practitioners of the MCMC algorithm is to store the posterior sample obtained while fitting a model and to use the same to learn different aspects regarding the problem in the future; in that case, the computations to obtain the item fit measures should not take too long. Another issue, that might make some potential users uneasy, is that these techniques are Bayesian in nature and use prior distributions. In the absence of substantial prior information, one can use noninformative prior distributions to reduce the effect of the prior distribution on the inference. Alternatively, one can perform sensitivity analysis to detect the effect of the prior distributions.

5.6 Assessing Identifiability of Model Parameters

The Bayesian networks deal with latent classes (as discussed earlier) and hence there is a possibility of nonidentifiability or weak identifiability, which means that data provides little information about some parameters. Lindley (1971) argues that even a weakly identifiable Bayesian model may be valid in the sense that it can still describe the data via its identifiable parameters and information supplied a priori. On the contrary, although weakly identifiable models can be valid, they are not optimal for summarizing population characteristics. By failing to recognize weak identifiability, one may make conclusions based on a weakly identified parameter when one has no direct evidence beyond prior beliefs to do so (Garrett & Zeger, 2000). Further, such a model increases the chances of nonapplicability of standard asymptotic theory (e.g., as follows from Haberman, 1981) and problems with MCMC convergence. Therefore, a diagnostic for model identifiability, although not directly a model checking tool, is a necessary tool for assessing the performance of a Bayesian network.

This paper uses plots of prior vs. posterior distributions of the model parameters $\boldsymbol{\pi}$s and $\boldsymbol{\lambda}$s to assess their identifiability. The closer the two densities, the less is the information provided the data on the parameter (because posterior $\propto$ likelihood $\times$ prior), and the weaker is the identifiability of the parameter. It is possible to use numerical measures like percent overlap between the prior and posterior (e.g., Garrett & Zeger, 2000), but this paper does not use those measures.

The following five sections apply the suggested model checking tools to the mixed-number subtraction data set and simulated data sets.

6 Application of Direct Data Display

Consider Figure 2, which shows the mixed-number subtraction data and seven replicated data sets from the above 10,000 (chosen randomly), discussed in Section 3.3.

There are certain patterns in the observed data that are not present in the replicated data sets. For example, the examinees with the lowest scores could answer only two items correctly (interestingly, these are Items 4 and 5, the idiocyncrasies of which were discussed earlier—both of these can be solved without any knowledge of mixed-number subtraction).

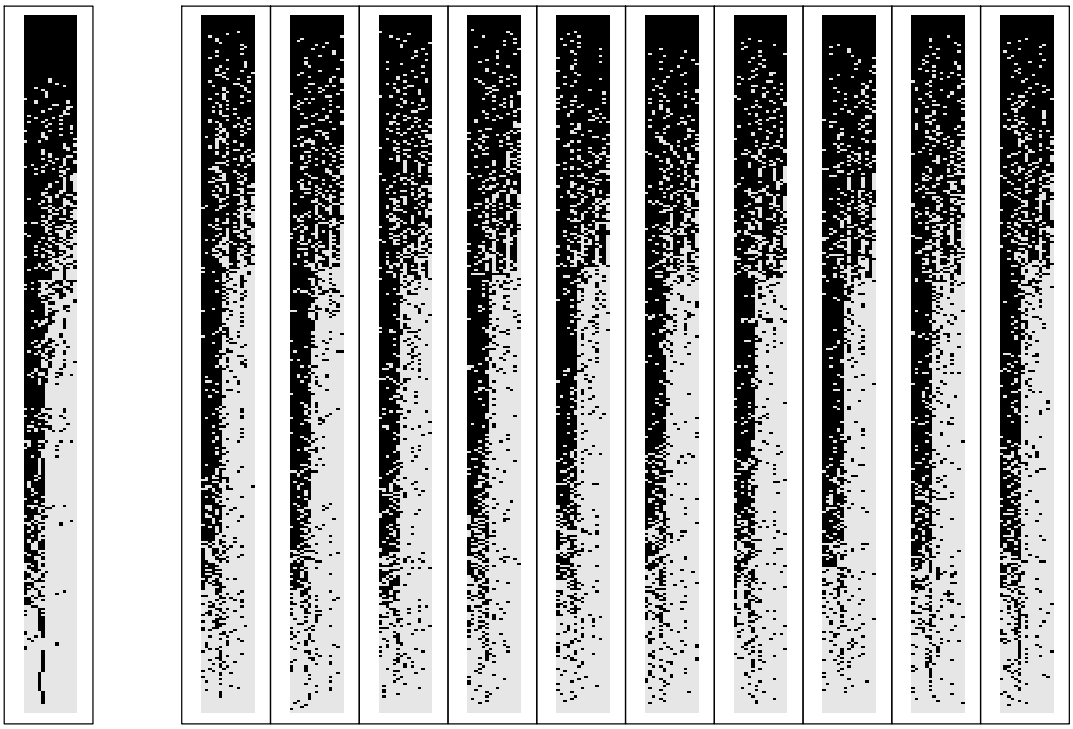

***Figure* 2. Mixed-number subtraction data and seven replicated data sets.**

Note. The left column in Figure 2 shows all observed responses, with a little black box indicating a correct response. The items, sorted according to decreasing proportion correct, are shown along the x-axis; the examinees, sorted according to increasing raw scores, are shown along the y-axis. Right columns show ten replicated data sets from the 2LC model. There are some clear differences between the observed and replicated data sets.

In the replicated data sets, however, these examinees get other items correct as well. The smartest examinees get all items correct in the observed data, which is not true for the replicated data sets. Further, the weakest half of the examinees rarely get any hard items (those belonging to Evidence Models 4 and 5) correct in the observed data, which is not true in the replicated data sets. These patterns suggest that the 2LC model is not entirely satisfactory for the data set.

To examine the performance of the 2LC model when data come from the same model, a

limited simulation study is performed. The above analysis of the mixed-number data using the 2LC model produces 10,000 posterior predictive data sets. We randomly pick one of them, fit the 2LC model to the data, and generate 10,000 posterior predictive data sets as before. Figure 3 shows the data simulated from the 2LC model and ten replicated data sets from its analysis. The data set on the left seem to fit in with the replicated data sets on the

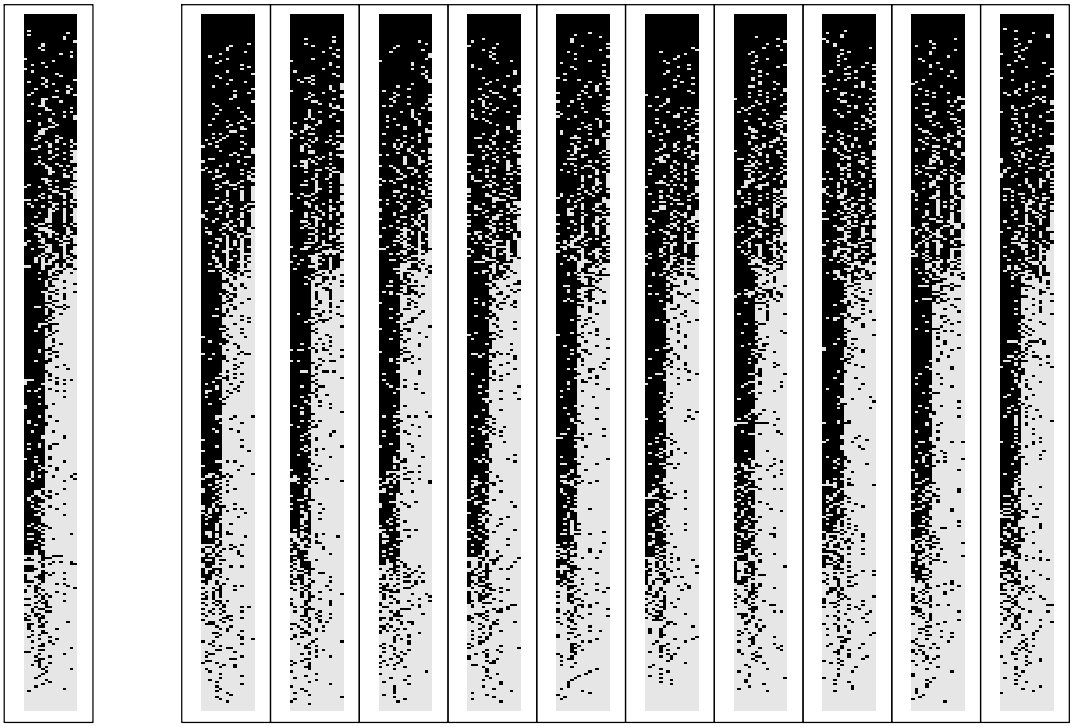

***Figure* 3. 2LC data and ten replicated data sets.**

Note. The left column shows data generated from the 2LC model, and the right columns show the ten replicated data sets from the 2LC model. The plot does not indicate any deficiency of the model.

right. The direct data display does not indicate any deficiency of the model, rightfully so.

Table 4.

P-values for Point Biserial Correlations When the 2LC Model Is Fitted to the Mixed-number Data Set

	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Observed value	.80	.50	.66	.52	.23	.78	.76	.56	.59	.71	.61	.74	.70	.75	.77
p-value	.05	.40	.04	.00	.08	.23	.45	.10	.14	.00	.02	.36	.08	.06	.43

7 Application of the Item Fit Measures

7.1 Fit of the 2LC Model to the Mixed-number Subtraction Data

This paper examines item fit for the 2LC model with the mixed-number subtraction data set.

7.1.1 Point Biserial Correlations

Table 4 shows the observed point biserials for each item and the p-values corresponding to them.

The table shows that the 2LC model underpredicts most of the point biserial correlations for this data set, and significantly so for Items 3, 4, 10, and 11.

Figure 4 shows a histogram for the predicted average biserial correlations (from the 10,000 replicated data sets). A vertical line shows the observed average (averaged over all items), which appears improbable under the 2LC model (PPP-value of 0.00).

This indicates some inadequacy of the 2LC model for these data.

7.1.2 Measures Based on Equivalence Classes

Figure 5 shows, for two items, plots of the realized discrepancy $D_j^{eq,\chi}(\boldsymbol{y}, \boldsymbol{\omega})$ and the predicted discrepancy $D_j^{eq,\chi}(\boldsymbol{y}^{rep}, \boldsymbol{\omega})$. A diagonal line is there for convenience—points consistently below the diagonal (showing that the realized discrepancy is mostly larger than the predicted discrepancy) indicate a problem. The plot towards the left shows that the model cannot explain the responses for the item adequately, the model-predicted

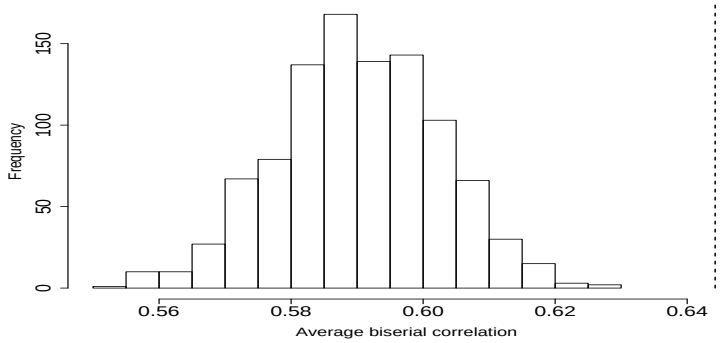

***Figure* 4. Observed and predicted average biserial correlations when the 2LC model is fitted to the mixed-number subtraction data.**

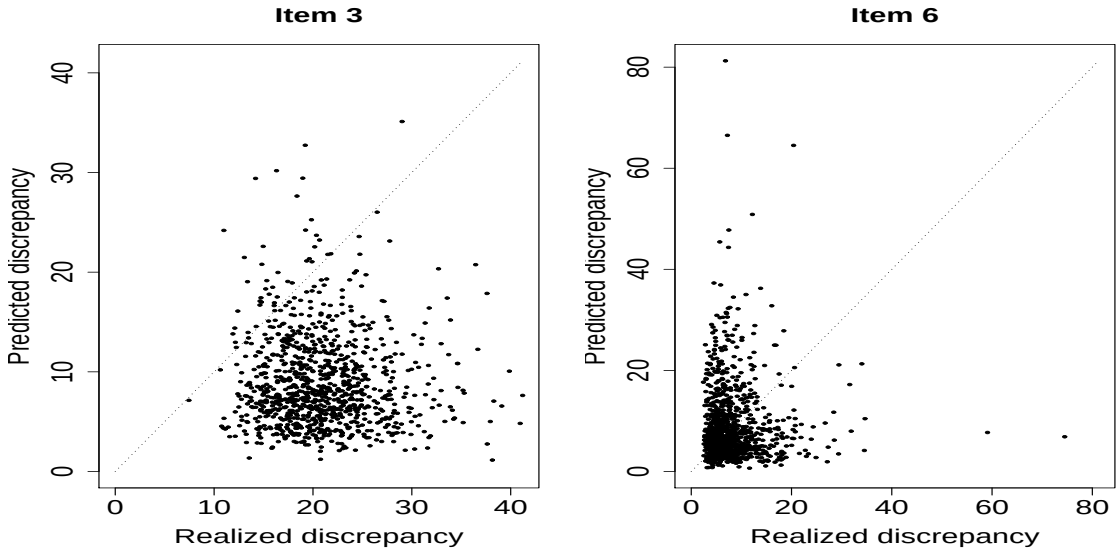

***Figure* 5. Comparison of realized and predicted discrepancies when the 2LC model is fit to mixed-number subtraction data.**

discrepancy being almost always smaller than the realized discrepancy. The PPP-value is 0.04. The plot towards the right shows the opposite—the model seems to reproduce the discrepancy and hence explain the responses for the item adequately (PPP-value = 0.49).

Table 5 shows the p-values for the suggested measures. The table first provides the overall p-values for the items. Because the p-values for the χ^2-type and G^2-type measures are very close, we will discuss the results for the former only. Then the table shows the p-values for each equivalence class. The overall p-value, which corresponds to the overall

discrepancy measure (8), is 0.01 and indicates that the model cannot explain the data set adequately.

Table 5.

Item Fit p-values Based on Equivalence Classes When the 2LC Model Is Fitted to the Mixed-number Subtraction Data

Item	p-value	Equivalence class and the average size of the equivalence class								
no.	for $D_j^{\chi}(\boldsymbol{y}, \boldsymbol{\omega})$	1 (52)	2 (9)	3 (12)	4 (2)	5 (3)	6 (12)	7 (120)	8 (34)	9 (80)
1	0.06	0.94	0.20	0.24	0.60	0.60	0.38	0.23	0.79	0.11
2	0.74	0.65	0.15	0.18	0.04*	0.03*	0.38	0.17	0.72	0.38
3	0.04*	0.99*	0.33	0.27	0.08	0.16	0.41	0.03*	0.73	0.53
4	0.00*	0.15	0.31	0.72	0.04*	0.02*	0.78	0.99*	0.07	0.00*
5	0.30	0.04*	0.39	0.47	0.51	0.11	0.42	0.66	0.59	0.32
6	0.49	0.81	0.28	0.19	0.07	0.08	0.32	0.39	0.67	0.38
7	0.56	0.65	0.26	0.50	0.11	0.08	0.25	0.39	0.69	0.27
8	0.43	0.58	0.48	0.36	0.10	0.16	0.44	0.57	0.22	0.17
9	0.28	0.82	0.26	0.27	0.08	0.00*	0.06	0.54	0.42	0.07
10	0.00*	1.00*	0.78	0.30	0.10	0.15	0.42	0.10	0.02*	0.45
11	0.19	0.70	0.28	0.44	0.08	0.02*	0.08	0.72	0.40	0.01*
12	0.49	0.76	0.33	0.21	0.07	0.06	0.24	0.45	0.68	0.45
13	0.04*	0.99*	0.69	0.27	0.51	0.12	0.12	0.06	0.46	0.56
14	0.52	0.86	0.28	0.34	0.08	0.16	0.36	0.26	0.41	0.33
15	0.41	0.54	0.22	0.29	0.47	0.19	0.14	0.68	0.70	0.37

*$p < 0.05$ or $p > 0.95$.

For convenience, the p-values less than 0.05 or more than 0.95 for the equivalence classes are marked with an asterisk. The table also shows the posterior means (up to the nearest integer) of the number of examinees in each equivalence class. An extreme p-value for a class with bigger size (e.g., Class 7) is more severe. On the other hand, extreme p-values in small classes, such as Classes 4 and 5, should not be of much concern.

Interestingly, the results are quite similar to those obtained using the rough χ^2-

approximation in Sinharay et al. (2004). The items that are significant at the 5% level are Items 3, 4, 10, and 13 in both the analyses, with Items 4 and 10 being significant at the 1% level in both. However, Item 1, with a p-value of 0.15 appeared to be fit reasonably by the model in Sinharay et al. (2004) while this analysis, with a p-value of 0.06, indicates that the item is not really fit well by the model and reveals the weakness of the χ^2 approximation. Remembering the conservativeness of the PPMC method, this is an important difference.

The last nine columns of the table provide deeper insight regarding items with extreme p-values. For example, the 2LC model seems to have failed for Equivalence Class 1, with a number of extreme p-values, and does not perform too well for Equivalence Class 9 either. The p-values corroborate the evidence found in Sinharay et al. (2004) that the model overpredicts the scores of individuals with low proficiency and underpredicts those for high proficiency.

7.1.3 Measures Based on Raw Scores

The item fit plots. Figure 6 provides item fit plots for all items for this example. The horizontal axis of each plot denotes the groups (i.e., the raw scores) of the examinees.

The vertical axis represents the proportion corrects for the groups. For any group, a point denotes the observed proportion correct and a box represents the distribution of the replicated proportion corrects for that group. A line joins the observed proportions for ease of viewing. The whiskers of the box stretch till the 5th and 95th percentiles of the empirical distribution and a notch near the middle of the box denotes the median. The width of the box is proportional to the square root of the observed number of examinees in the group (the observed number in a group provides some idea about the severity of a difference between the observed and replicated proportions; a significant difference for a large group is more severe than that for a small group). For any item, too many observed proportions lying far from the center of the replicated values or lying outside the range spanned by the whiskers (i.e., lying outside a 90% prediction interval) indicate a failure of the model to explain the responses to the item. In Figure 6, the plot for all the items other than 2, 6, 7, 14, and 15 provide such examples, with a number of proportions (out of a total of 14) lie outside the 90% prediction interval and a few others lie far from the center of the box.

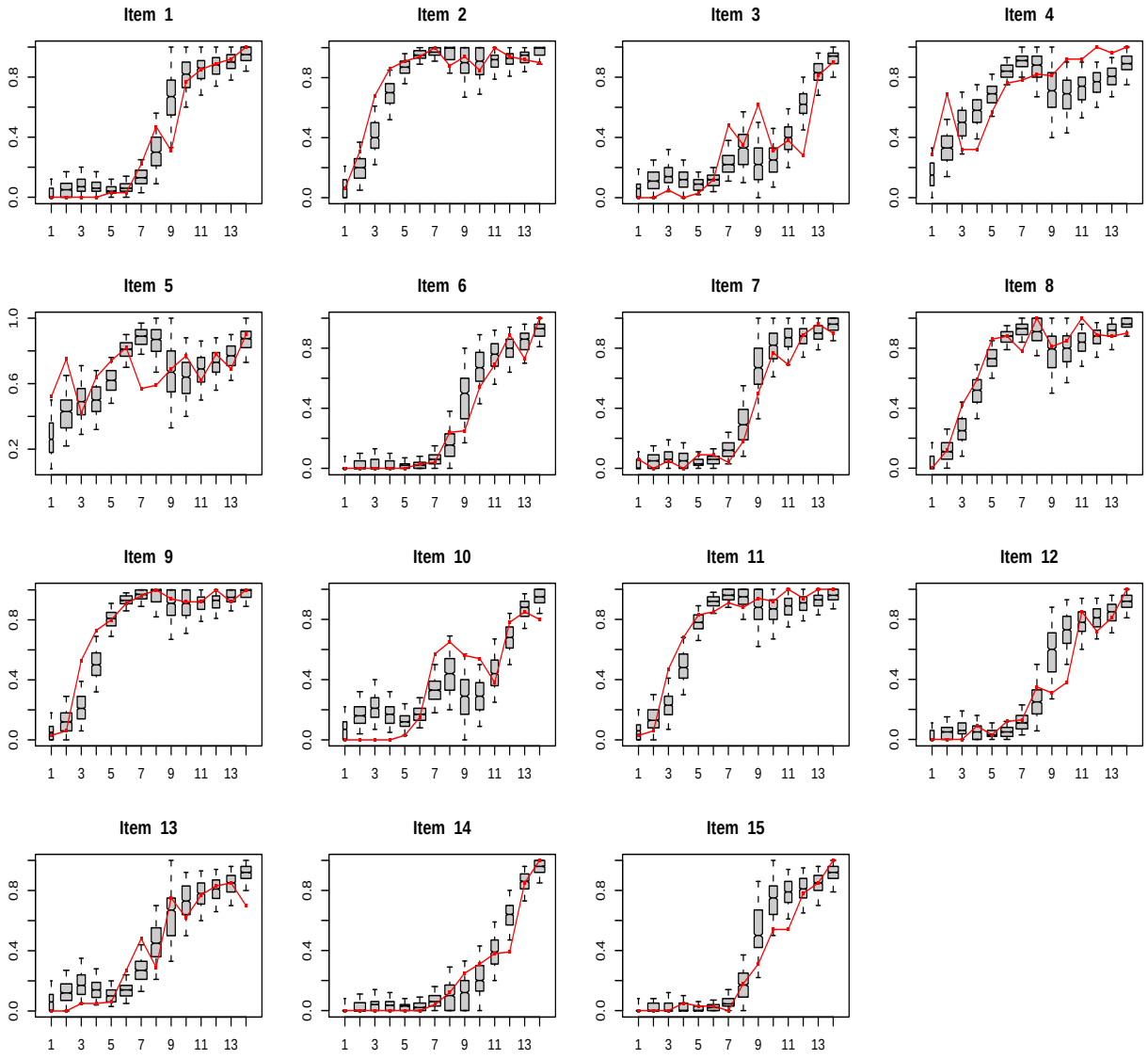

***Figure* 6. Item fit plots when the 2LC model is fit to the mixed number subtraction data.**

Other than providing an overall idea about the fit for an item, the suggested item fit plot also provides some idea about the region where the misfit occurs (e.g., for low- or high-scoring individuals), the direction in which the misfit occurs (e.g., whether the model overestimates/underestimates the performance for the discrepant regions), certain aspects of the item like its difficulty. One can create the plots with 95% prediction intervals as well.

The item fit plots point to an interesting feature of the Bayesian networks—the

model-predicted proportion correct score occasionally goes down as the raw score increases (e.g., for Items 4, 5, 8, 10)—this is an outcome of the multidimensional nature of the model.

The χ^2-type discrepancy measures and Bayesian p-values. Table 6 shows the item fit p-values corresponding to the discrepancy measures based on raw scores.

Table 6.

P-values for $D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ and $D_j^{raw,G}(\boldsymbol{y},\boldsymbol{\omega})$ When the 2LC Model is Fitted to the Mixed-number Data Set

Discrepancy	Item														
measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$	.07	.26	.00	.00	.00	.53	.42	.06	.05	.00	.03	.06	.01	.30	.30
$D_j^{raw,G}(\boldsymbol{y},\boldsymbol{\omega})$	.11	.19	.00	.00	.00	.52	.30	.23	.03	.03	.02	.08	.04	.38	.28

The table shows that the model cannot adequately explain the responses for Items 3, 4, 5, 9, 10, 11, and 13. It is interesting to note that this set of items includes items in Table 5 that misfit (3, 4, 10, 13), that is, items with the measures based on the equivalence classes. Additionally, this set has three more items. These responses might be due to the problems mentioned earlier about diagnostics based on latent quantities (what is called "observed" there is actually estimated and the estimation error may affect the results of the diagnostic) and probably indicates the superiority of the measures based on raw scores over those based on the equivalence classes—this issue needs further research. The p-value corresponding to the overall fit measure $\sum_j D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ is 0.00, indicating the poor overall fit of the 2LC model.

7.2 Fit of the 2LC Model to a Data Set Simulated From the 2LC Model

As before, to examine the performance of the 2LC model when data come from the same model, we randomly pick one of the posterior predictive data sets obtained while fitting the 2LC model to the mixednumber data set, and use BUGS to fit the 2LC model to the data set.

7.2.1 Point Biserial Correlations

The 2LC model explains the point biserials adequately as none of the p-values corresponding to the biserials for the items is significant at 5% level. The p-values corresponding to the average and variance of the biserials are not significant either.

7.2.2 Measures Based on Equivalence Classes and Raw Scores

Table 7 shows the p-values for the item fit discrepancy measures. The p-value corresponding to the overall discrepancy measure (8) is 0.56—therefore, no evidence for a model misfit is found.

Table 7.

P-values for $D_j^{eq,\chi}(\boldsymbol{y},\boldsymbol{\omega})$***,*** $D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ ***and*** $D_j^{raw,G}(\boldsymbol{y},\boldsymbol{\omega})$ ***When the 2LC Model Is Fitted to a Simulated Data Set***

Discrepancy	Item														
measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$D_j^{eq,\chi}(\boldsymbol{y},\boldsymbol{\omega})$	.37	.53	.56	.49	.28	.19	.74	.60	.41	.51	.72	.45	.66	.76	.52
$D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$	.58	.18	.33	.01	.50	.10	.21	.17	.51	.48	.65	.90	.27	.18	.12
$D_j^{raw,G}(\boldsymbol{y},\boldsymbol{\omega})$	.72	.50	.16	.00	.45	.34	.24	.06	.53	.48	.73	.84	.20	.16	.05

The item fit p-value for $D_j^{eq,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ is not significant for any item, which is expected because the model fitted is the correct model. There are a few extreme p-values for the item-equivalence class combinations, but that can be attributed to chance, and all but one of them are for the small equivalence classes (classes 2-6).

The item fit p-values corresponding to $D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ and $D_j^{raw,G}(\boldsymbol{y},\boldsymbol{\omega})$ show that the model cannot adequately explain Item 4, which can be attributed to chance. The p-values for a couple of other items are low as well, although not significant.

7.3 Fit of the 4LC Model to the Mixed-number Subtraction Data Set

This paper proceeds to examine item fit with the 4LC model, an extended version of the 2LC model and fit to the mixed-number subtraction data by Sinharay et al. (2004).

7.3.1 Brief Description of the 4LC Model

The examinees not having the necessary skills to solve an item are divided into two subgroups:

- Those who have not mastered Skill 1
- Those who have mastered Skill 1, but still lack one or more additional skills necessary for solving the item

Different success probabilities are assigned to each of these subgroups.

Similarly, the group of examinees having all the necessary skills to answer an item correctly are divided into two subgroups: those who have mastered all five skills and those who are yet to master one or more skills. Different success probabilities are assigned to each of these two subgroups.

In the end, the proficiency model for the 4LC model is the same as the 2LC model, but the link model is given by:

$$X_{ij}|\pi_{jm}, (\delta_{i(s)} + I_i^1 + I_i^{all} = m) \sim \text{Bern}(\pi_{jm}), \text{ for } m = 0, 1, 2, 3,$$

where I_i^1 is as defined earlier and I_i^{all} is the indicator function denoting whether the examinee i has all of the five skills or not. The prior distributions on the π_{jm}s are taken as:

$$\pi_{j0} \sim Beta(2, 25), \quad \pi_{j1} \sim Beta(6, 21), \quad \pi_{j2} \sim Beta(21, 6), \quad \pi_{j3} \sim Beta(25, 2)\cdot$$

This model is expected to explain the performance of the examinees with low and high proficiencies better than the 2LC model.

7.3.2 Point Biserial Correlations

The PPP-values corresponding to the point biserials, shown in Table 8, indicate that the 4LC model significantly underpredicts a few point biserial correlations for this data set

while it also overpredicts a few. The p-value for the average biserial correlation is 0.04, which is significant; the same is true for the p-value of 0.03 for the variance of the biserial correlations. The above results indicate that the 4LC model cannot adequately describe the biserial correlations for the data set.

Table 8.

P-values for Point Biserial Correlations, $D_j^{eq,\chi}(\boldsymbol{y},\boldsymbol{\omega})$, and $D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ When the 4LC Model Is Fitted to the Mixed-number Data Set

	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Biserial	.01	.66	.29	.18	.97	.08	.32	.29	.55	.00	.26	.38	.21	.07	.37
$D_j^{eq,\chi}(\boldsymbol{y},\boldsymbol{\omega})$	.46	.68	.37	.08	.16	.39	.60	.39	.51	.32	.49	50	.48	.56	.30
$D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$	.25	.64	.01	.14	.00	.34	.54	.22	.23	.01	.34	.10	.20	.18	.29

7.3.3 Item Fit Analysis Using Measures Based on Equivalence Classes

Table 8 also shows the PPP-values for the measure $D_j^{eq,\chi}(\boldsymbol{y},\boldsymbol{\omega})$. The overall p-value is 0.24, which indicates that the model underpredicts the variability in the data set, but the extent of underprediction is not significant.

The table shows that none of the p-values corresponding to $D_j^{eq,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ is extreme at 5% level. Item 4 appears borderline with a p-value of 0.08. There are only a handful extreme p-values corresponding to the item-equivalence class combinations (values not shown). The findings are mostly similar to those in Sinharay et al. (2004), except for Item 4, for which the rough χ^2 approximation in the latter led to a p-value of 0.02, whereas the p-value is 0.08 here.

7.3.4 Item Fit Analysis Using Measures Based on Raw Scores

The PPP-values for $D_j^{raw,\chi}(\boldsymbol{y},\boldsymbol{\omega})$ in Table 8 show that the model cannot adequately explain Items 3, 5, and 10, which is in contradiction to the findings with $D_j^{eq,\chi}(\boldsymbol{y},\boldsymbol{\omega})$. The overall p-value for the raw score-based measure is 0.00, indicating that even the 4LC model

does not perform adequately for the data set.

7.4 Discussion

The PPMC method with the suggested discrepancy measures seems to be powerful in detecting misfit of the 2LC model to the mixed-number subtraction data. The method has been criticized as being conservative, but seems to have enough power even for such a small data set. The examples show that the PPMC method can provide a useful solution to the difficult problem of assessing item fit for Bayesian networks. Of the two types of measures, those based on the raw scores of examinees seem to be more powerful than those based on equivalence classes.

8 Application of the Measure of Association Among the Items

Figure 7 shows the PPP-values, obtained by fitting the 2LC model to the mixed-number data set, corresponding to the odds ratio for each item pair. A triangle indicates a

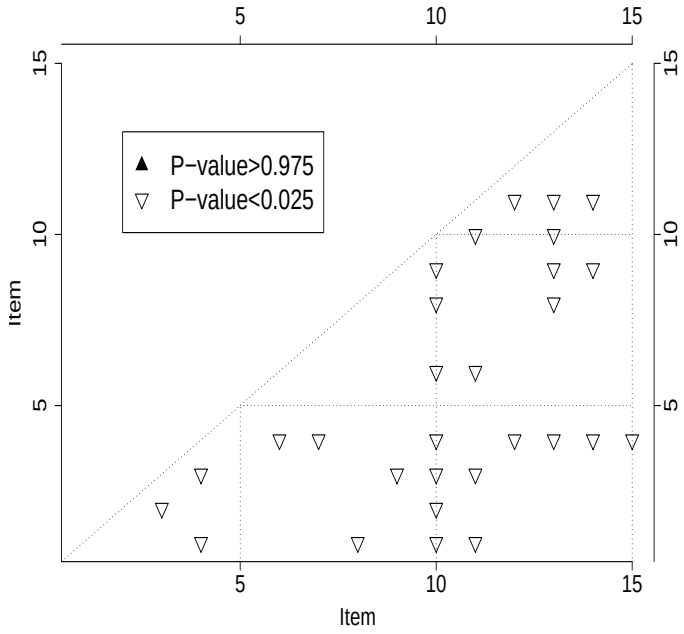

***Figure* 7. PPP-values for odds ratios for 2LC model fit to the mixed-number subtraction example.**

significant overprediction (high p-value) by the model, while an inverted triangle indicates

a significant underprediction (low p-value). Horizontal and vertical grid-lines at multiples of 5, vertical axis-labels on the right side of the figure, and a diagonal line are there for convenience. The plot has too many inverted triangles, indicating that the 2LC model underpredicts the associations among the items on too many occasions. The model especially performs poorly for item pairs involving Items 3, 4, 10, 11, and 13. Remember that four of these five were found problematic in item fit analysis earlier. The 2LC model could not capture adequately the dependence structure among the items adequately.

When the 2LC model is fitted to a data set simulated from the 2LC model (and analyzed earlier, during item fit analysis), a similar plot (not shown here) has no significant p-value at all, indicating that the 2LC model predicts the associations among the items of a data set generated adequately from the 2LC model.

Figure 8 shows the posterior predictive p-values, obtained by fitting the 4LC model to the mixed-number data set, corresponding to the odds ratio for each item pair. The

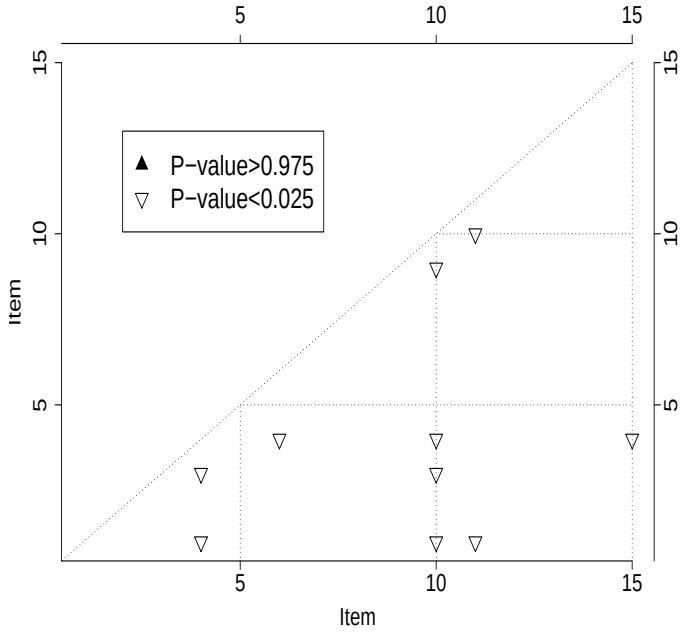

***Figure* 8. PPP-values for odds ratios for 4LC model fit to the mixed-number subtraction example.**

4LC model does a better job of predicting the associations among the items for the mixed-number subtraction data set, even though the proportion of extreme p-values at

5% level is more than what can be attributed to chance. The model struggles to explain associations for pairs involving Items 4 and 10.

9 Measuring Differential Item Functioning

9.1 Analysis of a Simulated Data Set With No DIF

Unfortunately, no demographic information regarding the examinees taking the mixed-number subtraction test are available. Therefore, this paper applies the DIF measures to data sets simulated from the 2LC model; randomly chosen subsets of the simulated data are used as the focal group and the reference group. Table 9 provides the p-values corresponding to the Mantel-Haenszel (MH) test (Holland, 1985) and the measure $D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega})$ when 150 randomly chosen examinees (from a total of 325) form the focal group and the remaining form the reference group. Because there is no DIF present, if the measures have good Type I error properties, the chance to observe a significant p-value is 0.05 only.

Table 9.

P-values for DIF Analysis for a Simulated Data Set With No DIF

	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MH	.64	.57	.99	.69	.20	.75	.59	.32	.82	.48	.53	.66	.27	.14	.15
$D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega})$	.85	.61	.82	.79	.14	.78	.56	.57	.60	.36	.61	.85	.37	.27	.19

The above table shows that there is a considerable similarity between the two sets of p-values. Most importantly, no item has significant p-value at 5% level under any of the two analyses, which is expected.

To study the issue further, the above process is repeated 100 times, each time treating a different random sample of 150 examinees as the focal group and the remaining as the reference group. The overall proportion of times a Mantel-Haenszel p-value is significant at the 5% level (which is like a Type I error rate) is 0.05, right at the nominal significance level. The same proportion is 0.04 for the posterior predictive p-values, probably pointing

to the slight conservativeness of them.

9.2 Analysis of a Simulated Data Set With DIF

As a next step, this paper analyzes a data set under conditions when DIF is present. The starting point is a data set simulated from the 2LC model (that was analyzed earlier). Table 9 shows that the p-value for Item 6 is quite high for both the methods—so DIF will be artificially introduced to that item. The overall proportion correct of the item is 0.31. A sample of 150 examinees (the same sample corresponding to Table 9) is chosen randomly, which will act as the focal group while the remaining examinees form the reference group. Of the wrong responses (score 0) to Item 6 in the focal group, 15 were reset to score 1. To ensure that the overall ability of the focal group does not increase by this process, a score of 1 on another randomly chosen item for each of these 15 examinees was reset to 0. This process artificially introduces a difference of proportion correct score of 0.1 for Item 6 between the two groups. The data set is analyzed using the two methods and the same focal and reference groups employed for the data generation are used in analyzing the data.

Table 10 provides the p-values corresponding to the Mantel-Haenszel test and the measure $D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega})$.

Table 10.

P-values for DIF Analysis for a Simulated Data Set With DIF

	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MH	.48	.93	.77	.46	.29	.01	.59	.32	.34	.67	.67	.66	.27	.14	.15
$D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega})$	.81	.73	.71	.89	.15	.00	.53	.57	.29	.51	.57	.87	.41	.28	.20

The results are encouraging. The only p-value significant at 5% level under both the methods is that for Item 6 (although that for $D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega})$ is slightly more extreme than that for the Mantel-Haenszel test). Comparing to Table 9, we see that the p-values for a number of items remain very close to the those in Table 9 while some are slightly different.

Further, few more data sets are simulated under conditions where DIF is present. The

data generating scheme is very similar to that employed above, except that the number of wrong responses to Item 6 which are reset to correct responses in the focal group is reduced gradually from 15 to lower numbers. The DIF p-values for Item 6 for both the statistics increase as an outcome, the p-value for the Mantel-Haenszel test statistic always being slightly larger than that for $D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega})$. For example, when 11 scores of 0 are reset to 1, the p-value for the Mantel-Haenszel test statistic is 0.051 while that for $D_j^{MH}(\boldsymbol{y}, \boldsymbol{\omega})$ is 0.039.

From the above limited simulations, both the Mantel-Haenszel DIF statistic and the posterior predictive p-value seem to perform reasonably, but the issue needs further investigation.

10 Assessing Identifiability of the Model Parameters

Figure 9 and 10 show the prior distributions (dashed lines) and posterior distributions for all the $\boldsymbol{\pi}$s and all the $\boldsymbol{\lambda}$s obtained by fitting the 2LC model to the mixed-number subtraction data.

The plots indicate identifiability problems with the 2LC model in regard to the mixed-number subtraction data set. From Figure 9, it is clear that π_{20}, π_{90}, and $\pi_{11,0}$ are barely identifiable; one reason for this is the presence of few examinees without Skill 1 (the only skill required to solve Item 2) who answer Item 2 correctly. The figures also indicate that the prior distributions for π_{j0}s for Items 4 and 5 (whose specialities are discussed earlier) are not in agreeement with the corresponding posteriors.

Figure 10 shows more severe problems with the model applied. Of the 18 $\boldsymbol{\lambda}$s plotted, the prior and posterior completely overlap for six ($\lambda_{2,0}$, $\lambda_{5,0}$, $\lambda_{WN,0,0}$, $\lambda_{WN,0,1}$, $\lambda_{WN,0,2}$, $\lambda_{WN,1,0}$). The two distributions are not too far for a few others ($\lambda_{5,1}$, $\lambda_{WN,1,1}$, $\lambda_{WN,1,2}$, $\lambda_{WN,2,1}$, $\lambda_{WN,2,2}$). The nonidentifiability of the parameters indicate that, for example, there is little information (because of the lack of such an item in the test) about the chance of having Skill 2 in the absence of Skill 1.

The above findings point to an undesirable characteristic of the Bayesian networks. It is apparent that even though a model may appear attractive from a cognitive perspective, its parameters may not be well-identified. A user of BNs and similar models like DINA (Junker & Sijtsma, 2001), NIDA (Maris, 1999), higher order latent trait models (de la

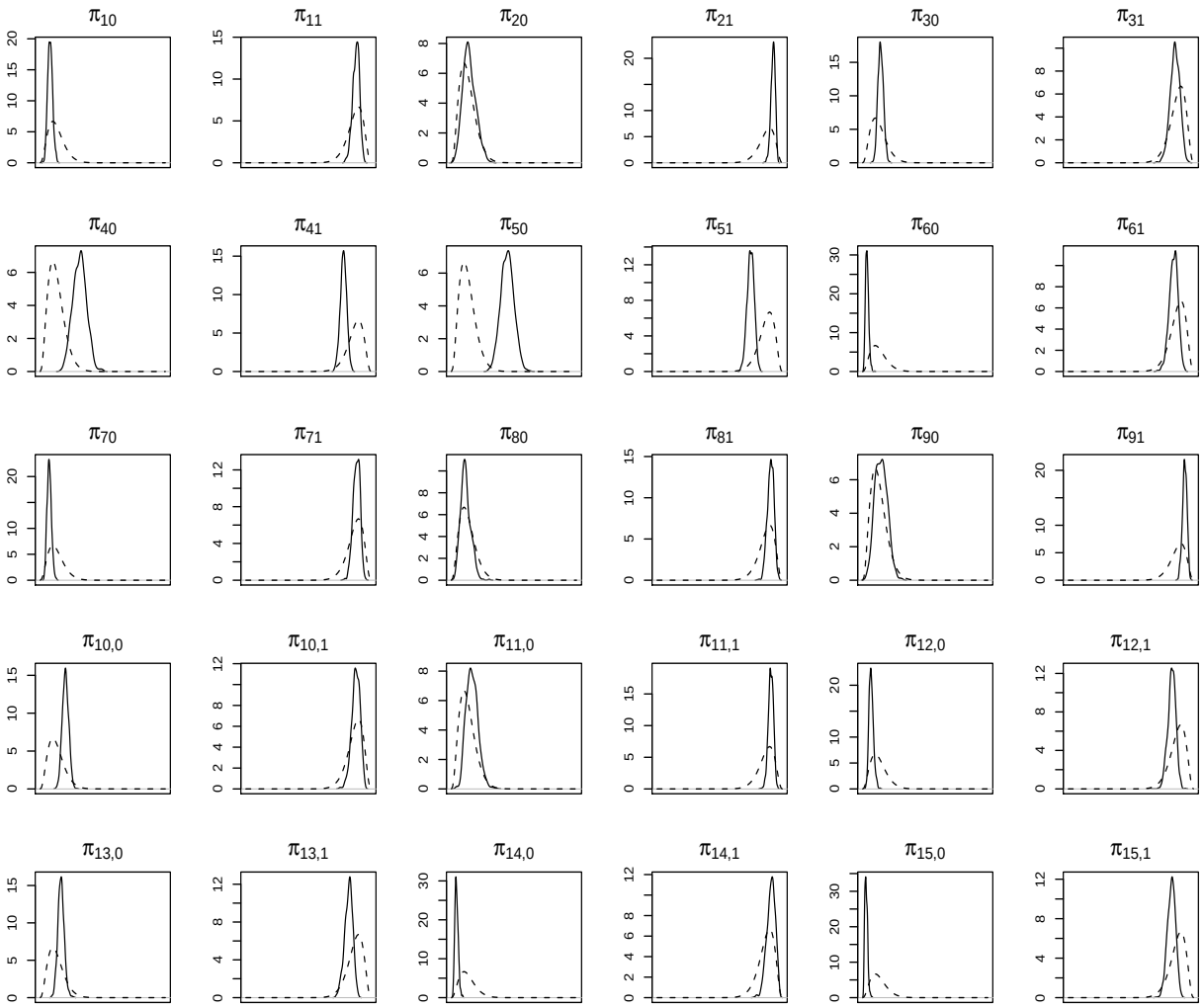

***Figure* 9. Prior vs. posterior distributions of πs for 2LC model fit to the mixed-number data.**

Note. Prior distributions are dashed lines and posterior distributions are solid lines. The x-axis, not labeled, ranges from 0 to 1.

Torre & Douglas, in press), the fusion model (Hartz et al., 2002) etc. should therefore exercise caution. The best option is to design an assessment that appropriately captures the necessary skills that interest the researcher and then apply a model whose parameters are well-identified. If not, one has to be careful in making conclusions from such models to prevent the possibility of drawing inferences based on the weakly identified parameters.

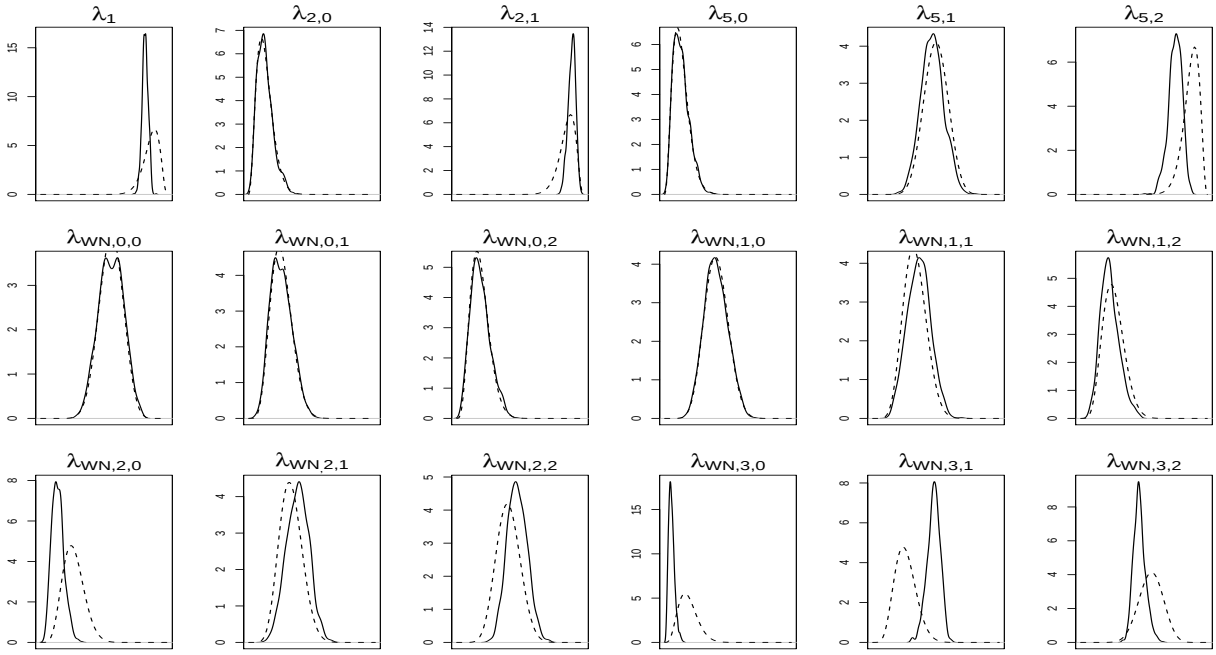

***Figure* 10. Prior vs. posterior distributions of λs for 2LC model fit to the mixed-number data.**

Note. Prior distributions are dashed lines and posterior distributions are solid lines. The x-axis, not labeled, ranges from 0 to 1.

11 Conclusions

There exists no unanimously-agreed-upon statistical techniques for assessing fit for Bayesian networks. Yan et al. (2003) and Sinharay et al. (2004) suggest a number of diagnostic measures for such models, but there is substantial scope of further investigation in the area. This paper first shows how to use the PPMC method to assess item fit for simple BNs.

First, a direct data display, comparing the observed data and a number of predicted data sets, is suggested as a quick check of the fit of the model. The display is found to provide useful insight about the fit of the model in the real data example.

Second, the posterior predictive p-values (PPP-values) corresponding to the point biserial correlations (a natural quantity for any assessment with dichotomous items) are shown to provide useful feedback about item fit. These can be seen as diagnostics for item fit.

Third, this paper suggests two sets of item fit diagnostics, one set using examinee

groups based on equivalence class membership (which is determined by the skills) of the examinees and another set using groups based on raw scores of the examinees. Another tool suggested is the item fit plot corresponding to the raw score-based item fit diagnostic. Limited simulation studies and real data analyses demonstrate the usefulness of the item fit plots and the corresponding PPP-value. In combination with the item fit plots suggested in Sinharay et al. (2004), the measures based on equivalence classes promise to provide useful feedback regarding item fit for Bayesian networks. The item fit diagnostics based on raw scores seem to have more power than those based on equivalence classes.

The PPMC method using the odds ratio measure is able to detect the inability of the BNs to explain the association among items. The graphical approach, intuitively appealing and straightforward to apply, provides useful information regarding the fit of the model regarding interactions between items.

This paper also investigates the issue of DIF in the context of BNs, using the Mantel-Haenszel test (Holland, 1985) and one based on posterior predictive checks. Both the statistics seem to perform reasonably well for BNs in very limited simulations, but the issue needs further investigation.

Finally, the paper suggests assessing identifiability of the model parameters, by comparing the prior and posterior distributions of the model parameters, as a part of any analysis fitting a BN. Such comparisons show that the 2LC model, applied to the mixed-number subtraction data set in a number of publications, has severe identifiability problems. This phenomenon warns the users of the BNs and other similar models to be cautious—in an application, it is highly undesirable to make conclusions based on weakly identified parameters.

Overall, this paper suggests useful model diagnostics for BNs to address the prime areas of model fit that are of concern to psychometricians. The PPP-values are known to be conservative (see, e.g., Robins et al., 2000)—so the Type I error rate for the suggested measures are not a concern. Rather, the concern is about the power of the measures. The measures seem to have considerable power in this study. They detect the misfit of the 2LC model even though the data set is rather small, supporting the findings from Sinharay et al. (2004) that the model is inadequate for the data set. The diagnostics promise to be

useful not only to the users of BNs, but also to those using other similar models regarding formative assessment, such as the DINA/NIDA models, the higher order latent trait models (de la Torre & Douglas, in press), and the fusion model (Hartz et al., 2002). Most of these models are fitted using the MCMC algorithm and hence the PPMC method with the suggested measures is a natural candidate for a model checking tool there.

There are a number of issues that need further examination, however. Extensive simulations examining the Type I error and power of the suggested measures may reveal interesting issues; this paper does not delve into that. For large numbers of items and or skills, the number of equivalence classes may be large—it will be interesting to examine whether it is feasible to use some of these methods within reasonable time limits. This paper discusses diagnostics in the context of a simple BN; application or extension of these to a more complicated BN (e.g., one where the skill variable/response might be polytomous) is not straightforward—this issue needs further research.

References

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223-238.

de la Torre, J., & Douglas, J. (in press). Higher-order latent trait models for cognitive diagnosis. *Psychometrika.*

Garrett, E. S., & Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics, 56,* 1055-1067.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* Chapman & Hall: New York.

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

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

Haberman, S. (1981). Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction *Annals of Statistics, 9,* 1178–1186.

Haertel, E. H., & Wiley, D. E. (1993) Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of test* (pp. 359–384). Hillsdale, NJ: Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Academic Publishers.

Hartz, S., Roussos, L., & Stout, W. (2002). *Skills diagnosis: Theory and practice. User Manual for Arpeggio software.* Princeton, NJ: ETS.

Holland, P. (1985). On the study of differential item performance without IRT. *Proceedings of the 27th annual conference of the Military Testing Association* (Vol I, pp. 282-287), San Diego.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with non-parametric item response theory. *Applied Psychological Measurement, 23*(3), 258-272.

Klein, M. F., Birnbaum, M., Standiford, S. N., & Tatsuoka, K. K. (1981). *Logical error analysis and construction of tests to diagnose student "bugs" in addition and subtraction of fractions* (Research Report 81-6). Urbana, IL: Computer-based Education Research Laboratory, University of Illinois.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local Computations with probabilities on graphical structures and their applications to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B, 50,* 157-224.

Lindley, D. V. (1971). *Bayesian statistics: A review.* Philadelphia: Society for Industrial and Applied Mathematics.

Lord, F. M. (1980). *Applications of item Response Theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 453–461.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187-212.

Mislevy, R. J. (1995). Probability-based Inference in Cognitive Diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (2001). Bayes nets in educational assessment: Where the numbers come from. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 437-446.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective, 1.*(1) 3–62.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48*(1), 49–72.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Pearl, J. (1988). *Probabilistic reasoning in intelligence systems: Networks of plausible inference*. San Mateo: Morgan-Kaufmann.

Reiser, M. (1996). Analysis of residuals for the multinomial item response models. *Psychometrika, 61*(3), 509-528.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). The asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association, 95*, 1143-1172.

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

Sinharay, S. (2003). *Bayesian item fit analysis for dichotomous item response theory models* (ETS RR-03-34). Princeton, NJ: ETS.

Sinharay, S., & Johnson, M. S. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models* (ETS RR-03-28). Princeton, NJ: ETS.

Sinharay, S., Almond, R., & Yan, D. (2004). *Model checking for models with discrete proficiency variables in educational assessment* (ETS RR-04-04). Princeton, NJ: ETS.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1995). BUGS: Bayesian inference using Gibbs sampling (Version 0.50) [Computer software]. Cambridge, MA: MRC Biostatistics Unit.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold & M. G. Shafto (Eds), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Williamson, D. M., Almond, R. G., & Mislevy, R. J. (2000). Model criticism of Bayesian networks with latent variables. *Uncertainty in Artificial Intelligence Proceedings 2000,* 634-643.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123-140.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment* (ETS RR-03-32). Princeton, NJ: ETS.