Research Report

Assessing Fit of Models With Discrete Proficiency Variables in Educational Assessment

Sandip Sinharay

Russell Almond

Duanli Yan

Research & Development

February 2004
RR-04-07

Assessing Fit of Models With Discrete Proficiency Variables in Educational Assessment

Sandip Sinharay, Russell Almond, and Duanli Yan

ETS, Princeton, NJ

February 2004

Abstract

Model checking is a crucial part of any statistical analysis. As educators tie models for testing to cognitive theory of the domains, there is a natural tendency to represent participant proficiencies with latent variables representing the presence or absence of the knowledge, skills, and proficiencies to be tested (Mislevy, Almond, Yan, & Steinberg, 2001). Model checking for these models is not straightforward, mainly because traditional χ^2-type tests do not apply except for assessments with a small number of items. Williamson, Mislevy, and Almond (2000) note a lack of published diagnostic tools for these models.

This paper suggests a number of graphics and statistics for diagnosing problems with models with discrete proficiency variables. A small diagnostic assessment first analyzed by Tatsuoka (1990) serves as a test bed for these tools. This work is a continuation of the recent work by Yan, Mislevy, and Almond (2003) on this data set. Two diagnostic tools that prove useful are Bayesian residual plots and an analog of the item characteristic curve (ICC) plots. A χ^2-type statistic based on the latter plot shows some promise, but more work is required to establish the null distribution of the statistic. On the basis of the identified problems with the model used by Mislevy (1995), the suggested diagnostics are helpful to hypothesize an improved model that seems to fit better.

Key words: Bayesian methods, Bayesian residual, item fit, Markov chain Monte Carlo, model fit, person fit, posterior predictive model checking

Acknowledgments

The authors thank Matthew Johnson (Baruch College, New York) for his invaluable advice resulting in significant improvement of the paper. The authors also thank Ting Lu for her assistance with computing in this work; Kim Fryer, Rochelle Stern, and Janice Lukas for their help with proofreading; and Hal Stern, Robert Mislevy, Shelby Haberman, Dan Eignor, and Paul Holland for their useful comments about the work.

1. Introduction

Model checking is a crucial part of any model-based statistical analysis, providing a vital sanity check that the theory underlying the model can actually predict the phenomena observed in the data. Model checking can identify individuals whose responses are not explained well by the model, and it can suggest improvements to the model and hence the underlying process that generated the data. Thus, model checking is an important part of the round trip between theory and empirical observation, which is the basis of the scientific method.

Model checking in educational testing presents special challenges as the part of the model describing student proficiency almost always consist purely of latent variables. The bulk of work to date, which is nowhere near completion, is based on unidimensional item response theory (IRT) models where the proficiency model consists of a single continuous latent trait (e.g., van der Linden & Hambleton, 1997).

The models with discrete proficiencies are of particular interest because of their ability to capture expert opinion about the proficiencies used to solve assessment problems and their interrelationships (Mislevy & Gitomer, 1996; Mislevy, Steinberg, & Almond, 2003). However, there is a severe lack of well-established diagnostic tools (Williamson, Almond, & Mislevy, 2000) for these models. The standard χ^2-type test does not apply, except for assessments with a small number of items.

This paper explores a number of approaches to assess the fit of models with student proficiency consisting of discrete variables, particularly those in which the distribution of the proficiency variables can be described using a Bayesian network. The paper then applies these techniques to a data set from Tatsuoka's (1990) research on mixed number subtraction with middle school students. This work is an extension of the recent work on model checking by Yan, Mislevy, & Almond (2003) on this data set. Appropriately created Bayesian residual plots help us to improve upon a simple model with discrete proficiency variables fit to the data set by Mislevy (1995). This paper suggests an "item fit plot," an equivalent of the standard item characteristic curve (ICC) applicable to models with discrete proficiency variables, and an attached χ^2-type test statistic. These plots and the test statistics detect a number of problems with the model and flag two problematic items in the test. The posterior predictive model checking method (Rubin, 1984; Gelman, Meng, & Stern, 1996) is also applied to the model, but the discrepancy measures (which are equivalent to the classical "test statistics") used with the method do not seem to have enough power to detect item fit or overall model fit in this example. However, the measures appear to be promising in diagnosing person misfits.

The next section begins with a description of the mixed number subtraction problem, which motivates this work and will be used throughout the paper. The section then reviews the Almond and Mislevy (1999) framework for educational testing models and the particular use of Bayesian networks to model student proficiencies and item outcomes. The section introduces a data set from Tatsuoka's (1990) work and then describes a specific model with discrete proficiency variables, called the two-parameter latent class model or "2LC" model hereafter, fit to the data set by Mislevy (1995). Finally, it gives a brief overview of Bayesian analysis and the Markov chain Monte Carlo (MCMC) algorithm, which is used for fitting all the models in this work. Section 3 reviews a number of approaches to model checking for these or related models. Section 4 starts by providing a summary of the method for fitting the 2LC model to the data set described in Section 2 and the results obtained. The section then applies a number of diagnostic procedures, including item fit plots and related χ^2-type test statistics, to the 2LC model. The diagnostics indicate that the 2LC model is inadequate to explain the variability in the data set. Section 5 introduces three new models, all involving discrete proficiency variables, that are possible improvements to the 2LC model, and applies the model diagnostics discussed earlier to those models. An extended version of the 2LC model, referred to as the "3LC" model here, seems to explain the data set satisfactorily. Section 6 discusses the performance of the diagnostics and makes recommendations for both practical application and future research.

2. Background

This section reviews some background material and introduces the small example assessment (based on mixed number subtraction), which we will analyze in later sections.

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 rule space method of Tatsuoka (1983). Rule space 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 rule space method, we introduce what will be a running example used through the paper—one regarding a test on mixed number subtraction.

This example is grounded in a cognitive analysis of middle school students' solutions of mixed-number subtraction problems. Klein, Birnbaum, Standiford, & Tatsuoka (1981) identify two methods of solution for these problems:

- Method A: Convert mixed numbers to improper fractions, subtract, then reduce if necessary.
- Method B: Separate mixed numbers into whole number and fractional parts, subtract as two subproblems, borrowing one from the whole-number minuend if necessary, then simplify and reduce if necessary.

We focus on students learning to use Method B (giving us 325 students). 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 are 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 remaining 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, there are no students who have Skill 4 but not 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 the data set collected by Tatsuoka (1990), characterized by the skills they require. The part of the table marked "Skills required" represents the Q-matrix.

Table 1.

Skill Requirements for the Mixed Number Subtraction Problems

Item	Text of	Skills required					Evidence
no.	the item	1	2	3	4	5	model
2	$\frac{6}{7} - \frac{4}{7}$	x					1
4	$\frac{3}{4} - \frac{3}{4}$	x					1
8	$\frac{11}{8} - \frac{1}{8}$	x	x				2
9	$3\frac{4}{5} - 3\frac{2}{5}$	x		x			3
11	$4\frac{5}{7} - 1\frac{4}{7}$	x		x			3
5	$3\frac{7}{8} - 2$	x		x			3
1	$3\frac{1}{2} - 2\frac{3}{2}$	x		x	x		4
7	$4\frac{1}{3} - 2\frac{4}{3}$	x		x	x		4
12	$7\frac{3}{5} - \frac{4}{5}$	x		x	x		4
15	$4\frac{1}{3} - 1\frac{5}{3}$	x		x	x		4
13	$4\frac{1}{10} - 2\frac{8}{10}$	x		x	x		4
10	$2 - \frac{1}{3}$	x		x	x	x	5
3	$3 - 2\frac{1}{5}$	x		x	x	x	5
14	$7 - 1\frac{4}{3}$	x		x	x	x	5
6	$4\frac{4}{12} - 2\frac{7}{12}$	x	x	x	x		6

A number of features of this data set can be learned 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 (Mislevy et al., 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 (even assuming no chance errors). For example, because every item requires Skill 1, the 12 profiles that lack Skill 1 are indistinguishable on the basis of this data. Similar logic reveals that there are only nine *equivalence classes* of student profiles. Table 2 describes the classes by relating them to the evidence models.

Table 2.

Skill Combinations for Each 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

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.

Tatsuoka (1990) analyzes this data set using her rule-space methodology (Tatsuoka, 1983), which uses a pattern-matching approach to handle uncertainty. Mislevy (1995) later reanalyzed the data using Bayesian networks (Section 2). Instead, we will develop a more formal item response (IR) model for this problem.

Evidence-centered Design Framework

Mislevy (1995) recasts the mixed number subtraction example as an IR model. Unlike the rule space model where the error model is implicit in distance metric used for matching, we explicitly model the probability that a participant with attributes $\boldsymbol{\theta}_i = \{\theta_{i1}, \ldots, \theta_{iK}\}$ will get a response vector $\mathbf{X}_i$. Using an explicit error model makes it easy to check how well it models the data and possibly suggests improvements to the model on the basis of diagnostic statistics and graphs. Thus, we will follow this formulation of the example throughout the paper.

Almond and Mislevy (1999) lay out a general formulation for educational testing models, which form the basis of evidence-centered design (Mislevy et al., 2003). The model starts by postulating a number of *proficiency variables*, $\boldsymbol{\theta}_i = \{\theta_{i1}, \ldots, \theta_{iK}\}$. These are latent variables describing knowledge, skills, and abilities of the participant we wish to draw inferences about. (Note that these are sometimes referred to as *person parameters* in the IRT literature. However, as there is no difference in Bayesian statistics between unknown parameters and latent variables, 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.

Let $\mathbf{X}_{ij}$ denote the scored outcome of the i-th participant to the j-th task. Note that in general this can be a vector valued quantity; that is, a single "task" could consist of multiple "items." Note also that these outcomes are *scored*, which implies some level of processing from the raw response. In our example, each item produces a single dichotomous outcome that is 1 if the response is correct and 0 if it is incorrect. One could imagine other ways of processing the responses in this situation, for example, providing two outcomes: one for whether the response was correct or not and one for whether the response was reduced to the simplest form. Such a scheme might be more useful for producing diagnostic feedback, but is not further considered here.

Next, we make two critical assumptions. The first is that all participants are statistically independent. The second is that given the proficiency variables, the observed outcomes for different tasks are independent. (In the case of several items that were dependent, for example a reading testlet, we would group them into a single "task" to make them independent of other tasks). Under this assumption, the joint distribution of the proficiency variables and all of the outcome variables is:

$$\prod_{i=1}^{I} P(\boldsymbol{\theta}_i|\boldsymbol{\lambda}) \prod_{j=1}^{J} P_j(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\pi}_j),$$

where $\boldsymbol{\pi}_j$ is the parameters of the distribution of X_{ij} given $\boldsymbol{\theta}_i$ and $\boldsymbol{\lambda}$ is the parameters of $P(\boldsymbol{\theta}_i)$. We call the term $P_j(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\pi}_j)$ the *link model* because it provides a link between the latent proficiency variables and the observable outcomes.

If $P(\boldsymbol{\theta}_i)$ and $P_j(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\pi}_j)$ are known, then it is simple to compute the probability of a proficiency profile for the participant. Applying Bayes theorem, we find $P(\boldsymbol{\theta}_i, \boldsymbol{\pi}_j|\mathbf{X}_i)$ and make diagnostic recommendations on the basis of this distribution.

In the usual situation, however, $P(\boldsymbol{\theta}_i)$ and $P_j(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\pi}_j)$ are only known up to the values of certain parameters, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$. We make two assumptions about the independence of the parameters. We assume that a priori $\boldsymbol{\lambda}$ is independent of $\boldsymbol{\pi}_j$, and $\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_{j'}$ are independent for $j \neq j'$. In this case, our model becomes:

$$\left(\prod_{i=1}^{I} P(\boldsymbol{\theta}_i|\boldsymbol{\lambda}) \prod_{j=1}^{J} P_j(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\pi}_j) \right) \; P(\boldsymbol{\lambda}) \prod_{j=1}^{J} P_j(\boldsymbol{\pi}_j) \tag{1}$$

Figure 1 shows this model graphically.

Note that while the link model for each item differs in its parameters, the functional form is often the same. Table 1 identified six *evidence models* in the mixed number subtraction data. The functional form differs across evidence models (each will rely on a different number of θ_{ik}s); however, it will be the same within a given evidence model. It should be possible to exploit the evidence model structures to build hierarchical models for the parameters, but we have not done this here.

Mixed Number Subtraction Proficiency Model

The Mislevy (1995) model for the mixed number subtraction problem follows exactly this framework laid out above. It starts with five proficiency variables, $\{\theta_{i1}, \ldots, \theta_{i5}\}$,

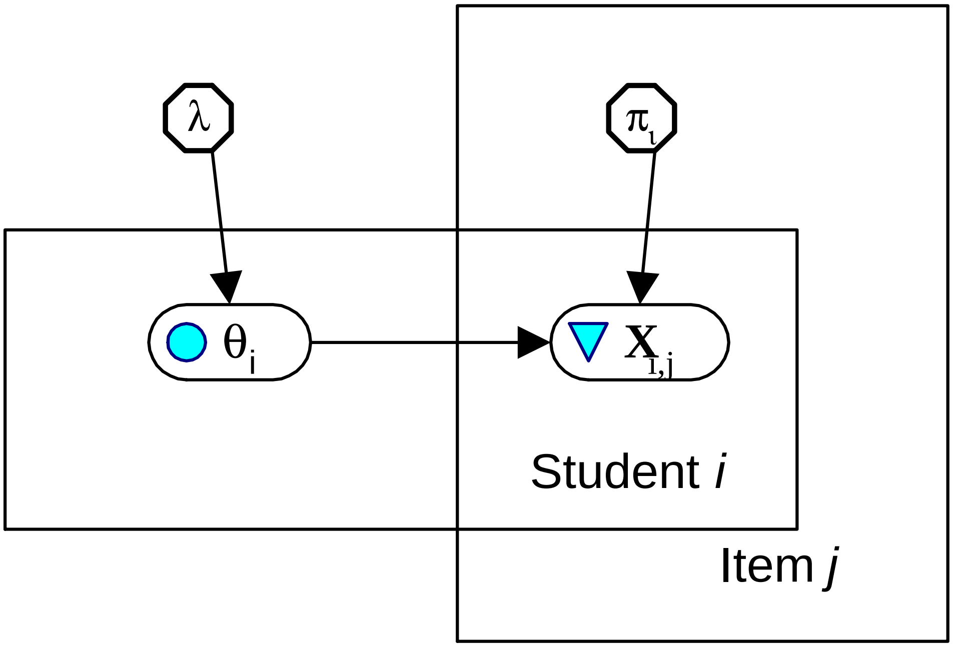


***Figure* 1. The graphical representation of Equation 1.**

corresponding to the five skills identified above. Each of these is an indicator variable, which takes on the value 1 if the participant has mastered the skill and the value 0 otherwise. The prior (population) distribution $P(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is expressed as a discrete Bayesian network or graphical model (Pearl, 1988; Lauritzen & Spiegelhalter, 1988). The Bayesian network uses a graph to specify the factorization of the joint probability distribution over the skills. Note that the Bayesian network entails certain conditional probability conditions, which we can exploit when developing the Gibbs sampler for this model. Figure 2 shows the dependence relationships among the skill parameters 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

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

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. The relationship between θ_{WN} and θ_3 and θ_4 are logical rather than probabilistic; but they can be represented with probability tables with 1s and 0s.

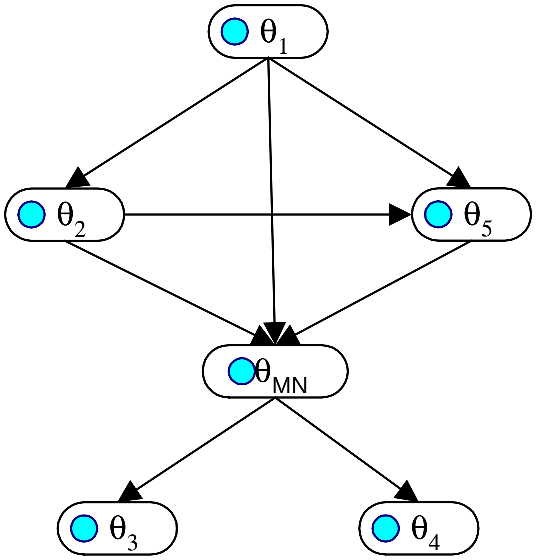


***Figure* 2. The graphical representation of the student model for mixed number subtraction example,** $p(\boldsymbol{\theta}) = p(\theta_3|\theta_{WN})p(\theta_4|\theta_{WN})p(\theta_{WN}|\theta_1,\theta_2,\theta_5)p(\theta_5|\theta_1,\theta_2)p(\theta_2|\theta_1)p(\theta_1)$.

The parameters $\boldsymbol{\lambda}$ of the graphical model 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, we require prior distributions $P(\boldsymbol{\lambda})$. We assume that λ_1, $\boldsymbol{\lambda}_2$, $\boldsymbol{\lambda}_5$, and $\boldsymbol{\lambda}_{WN}$ are a priori independent. They will be a posteriori dependent because the $\boldsymbol{\theta}$ variables are latent (Madigan & York, 1991). However, the MCMC analysis will take that dependence into account.

The natural conjugate priors for the components of $\boldsymbol{\lambda}$ are either beta or Dirichlet distributions. In all cases, we chose the hyper-parameters 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. Here, we have given numbers that correspond to 87% for acquiring a skill when the previous skills are mastered and 13% for acquiring the same skill when the previous skills are not mastered. They are as follows:

$$\begin{aligned}
\lambda_1 &\sim \text{Beta}(23.5, 3.5) \\
\lambda_{2,0} &\sim \text{Beta}(3.5, 23.5) \\
\lambda_{2,1} &\sim \text{Beta}(23.5, 3.5) \\
\lambda_{5,0} &\sim \text{Beta}(3.5, 23.5) \\
\lambda_{5,1} &\sim \text{Beta}(13.5, 13.5) \\
\lambda_{5,2} &\sim \text{Beta}(23.5, 3.5)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\lambda}_{WN,0,\cdot} &= (\boldsymbol{\lambda}_{WN,0,0}, \boldsymbol{\lambda}_{WN,0,1}, \boldsymbol{\lambda}_{WN,0,2}) \sim \text{Dirichlet}(15, 7, 5) \\
\boldsymbol{\lambda}_{WN,1,\cdot} &\sim \text{Dirichlet}(11, 9, 7) \\
\boldsymbol{\lambda}_{WN,2,\cdot} &\sim \text{Dirichlet}(7, 9, 11) \\
\boldsymbol{\lambda}_{WN,3,\cdot} &\sim \text{Dirichlet}(5, 7, 15)
\end{aligned}$$

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 $\boldsymbol{\theta}$ that have non-zero 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 over all 24 possible latent classes, only 9 of them are identifiable from the data (Table 2). This property of the test design will manifest itself later in the analysis.

Mixed Number Subtraction Link Models

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 (or item from an evidence model), we say that student has mastered the item (evidence model). In general, students will not behave according to the ideal model; we will get false positive and false negative results.

We now build a series of link models, which follow that intuition. The 2LC model uses two parameters per link model: the true positive and false positive probabilities. That is:

$$P(X_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\pi}_j) = \begin{cases} \pi_{j1} \text{ if Examinee } i \text{ mastered all the skills needed to solve Item } j, \\ \pi_{j0} \text{ otherwise.} \end{cases} \tag{2}$$

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. 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 taken as

$$X_{ij}|\pi_{j\delta_{i(s)}}, \boldsymbol{\theta}_i \sim \text{Bernoulli}(\pi_{j\delta_{i(s)}}) \tag{3}$$

The local independence assumption is made; that is, given the proficiency $\boldsymbol{\theta}_i$, the response of an examinee to the difference items are assumed independent. The probability π_{j1} represents a "true-positive" probability for the item; that 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), both within and across evidence models. However, we use the same priors for all items:

$$\begin{aligned} \pi_{j0} &\sim Beta(3.5, 23.5) \\ \pi_{j1} &\sim Beta(23.5, 3.5) \end{aligned} \tag{4}$$

This model is very similar to the "noisy-and" model discussed in Pearl (1988) and Junker and Sijtsma (2000) and also to the fusion model of Hartz, Roussos, and Stout (2002). However, both the noisy-and and fusion model include additional terms for modeling the effect of missing each of the individual skills. Thus, their models are somewhat softer than the mastered/not-mastered approach of the 2LC model described here. The fusion model also includes an additional continuous proficiency model variable for proficiency to apply the skills to solve the problem. In the later sections we will soften the 2LC model in several ways.

Bayesian Analysis and Markov Chain Monte Carlo Algorithm

Although a substantial amount of prior information about the values of the parameters of the 2LC model ($\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$) is available, we still would like to refine that knowledge from data. In particular, our interest is to know about their posterior distributions for modeled parameters based on the observed test outcomes. A knowledge of the whole distribution is essential in order to be able to properly criticize the model. However, simply applying the Bayes theorem to learn about the posterior distribution (which is the first step in a typical Bayesian analysis) leaves us with an integral, which is impossible to compute analytically. The fact that all of the proficiency variables are latent and hence missing causes many of the convenient independence properties of our model to disappear.

Imputing values for the latent variables and parameters allows us to exploit those independence conditions. In particular, the MCMC simulation repeatedly samples from the distributions of the latent variables and parameters using a Markov chain whose stationary distribution is equal to the posterior distribution. As long as the Markov process is run long enough so that the distribution of the draws is close enough to the stationary distribution, we can calculate quantities of interest using Monte Carlo integration with this sample.

The Gibbs sampler and the Metropolis–Hastings algorithm (see, e.g., Gelman et al., 1995) are two of the most common MCMC algorithms. A number of books, such as Gelman et al. (1995), give a detailed discussion of the MCMC methods. We use the BUGS software (Spiegelhalter, Thomas, Best, & Gilks, 1995), which builds a Gibbs sampler (or, if necessary, a Metropolis–Hastings algorithm) based on the description of the problem (implicitly the model graph). The Gibbs sampling algorithm can exploit the conditional independence implicit in the graphs in Figures 1 and 2. Mislevy, Senturk, et al. (2001) and Yan et al. (2003) describe their applications using the algorithm for the mixed number subtraction example.

3. Diagnostics for Models With Discrete Proficiency Variables

Although the model described in the previous section has a quite complex latent structure, it is also a good reflection of the cognitive theory of the domain as expressed by Klein et al. (1981). Therefore, studying model fit and attempting model improvement will help us refine not only the measurement properties of the model but also the underlying cognitive model.

In evaluating the fit of educational assessment models, an investigator can look at three kinds of model fit tests:

- tests of global fit, indicating the overall fit of data to the model
- tests of item fit, identifying assessment items that the model does not predict well
- tests of person fit, identifying participants whose response patterns are not predicted well by the model.

Many diagnostic tests are based on analysis of residuals—the difference between the predicted and observed values. We perform Bayesian model fitting in this work—so the standard residual plots do not apply directly. This section first discusses how a Bayesian version of the classical residual plot (Chaloner & Brant, 1988) may be useful for our problem. Then this section reviews a number of plotting techniques aimed at analyzing item performance. Because we fit the model using an MCMC algorithm, we have access to the generated values of all the parameters and variables of the model. As an obvious outcome, we can apply some additional model fit techniques. One of them is the posterior predictive model checking method, which is also discussed in this section.

Bayesian Residual Analysis

In linear models, analysis of residuals has proved to be a robust tool for diagnosing problems with the model fit. Because the model described above is a full Bayesian model, we use the Bayesian approach to residual analysis (Chaloner & Brant, 1988). Suppose X_i denotes the observation for individual i. Suppose further that $E(X_i|\boldsymbol{\omega}) = E_i$, where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots \omega_M)$ denotes the vector of all parameters in the model. Consider the realized residual $\epsilon_i = X_i - E_i$. After data has been collected and a Bayesian model has been fitted, X_i is considered outlying if the posterior distribution for the residual ϵ_i is located far from zero (Chaloner & Brant, 1988).

For the mixed number subtraction data, all the individual observations are binary. The residuals from binary response models are difficult to define and interpret (Albert & Chib, 1995), mainly because the distribution of an individual observation is far from the normal distribution. Therefore, in this work, we look at residuals after some pooling, which creates more meaningful and more stable residuals.

Let O_i denote the raw score (or number-correct score) for Examinee i. The raw score is a very natural quantity to examine in the analysis of any test data—the classical test theory revolves around the raw score and the raw score plays an important role in the IRT as well (e.g., van der Linden and Hambleton, 1997). The mixed number subtraction link model in (2) provides the basis of computing the expectation of the raw score of an examinee conditional on the parameters of the model, resulting in an expression for the realized residuals ϵ_i. Examining the posterior distribution of the realized residuals may provide some insights about the fit of the model. The idea is described in much more details in the next section, where we apply the idea to the mixed number data.

Item Fit Plots

For IRT models, one way to detect lack of item fit is to compare the average item performance levels of various proficiency groups to the performance levels predicted by the model (see, for example, Hambleton and Swaminathan, 1985 and Hambleton, 1989). The comparison is made mostly by plotting an item characteristic curve (ICC), which shows the observed vs. predicted proportion correct scores for the various proficiency groups; a χ^2-type test statistic is also used to make the comparison. Yen (1981) used groups based on the likelihood estimates of proficiency of the examinees while Orlando and Thissen (2000) formed the groups based on the raw scores of the examinees. An ICC is plotted for each item in a test to obtain the fit of that item. Too many misfitting items indicate a problem with the fit of the model to the data; on the other hand, very few misfitting items usually indicate that those particular items are outliers in the sense that they cannot be explained by the model.

When extending the IRT item fit ideas to models with discrete proficiency variables, the first task is to identify the groups comparable to the proficiency groups used in IRT. The equivalence classes of states of the proficiency variables (see Table 2) form natural groups, but they rely on the state of unobserved variables. However, because we use the MCMC algorithm to fit the model, there is a way to form the groups and hence to judge item fit. Looking at the draws of the proficiency variables for the individuals in each iteration of the Markov chain, we can classify the individuals into different groups according to different combinations of the proficiency variable values. By comparing the observed proportion correct for an item to the expected proportion correct for each group of individuals, we may have an idea about the fit of the item. As with IRT models, this can be done graphically or by using a χ^2-type test statistic.

The measure may have low power as it depends on unobserved quantities, but provides us with one way to assess item fit, and is proved to be useful in the real data example later.

Posterior Predictive Model Checking

Let $\boldsymbol{y}$ represent all of the observed variables in our model and let $\boldsymbol{\omega}$ represent all of the parameters and unobserved (proficiency) variables. Let $\boldsymbol{y}^{rep}$ denote replicate data that we might observe if the experiment that generated $\boldsymbol{y}$ is replicated with the same value of $\boldsymbol{\omega}$ that generated the observed data. Since the value $\boldsymbol{\omega}$ that generated the observed data is unknown, we derive the posterior (given $\boldsymbol{y}$) predictive distribution of $\boldsymbol{y}^{rep}$ by averaging over the plausible values of $\boldsymbol{\omega}$, given by the posterior distribution $p(\boldsymbol{\omega}|\boldsymbol{y})$,

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

Guttman (1967) applies the posterior predictive distribution in a goodness-of-fit test. Rubin (1984) suggests simulating replicate data sets from the posterior predictive distribution for model checking. Any significant difference between the replications and the observed data indicates a possible failure of the model.

In practice, for a given diagnostic measure, $D(\boldsymbol{y})$, any significant difference between the observed value $D(\boldsymbol{y})$ and the reference distribution of $D(\boldsymbol{y}^{rep})$ indicates a possible model failure. Gelman et al. (1996) extend the posterior predictive approach to use diagnostic measures $D(\boldsymbol{y}, \boldsymbol{\omega})$ that depend on the data and the parameters. The divergence of the data from the posterior predictive distribution can be determined by comparing the posterior predictive distribution of $D(\boldsymbol{y}^{rep}, \boldsymbol{\omega})$ with the posterior distribution of $D(\boldsymbol{y}, \boldsymbol{\omega})$. The comparison can be carried out easily by simulation. We draw N simulations $\boldsymbol{\omega}^1$, $\boldsymbol{\omega}^2$,...,$\boldsymbol{\omega}^N$ from the posterior distribution of $\boldsymbol{\omega}$, and then draw one $\boldsymbol{y}^{rep}$ from the predictive distribution $p(\boldsymbol{y} \mid \boldsymbol{\omega})$ using each simulated $\boldsymbol{\omega}$. We then have N draws from the joint posterior distribution $p(\boldsymbol{y}^{rep}, \boldsymbol{\omega} \mid \boldsymbol{y})$. The posterior predictive check boils down to comparing the values of the realized discrepancy $D(\boldsymbol{y}, \boldsymbol{\omega}^n)$ and the replicated discrepancy measures $D(\boldsymbol{y}^{rep,n}, \boldsymbol{\omega}^n)$, $n = 1, 2, \ldots N$, perhaps by plotting the pairs $\big(D(\boldsymbol{y}, \boldsymbol{\omega}^n), D(\boldsymbol{y}^{rep,n}, \boldsymbol{\omega}^n)\big)$ in a scatter-plot. One popular summary of the comparison is the tail-area probability or Bayesian p-value,

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

where $I_{[A]}$ is the indicator function for the event A. The p-value is estimated from the simulations as the proportion of the N replications for which $D(\boldsymbol{y}, \boldsymbol{\omega}^n) \geq D(\boldsymbol{y}^{rep,n}, \boldsymbol{\omega}^n)$. Very extreme posterior predictive p-values (close to 0 or 1) indicate model misfit.

Posterior predictive checks have been criticized for being conservative (see for example, Bayarri & Berger, 2000, and Sinharay & Stern, 2003). Still, they are easy to carry out and interpret. They are especially useful if we think of the current model as a plausible ending point with modifications to be made only if substantial lack of fit is found. Successful applications of the technique in psychometrics include Johnson, Cohen, and Junker (1999), Hoijtink and Molenaar (1997), Mislevy, Senturk, et al. (2001), and Sinharay and Johnson (2003) and the references therein.

For this work, we use as discrepancy measures the sum of squares of standardized residuals over the persons (to detect item fit) or over the items (to detect person fit) and proportion correct for items and persons.

4. The Diagnostics Applied to the 2LC Model

We apply the different diagnostics discussed in Section 3 to the mixed number subtraction example in an attempt to find out whether the 2LC model adequately explains the variability in the data set. The first part of this section describes briefly about fitting the 2LC model using an MCMC algorithm. We then examine Bayesian residual plots for the 2LC model. What follows is an analog of ICC plots for assessing item fit. We then attempt to build a fit statistic to identify the items with problems. Then we look at the posterior predictive model check statistics. Finally, we summarize what the diagnostics applied tell us about the 2LC model.

Fitting the Model Using MCMC Algorithm

Mislevy, Almond, et al. (2001) describe fitting the 2LC model to the data set using the MCMC algorithm. The joint posterior distribution of the parameters of the model given the data $\boldsymbol{X}$ is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\pi} | \boldsymbol{X}) \propto \left\{ \prod_i \prod_j p(X_{ij} | \boldsymbol{\theta}_i, \boldsymbol{\pi}_j) \right\} \left\{ \prod_i p(\boldsymbol{\theta}_i | \boldsymbol{\lambda}) \right\} \left\{ \prod_j p(\boldsymbol{\pi}_j) \right\} p(\boldsymbol{\lambda}).$$

As mentioned before, we use the BUGS program (Spiegelhalter et al., 1995) to run the MCMC algorithm that fits the 2LC model to the data.

The BUGS program is used to generate five chains of size 3,000 with dispersed starting values. Looking at the plot of the Gelman-Rubin diagnostic measure (e.g., Gelman et al., 1995), we find that convergence is achieved within a few hundred iterations. We retain the last 2,000 values in each chain to obtain a total posterior sample of size 10,000. The posterior summary, the posterior mean, sd, and three quantiles (2.5%, median, and 97.5%) of a few parameters are given in Table 3.

Table 3.

Posterior Summary of a Few Parameters for the 2LC Model

Parameter	Interpretation of	Mean	sd	Quantiles		
	the parameter			2.5%	Median	97.5%
λ_1	$P(\theta_1 = 1)$	0.82	0.02	0.78	0.83	0.87
λ_{20}	$P(\theta_2 = 1 \vert \theta_1 = 0)$	0.13	0.06	0.03	0.12	0.26
λ_{21}	$P(\theta_2 = 1 \vert \theta_1 = 1)$	0.91	0.03	0.85	0.91	0.96
$\pi_{1,0}$	$P(X_1 = 1 \vert \delta_{i(4)} = 0)$	0.08	0.02	0.04	0.08	0.12
$\pi_{1,1}$	$P(X_1 = 1 \vert \delta_{i(4)} = 1)$	0.88	0.03	0.82	0.88	0.93
$\pi_{4,0}$	$P(X_4 = 1 \vert \delta_{i(1)} = 0)$	0.32	0.05	0.21	0.32	0.42
$\pi_{4,1}$	$P(X_4 = 1 \vert \delta_{i(1)} = 1)$	0.78	0.02	0.73	0.78	0.83
$\pi_{15,0}$	$P(X_{15} = 1 \vert \delta_{i(4)} = 0)$	0.03	0.01	0.01	0.03	0.05
$\pi_{15,1}$	$P(X_{15} = 1 \vert \delta_{i(4)} = 1)$	0.82	0.03	0.76	0.83	0.89

The posterior mean of 0.82 for λ_1 suggests that, on an average, 82% students have the Skill 1 (basic fraction subtraction). Note that this is a key skill in the sense that all the 15 items in this test require the presence of this skill to solve the item successfully. We also notice that of those who do not have Skill 1, about 13% have Skill 2; and approximately 91% of those having Skill 1 have Skill 2. Note that among the λs shown in the table, the posterior sd is largest for λ_{20} indicating the presence of the least number of students with this particular combination. A closer look reveals that the posterior distribution of λ_{20} is very close to the $Beta(3.5, 23.5)$ prior distribution (Note that the mean and sd of the prior distribution are 0.13 and 0.06, the same as the corresponding posterior quantities). This is because there are no tasks in the test that assess the presence of Skill 2 in the absence of Skill 1 (this can be verified in Table 1), making it impossible to distinguish between those two sets of latent classes. This should not make a practical difference, as once we have assessed that a student lacks Skill 1, basic fraction subtraction, we would assign remedial exercises to address this lack and then reassess for the presence of Skill 2.

From the summary of the πs, we see, for example, that for the individuals who do not have the necessary skills for solving Item 4 (which requires Skill 1 only), the chance of getting the item correct is about 32%, whereas someone having Skill 1 will have a 78% chance of solving that item correctly.

Bayesian Residual Plots

As mentioned earlier, this paper examines residuals based on the number correct scores of the examinees. Let O_i denote the observed number correct score of Examinee i. For this data set, O_i will range from 0 to 15.

Suppose we know the parameters of the model. Using $\theta_1, \theta_2, \ldots \theta_5$, the values of the proficiency variables, it is possible to compute the expected number correct score of Examinee i as

$$E(O_i|\boldsymbol{\theta}_i, \boldsymbol{\pi}, \boldsymbol{\lambda}) \equiv E_i = \sum_{j \in \text{items}} \pi_{j\delta_{i(s)}}, \tag{5}$$

where $\delta_{i(s)}$ is the indicator for mastery of the skills in the Evidence Model s for Item j. The value of $\delta_{i(s)}$ is determined by the values of $\theta_1, \theta_2, \ldots \theta_5$ for a particular examinee. Then, as in Chaloner and Brant (1988), define the realized residual for Examinee i as

$$R_i = O_i - E_i \cdot$$

An examination of the posterior distribution of R_is may be a useful tool in this context. Although we don't know the values of the parameters or of the latent variables, we have the draws from the posterior distribution (obtained by the MCMC algorithm) of the parameters given the data. Therefore, we can compute, for each examinee, values of R_i for each iteration of the MCMC algorithm, and then, a 95% posterior credible interval for R_is (formed by the 2.5th and 97.5th percentiles).

For each examinee, a vertical line in Figure 3 plots the 95% posterior credible interval for the residual (R_i). The horizontal axis corresponds to the posterior mean of the corresponding E_i (the latter is like a predicted/fitted value for the raw score for Examinee

i) for the 2LC model. We jitter the latter quantities to avoid having too many overlapping points. The dots show the posterior mean of the R_is. The horizontal line in the middle of the plot shows the 0-line, that is, the line for R_i=0.

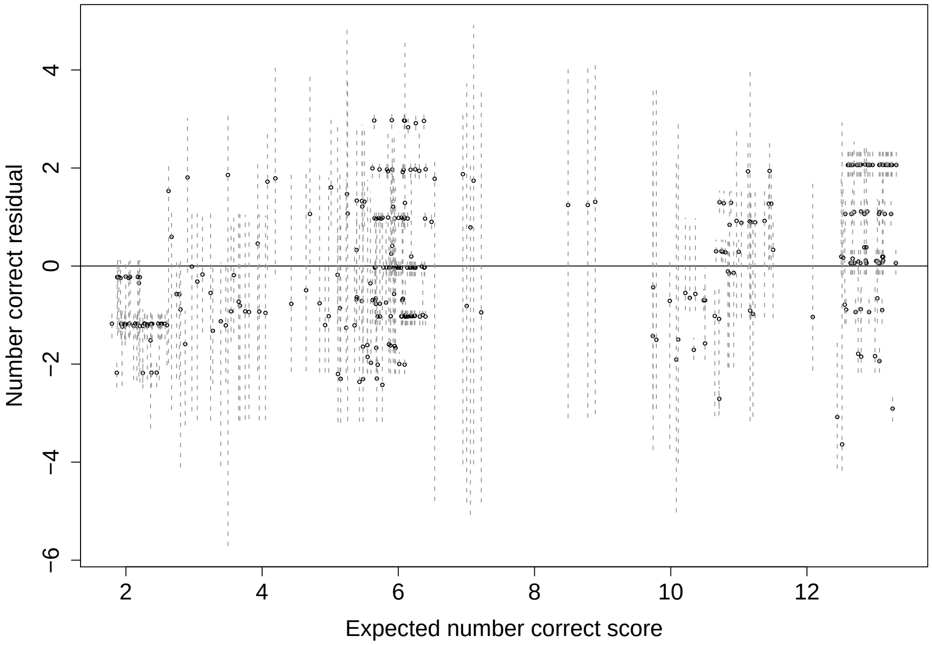


***Figure* 3. Plot of the posterior distributions of the number correct score residuals vs. the predicted number correct scores for the 2LC model.**

Figure 3 indicates a potential problem with the 2LC model. Note that towards the left of the plot (i.e., for low estimated expected scores), the residuals are mostly distributed below the 0-line. Towards the right side (i.e., for high estimated expected scores), the residuals are mostly distributed above the 0-line. These plots then suggest that the 2LC model over-predicts the scores of the individuals who have low proficiency to solve mixed number subtraction problems; on the other hand, the model under-predicts the scores of the individuals who have high proficiency. Clearly this indicates that the 2LC model does not explain the data adequately—it seems to pull the number correct score towards the middle.

Another approach to looking at overall fit is to plot the distribution of the residual scores vs. some measure of overall proficiency. Intuitively, the more skills a participant has mastered, the more items the participant should be able to solve correctly. The θ-score of an individual, $\sum_{k=1}^{5} \theta_{ik}$, represents the number of skills an examine has mastered. Although the values of $\boldsymbol{\theta}_i$ are unobservable, we have a number of imputed values $\boldsymbol{\theta}_i^t$ from each cycle of the MCMC loop. The posterior mean of these values over all the iterations (after an initial burn-in) of the MCMC gives us a point estimate of the θ-score of an examinee. We can use this estimated θ-score instead of the predicted score of the examinees to create a residual plot. Figure 4 shows a residual plot for the 2LC model created using this method.

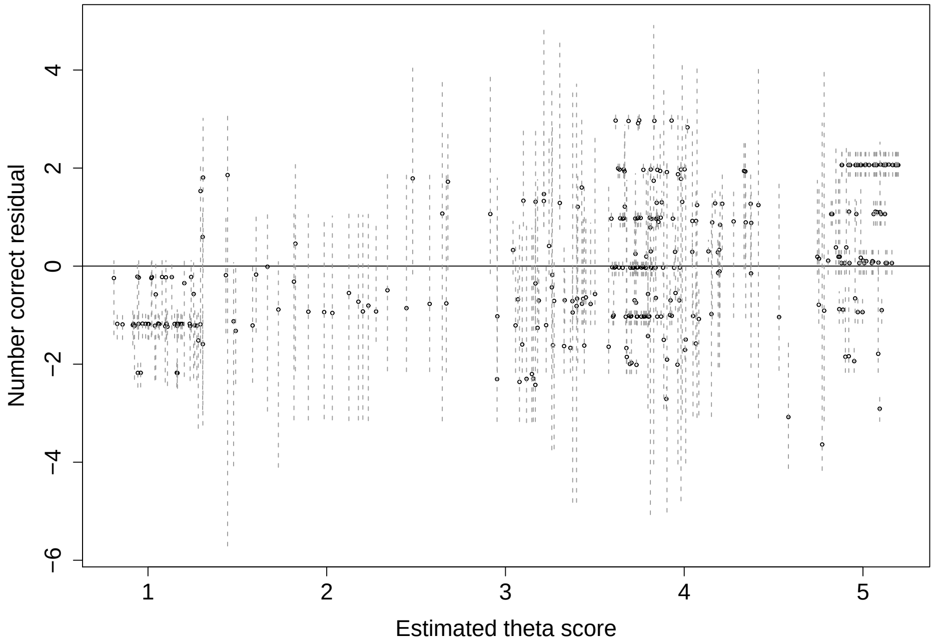


***Figure* 4. Plot of the posterior distributions of the number correct score residuals vs. the estimated θ-scores for the 2LC model.**

Figure 4 indicates the same problem as seen in Figure 3; in particular, it indicates a lack of fit. Participants who have mastered few skills are performing worse than predicted, and participants who have mastered all of the skills are performing better than expected. This may indicate a problem with the link models. Section 5 looks at some possible remedies.

A further refinement of the Bayesian residual plot described above may be achieved by using $E(O_i|\boldsymbol{\pi}, \boldsymbol{\lambda})$, taking the expectation of $E(O_i|\boldsymbol{\theta}_i, \boldsymbol{\pi}, \boldsymbol{\lambda})$ over $\boldsymbol{\theta}_i$, in (5) and examining residuals based on this refined expectation. The refined residuals will form the basis of a more powerful model diagnostic tool. However, for our purpose, residuals using $E(O_i|\boldsymbol{\theta}_i, \boldsymbol{\pi}, \boldsymbol{\lambda})$ allowed us to detect some problems with the 2LC model, and hence we do not pursue those based on $E(O_i|\boldsymbol{\pi}, \boldsymbol{\lambda})$.

Item Fit Plots

Before discussing possible remedies, we first explore in detail the item fit diagnostics introduced in Section 3. As mentioned before, the key idea will be to divide the examinees into different groups based on their proficiencies (so that different groups will have different success probabilities for an item) and then compare the observed proportion-correct scores against the predicted proportion-correct scores of the different groups.

The natural groups are 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.

Though the true class membership for any examinee is unobservable, we can classify that examinee on the basis of a set of imputed values of $\boldsymbol{\theta}^k$ from the MCMC algorithm. Yan et al. (2002) use groups based on a single iteration to search for item misfit. However, using only one iteration from the MCMC ignores the posterior variability of the model parameters and class assignments. We extend the plots to average over all the iterations of the MCMC.

In each iteration of the MCMC algorithm, we calculate, for each examinee, a vector of indicators of membership in each of the nine equivalence classes. Averaging these indicators over cycles gives us the a series of vectors $\boldsymbol{\tau}_i = (\tau_{i1}, \tau_{i2}, \ldots \tau_{i9})$, where τ_{ik} represents the proportion of iterations in which Examinee i was assigned to Equivalence Class k. The vector $\boldsymbol{\tau}_i$ provides a probabilistic classification of the examinee into an equivalence class.

To determine the “observed” proportion correct for Item j for examinees in Equivalence Class k, we take the weighted average over examinees using the classification probabilities $\boldsymbol{\tau}_i$ as weights. Thus, the observed proportion correct $\widehat{p_{kj}}$ for the k-th equivalence class and the

j-th item is given by

$$\widehat{p_{kj}} = \frac{1}{\sum_i \tau_{ik}} \sum_i \tau_{ik} X_{ij}. \tag{6}$$

Note that $\widehat{p_{kj}}$ is not observed in the true sense of the term because we do not really observe, but rather estimate the τ_{ik}s. The predicted proportion correct for the k-th equivalence class and the j-th item is the posterior mean of the appropriate $\pi_{j\delta_{(k)(s)}}$ for that equivalence class and item combination. Here, $\delta_{(k)(s)}$ is an indicator that tells whether examinees in Equivalence Class k have mastered the skills necessary to solve items from Evidence Model s (where s depends on the Item j, see Table 1). For example, for examinees in Equivalence Class 1, $\delta_{(1)(s)} = 0$ for all s, and hence the predicted proportion correct score for any item will be the posterior mean of π_{j0}. If the model fits the data well, we can expect the observed proportion correct $\widehat{p_{kj}}$ to be close to the predicted proportion correct for all combinations of equivalence classes and items.

Comparing the $\widehat{p_{kj}}$s to the corresponding predicted values for all equivalence classes should provide information about how well the link model for Item j fits the data. Many large deviations indicate a problem with the model. We use a rough confidence interval based on a presumption of normality. Using the fact that the responses of different examinees to an item are independent, we obtain

$$\begin{aligned} V_{kj} = Var(\widehat{p_{kj}}|\boldsymbol{\theta}, \boldsymbol{\pi}) &= \frac{1}{(\sum_i \tau_{ik})^2} \sum_i \tau_{ik}^2 Var(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\pi}_j) \\ &= \frac{1}{(\sum_i \tau_{ik})^2} \sum_i \tau_{ik}^2 \pi_{j\delta_{i(s)}}(1 - \pi_{j\delta_{i(s)}}). \end{aligned}$$

We estimate the πs in the above quantity by their posterior means to obtain an estimated variance of $\widehat{p_{kj}}$. We make a normal approximation of the proportions to take $(\widehat{p_{kj}} \mp 2 \times V_{kj}^{1/2})$ as a rough 95% confidence interval. Because there are only nine equivalence classes and as many as 325 examinees, normal approximation of the proportion correct is not entirely unreasonable.

Figures 5 and 6 show the item fit plots for the 15 items. The horizontal lines in each plot are the posterior means of π_{j0} and π_{j1} for the items (which are the predicted proportion correct for the equivalence classes). The vertical lines constitute the rough 95% confidence

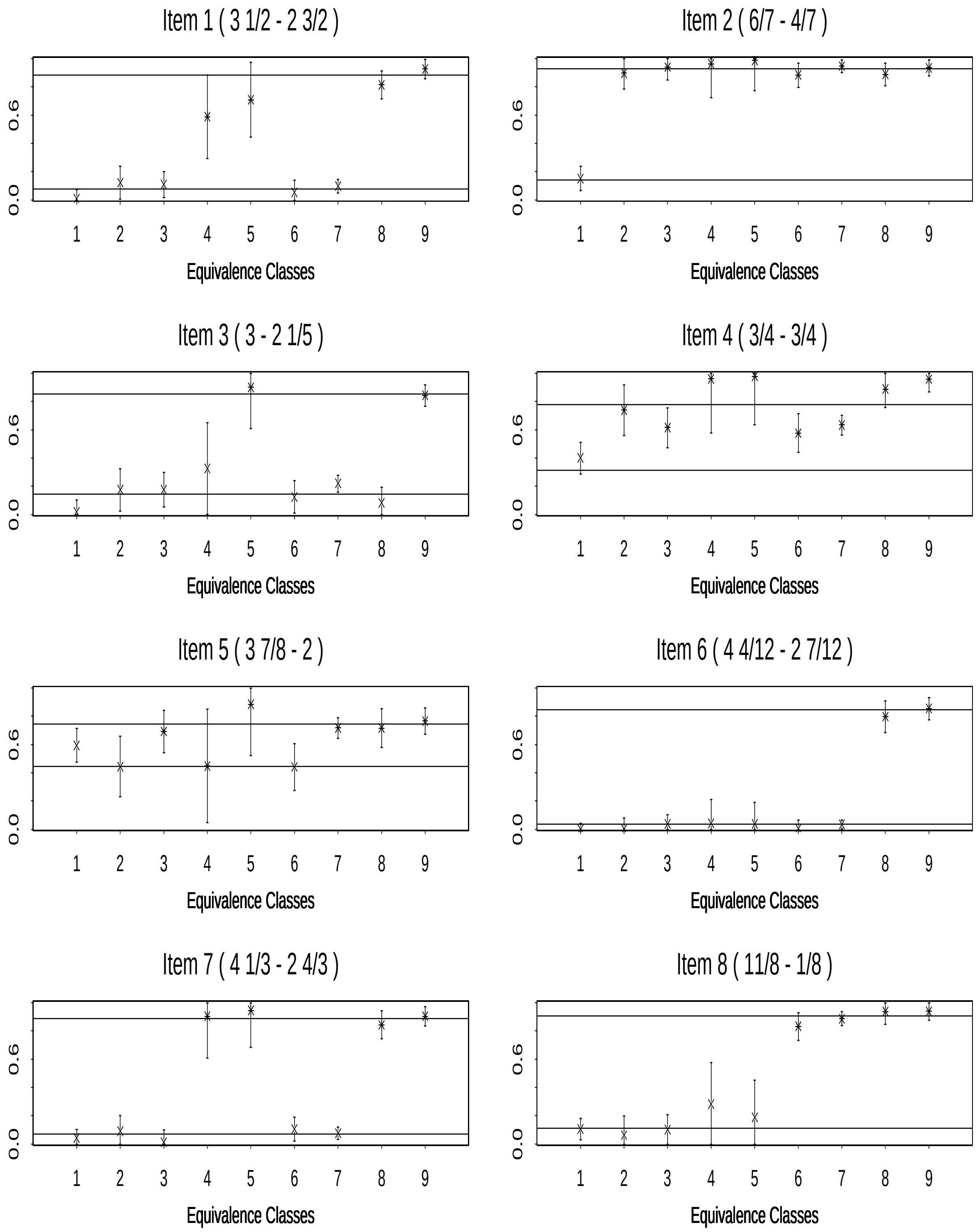


***Figure* 5. Item fit plots for Items 1–8 in the 2LC model.**

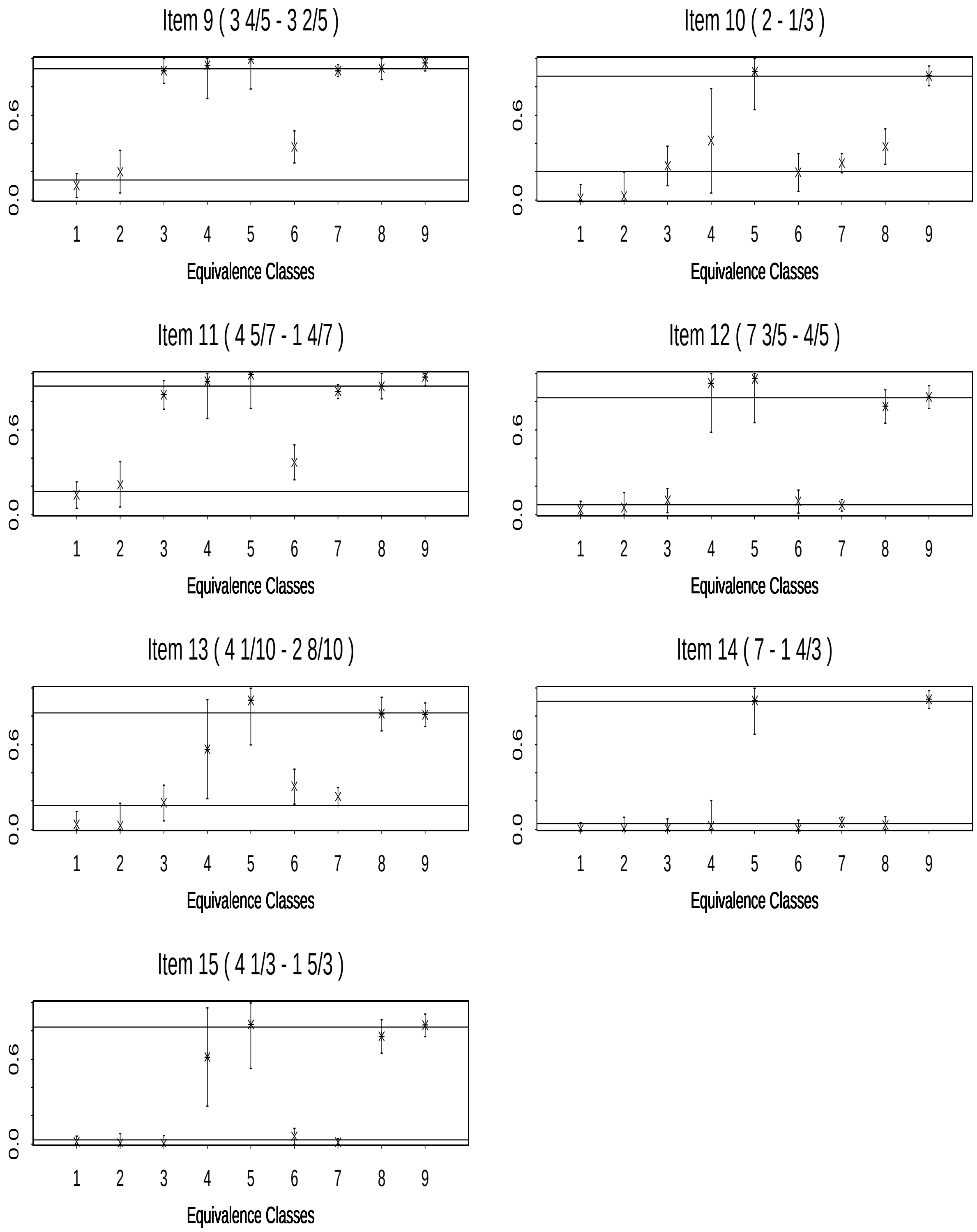


***Figure* 6. Item fit plots for Items 9–15 in the 2LC model.**

intervals attached to the observed proportion correct for the examinees in each equivalent class. The glyph in the middle of an interval represents the observed proportion correct for that equivalence class. The glyph depends on the value of $\delta_{(k)(s)}$; it is an asterisk (*) for equivalence classes that have mastered the skills necessary to solve the item, and the glyph is an x for equivalence classes lacking one or more skills. Confidence intervals having an x at the center, but not covering the lower horizontal line indicate possible misfit, as do confidence intervals having an asterisk at the center, but not covering the upper horizontal line.

The total weight of Equivalence Class k, $N_k = \sum_i \tau_{ik}$, plays roughly the same role as the sample size if we could observe the classes. Because the total weight is quite small for Classes 2 to 6 (ranging from about 2 to 11), a difference in the observed and predicted proportions for those classes could be due to sampling variability. For the other equivalence classes, such differences are more likely to imply a serious problem.

In Table 4, all the combinations of equivalence class and item for which there is a discrepancy between the observed and predicted proportions are marked with a "$\surd$." A look at the table suggests that there is a problem with about half the items in the data set. This indicates that the model cannot explain the data set adequately. Looking more closely, we see that many of the problems occur for Equivalence Class 1 (people who have not mastered any skills) and Equivalence Class 9 (people who have mastered all of the skills). This is consistent with the findings from the residual plots.

A Test Statistic to Detect Item Fit

To quantify the item fit plots discussed above, we define a test statistic. The quantity $\widehat{p_{kj}}$ is the estimated observed proportion correct for the j-th item for participants in the k-th equivalence class. So the observed number of examinees in the k-th equivalence class getting the item j correct, denoted O_{kj}, is given by

$$O_{kj} = \widehat{p_{kj}} \sum_i \tau_{ik} = \sum_i \tau_{ik} X_{ij}.$$

The predicted number of examinees in the k-th equivalence class getting the item j correct, denoted E_{kj}, is obtained by multiplying the predicted proportion correct score for that

Table 4.

Problematic Cases Detected by the Item Fit Plots

Item	Equivalence class									Value
no.	1	2	3	4	5	6	7	8	9	of χ_j^2
1	✓									10.7
2										2.2
3	✓						✓			15.0^a
4							✓		✓	39.5^b
5	✓									7.6
6										3.3
7										2.7
8										3.7
9						✓				8.5
10	✓							✓		24.1^b
11								✓	✓	10.6
12										2.8
13	✓									14.4^a
14										4.0
15										3.9

[a] Larger than 95 percentile of $\chi_7^2 \approx 14$.

[b] Larger than 99 percentile of $\chi_7^2 \approx 18.5$.

combination (which is the posterior mean of the suitable $\pi_{j\delta_{(k)(s)}}$) by $N_k = \sum_i \tau_{ik}$.

We now define χ_j^2, the item fit statistic for the j-th item, as

$$\begin{aligned} \chi_j^2 &= \sum_{k=1}^{9} \frac{(O_{kj} - E_{kj})^2}{E_{kj}} + \sum_{k=1}^{9} \frac{((N_k - O_{kj}) - (N_k - E_{kj}))^2}{N_k - E_{kj}} \\ &= \sum_{k=1}^{9} \frac{N_k (O_{kj} - E_{kj})^2}{E_{kj}(N_k - E_{kj})} \end{aligned} \tag{7}$$

Although the statistic is inspired by classical χ^2 tests, the reference distribution is unknown. If the membership of the examinees to the equivalence classes were known, the statistic χ_j^2 would follow a χ^2 distribution with 7 d.f. (because there are nine equivalence classes for any item, and we are estimating two πs for each item) under the null hypothesis that the 2LC model fits the data set. But the membership is estimated making the true null distribution difficult to compute. The same issue arises with the χ^2 statistics suggested by Hambleton and Traub, 1973, and Yen, 1981, in the context of IRT models. However, even though the null distribution of these statistics is unknown, we can compare the values of χ_j^2s computed for a number of competing models by keeping the number of groups (equivalence classes in this problem) constant to judge which model is preferable over the others (in the same vein as suggested by Hambleton & Traub, 1973, for comparing χ^2 statistics to detect model misfit for IRT models).

We use the χ_7^2 distribution as a rough reference to provide a heuristic to flag items that are not fit well by the model. The computed values of the χ_j^2 statistics along with them are given in the last column of Table 4. Values greater than the 95 and 99 percentiles of the χ_7^2 distribution are flagged. Note that the χ^2 statistics and items fit plots flag the same items. We see low values of χ_j^2s for the items with no $\surd$ for them in the table (e.g., Items 2, 6, 7, 8) while the values are high for Items 3, 4, 10, and 13 for which we see a few $\surd$ (i.e., discrepant cases, as we discussed earlier) in the table.

Posterior Predictive Model Checking

The posterior predictive model check diagnostics are based on comparing observed data X_{ij} to replicated data, X_{ij}^{rep}. As each iteration of the MCMC algorithm generates values for all of the parameters as well as the latent proficiency variables, the only additional work involved in generating the replicates is that of generating X_{ij}^{rep}s from the model (3) using the generated values. The diagnostic measures we use, $D(\boldsymbol{y}, \boldsymbol{\omega})$, are the proportion correct (for both items and examinees) and the mean squared Pearson residuals (taking the averages across items, examinees, and both).

Consider $T_{ij}(\boldsymbol{X}, \boldsymbol{\omega})$ given by

$$\begin{aligned} T_{ij}(\boldsymbol{X}, \boldsymbol{\omega}) &= \frac{(X_{ij} - E(X_{ij}))^2}{V(X_{ij})} \\ &= \frac{(X_{ij} - \pi_{j\delta_{i(s)}})^2}{\pi_{j\delta_{i(s)}}(1 - \pi_{j\delta_{i(s)}})} \ , \end{aligned}$$

where $\boldsymbol{\omega}$ represents the parameters and latent variables. The quantity $T_{ij}(\boldsymbol{X}, \boldsymbol{\omega})$ measures the error involved with the estimation of the i-th examinee and j-th item and involves both the data and parameters. Summing it over all items, j, produces a discrepancy measure

$$T_i^{person}(\boldsymbol{X}, \boldsymbol{\omega}) = \sum_j T_{ij}(\boldsymbol{X}, \boldsymbol{\omega})$$

for Examinee i. Summing it over all examinees, i, produces a discrepancy measure

$$T_j^{item}(\boldsymbol{X}, \boldsymbol{\omega}) = \sum_i T_{ij}(\boldsymbol{X}, \boldsymbol{\omega})$$

for Item j. Summing the $T_{ij}(\boldsymbol{X}, \boldsymbol{\omega})$s over both items and participants produces an overall discrepancy measure

$$T^{overall}(\boldsymbol{X}, \boldsymbol{\omega}) = \sum_j \sum_i T_{ij}(\boldsymbol{X}, \boldsymbol{\omega}) \ .$$

These discrepancy measures resemble the classical χ^2 goodness-of-fit measure. The proportion of iterations in which the values of $T_i^{person}(\boldsymbol{X}, \boldsymbol{\omega})$ from the actual data exceed the values of $T_i^{person}(\boldsymbol{X}^{rep}, \boldsymbol{\omega})$ is a person fit p-value p_i^{person} for each person. Similarly, we can get an item fit p-value p_j^{item} for each item and an overall p-value $p^{overall}$.

To look at some discrepancy measures that do not depend on the estimated parameters and latent variables, we also calculated the proportion correct scores for both the actual and replicated data. $Prop_j^{item}(\boldsymbol{X})$ is the proportion correct for Item j, and $Prop_i^{person}(\boldsymbol{X})$ is the proportion correct for participant i.

The item fit p-values do not indicate any misfit, with p-values lying between 0.23 to 0.73. The overall fit p-value is 0.35, which does not provide evidence against the model. Thus the posterior predictive model checks fail to detect the problems noted with the earlier diagnostics; however, the posterior predictive model checking method is known to be conservative. Alternatively, the discrepancy measures used may not have been effective.

The person fit p-values flag a number of persons with unusual response patterns. Using the measure $T_i^{person}(\boldsymbol{X}, \boldsymbol{\omega})$, we observe extreme p-values (more than 0.95 or less than 0.05) for Examinees 36, 75, 77, 101, 113, 116, 137, 179, 272, 298, 315, and 319. Looking at the response patterns of these examinees, we find that their response patterns are unusual. For example, Examinee 315, who has the most extreme p-value of 0.996, gets only Items 8, 11, 12, 13, and 15 correct. Looking back at Table 1, we find this is quite unusual. This examinee gets wrong the only two items (2 and 4) in Evidence Model 1 (requiring Skill 1 only). This suggests that he should get all 15 items wrong because all of them require Skill 1. However, he gets correct Item 8, (requiring Skills 1 and 2). He also gets 3 out of 5 items correct in Evidence Model 4 (requiring Skills 1, 3 and 4), but gets only 1 correct out of the 3 in Evidence Model 3 (requiring Skills 1 and 3), which requires a subset of the skills need for Evidence Model 4. As a result, the model cannot explain well the response of Examinee 315, resulting in an extreme p-value.

For $Prop_i^{person}(\boldsymbol{X})$, extreme p-values were obtained for Examinees 40, 41, 42, 44, 45, 50, 52, 53, 79, 82, 96, 113, 116, 136, 137, 148, 161, 178, and 288. Again, these examinees have unusual response patterns under the model. For example, Examinee 40, with a p-value of 0.998 gets Items 1, 6, 7, 12, 14, and 15 wrong. That means he gets 4 out of 5 items in Evidence Model 4 wrong. Since Evidence Model 5 requires a set of skills that includes all those required for Evidence Model 4, we would expect that he would get all three items in Evidence Model 5 wrong under the model; but he gets two of those correct. Naturally, the model underestimates the proportion correct for this examinee and we see a large p-value.

Notice that only three examinees are flagged as unusual by both the measures considered here, indicating that they measure different types of discrepancies and that both of them may be useful. One of the three common examinees, Examinee 113, gets Items 5, 7, 8, 10, 12, and 15 wrong. The examinee gets Item 8 wrong (the only one in Evidence Model 2, requiring Skills 1 and 2), indicating that he probably does not have Skill 2. He gets 3 out of 5 wrong in Evidence Model 4 (requiring Skills 1, 3 , and 4), indicating that he probably does not have Skill 4. From these observations, we would expect him to get Item 6 wrong under the model as this item requires Skills 1, 2, 3, and 4. However, he gets the item correct.

Since not many examinees are found to have extreme p-values (as we are using a 10% level test, we expect about 32.5 examinees to be flagged compared to the 28 flagged by both person-fit diagnostics), the person-fit indices do not indicate any major failure of the model. They do flag some response patterns that are unusual under the model, but they may suffer from the same lack of power that we saw with the item fit tests. Furthermore, the test length, 15 items, is rather small, which would also contribute to a lack of power.

Limitations of the 2LC model

The overall fit plots clearly indicate that the 2LC model cannot explain the data satisfactorily. The item fit plots give us a clue as to what might be happening—the model seems to predict Equivalence Classes 1 and 9 unsatisfactorily.

One limitation of the 2LC model is that it uses an all-or-nothing approach to explain the probability of a correct outcome from an item (i.e, it divides the examinees into two groups based on whether or not they have mastered all the necessary skills for solving an item and assigns the same probabilities to all equivalence classes within a group). Actually, that may not be the case—it may be easier to compensate for the lack of one skill than for the lack of many. There also is no latent variable representing overall mathematical proficiency. It is possible that such a skill would be related to both how quickly a student could master the skills and how readily the participant could apply them to a given problem.

5. Three Revised Models

Increasing the number of parameters in the link models will soften the all-or-nothing nature of the 2LC link model. By partitioning the equivalence classes into two groups, the 2LC model only fits two $\pi_{i\delta_{i(s)}}$ values for each item. Expanding the effective number of groups has an appropriate softening effects.

To form one new model, we divide the examinees not having the necessary skills to solve an item into two groups:

- 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

There is also the group who have the necessary skills for an item. We assign different success probabilities to each of the three groups. Mathematically, the proficiency model remains the same as the 2LC model, but the link model for examinee i and Item j (that uses evidence model s) is now:

$$X_{ij}|\pi_{jm}, (\delta_{i(s)} + I_i^1 = m) \quad \sim \quad \text{Bern}(\pi_{jm}), \text{ for } m = 0, 1, 2, \tag{8}$$

where $\delta_{i(s)}$ is the indicator of whether the Examinee i has mastered the all the skills required for solving items using evidence model s, and I_i^1 is the indicator function denoting whether Examinee i has the Skill 1. This model is called the "3LC" model as it has three parameters, π_{j0}, π_{j1}, and π_{j2}, for each item.

Treating the students who lack Skill 1 specially makes sense from both a empirical and a cognitive perspective. Empirically, Table 4 flagged Equivalence Class 1 (lacking Skill 1) more often than any other. However, according to our cognitive model, Skill 1 is a prerequisite for all of the others. Students who have yet to master Skill 1 are probably struggling with the very basics of fraction subtraction, and it makes sense that they would be less readily able to solve any problem. Consequently, we use a lower prior distribution for π_{j0}, a $Beta(3.5, 23.5)$, than we use for π_{j1}, a $Beta(6, 21)$. The prior for true positives, π_{j2} remains a $Beta(23.5, 3.5)$. Note that Items 2 and 4 only require Skill 1. Therefore, we set $\pi_{j0} = \pi_{j1}$ for those items.

Extension of the 3LC model by adding another π for each item produces the 4LC model. We divide the group of examinees having all the necessary skills to get an item correct into two subgroups: those who have mastered all five skills and those who are yet to master one or more skills. We assign different success probabilities to each of these two sub-groups. The proficiency model is the same as the 2LC and 3LC models, but the link model is now:

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

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.

$$\begin{aligned} \pi_{j0} &\sim Beta(2, 25) \\ \pi_{j1} &\sim Beta(6, 21) \\ \pi_{j2} &\sim Beta(21, 6) \\ \pi_{j3} &\sim Beta(25, 2) \end{aligned}$$

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

Yan et al. (2003) consider another extension of the 2LC model, which introduces a new latent variable η_i, which represents the examinee's propensity to solve problems with or without the requisite skills. In this model, the success probability of Examinee i for Item j (Item j uses the Evidence Model s, and $\delta_{i(s)}$ is the indicator denoting whether the Examinee i has mastered the skills needed for items using Evidence Model s) is:

$$\pi^*_{j\delta_{i(s)}} = \frac{\exp(\text{logit}(\pi_{j\delta_{i(s)}}) + s_j\eta_i)}{1 + \exp(\text{logit}(\pi_{j\delta_{i(s)}}) + s_j\eta_i)},$$

where, $\text{logit}(x) = \log \frac{x}{1-x}$, s_j is an item slope parameter for Item j. We refer to this as the "2LC+η" model. The prior distribution assumed for η_is is independent $N(0, 1)$ and that for s_js is independent $N(-2, 0.5)$.

The results of the diagnostics for the three expanded models are discussed below.

Bayesian Residual Plots

We create the residual plots for these models in the same way as discussed in Section 4. Figure 7 shows a plot for the posterior distribution of the number correct score residuals vs. the jittered predicted score estimates for the four models.

As in Section 4, we also create residual plots using the estimated θ-score of an examinee instead of the predicted number correct score. Figure 8 shows residual plots for all four models using this method. We draw a horizontal line at y-coordinate 0 in these residuals plots for ease of viewing.

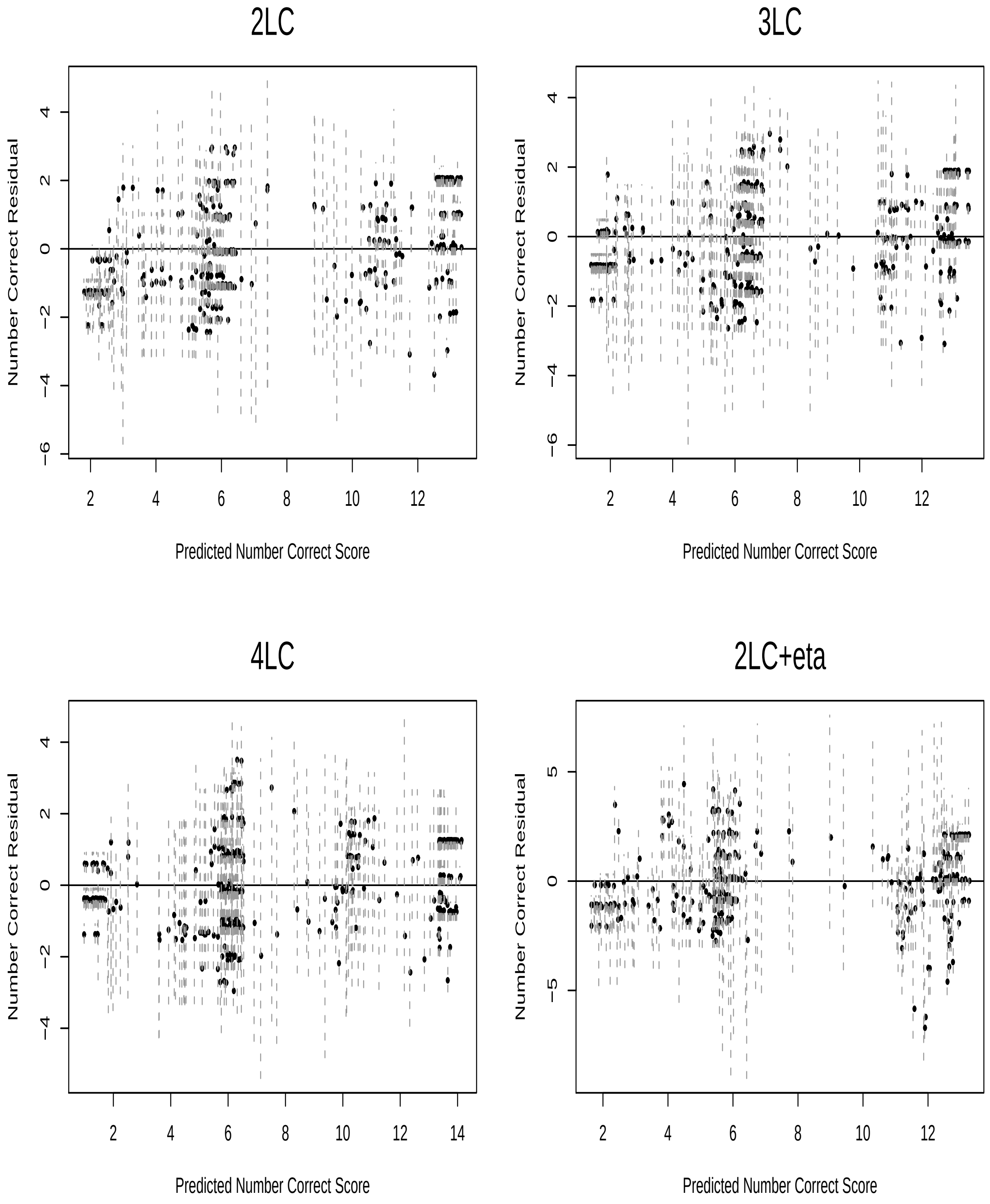


***Figure* 7. Plot of the posterior distributions of the number correct score residuals vs. the predicted number of correct scores for the four models.**

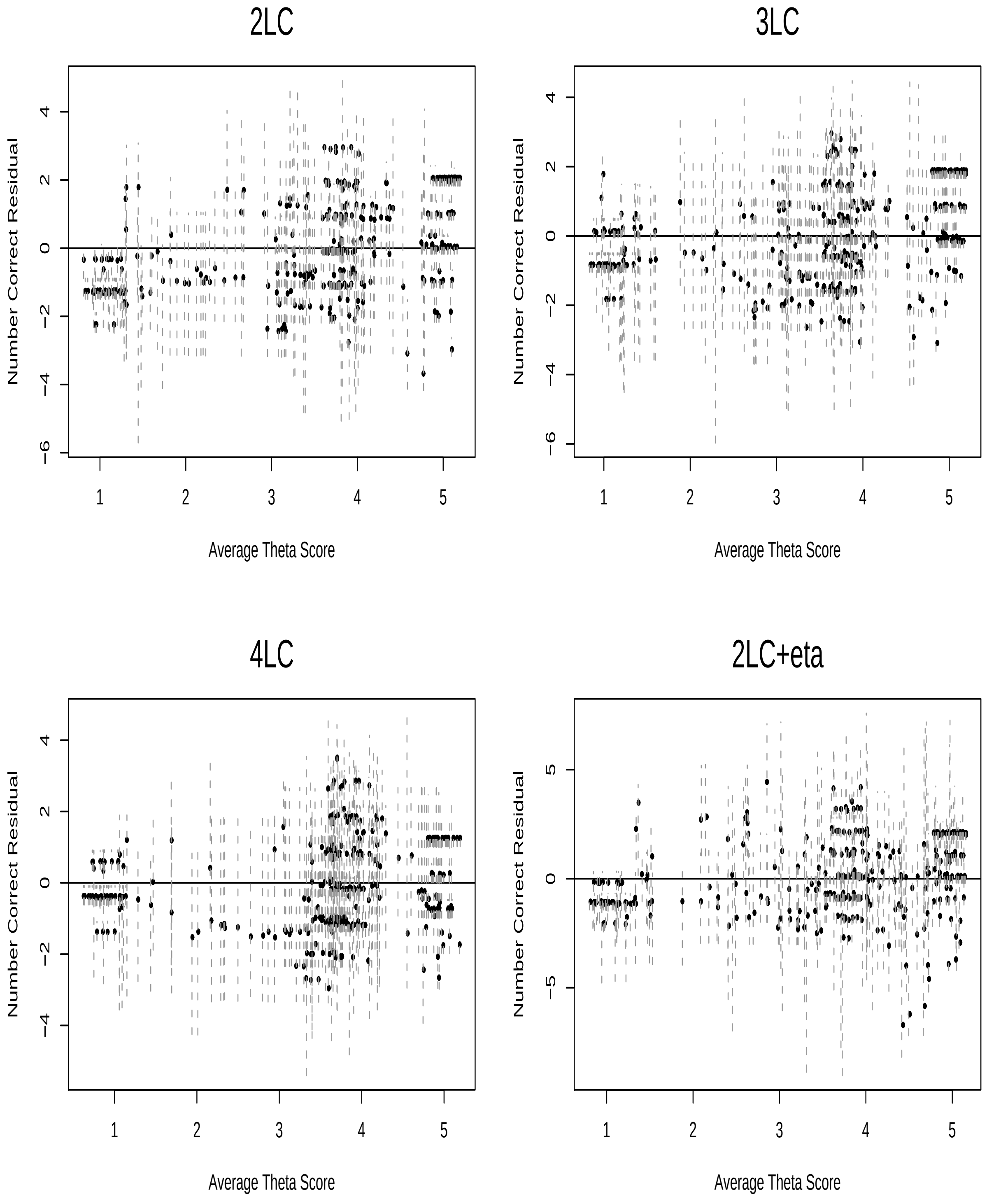


***Figure* 8. Plot of the posterior distributions of the number correct score residuals vs the estimated θ-scores for the four models.**

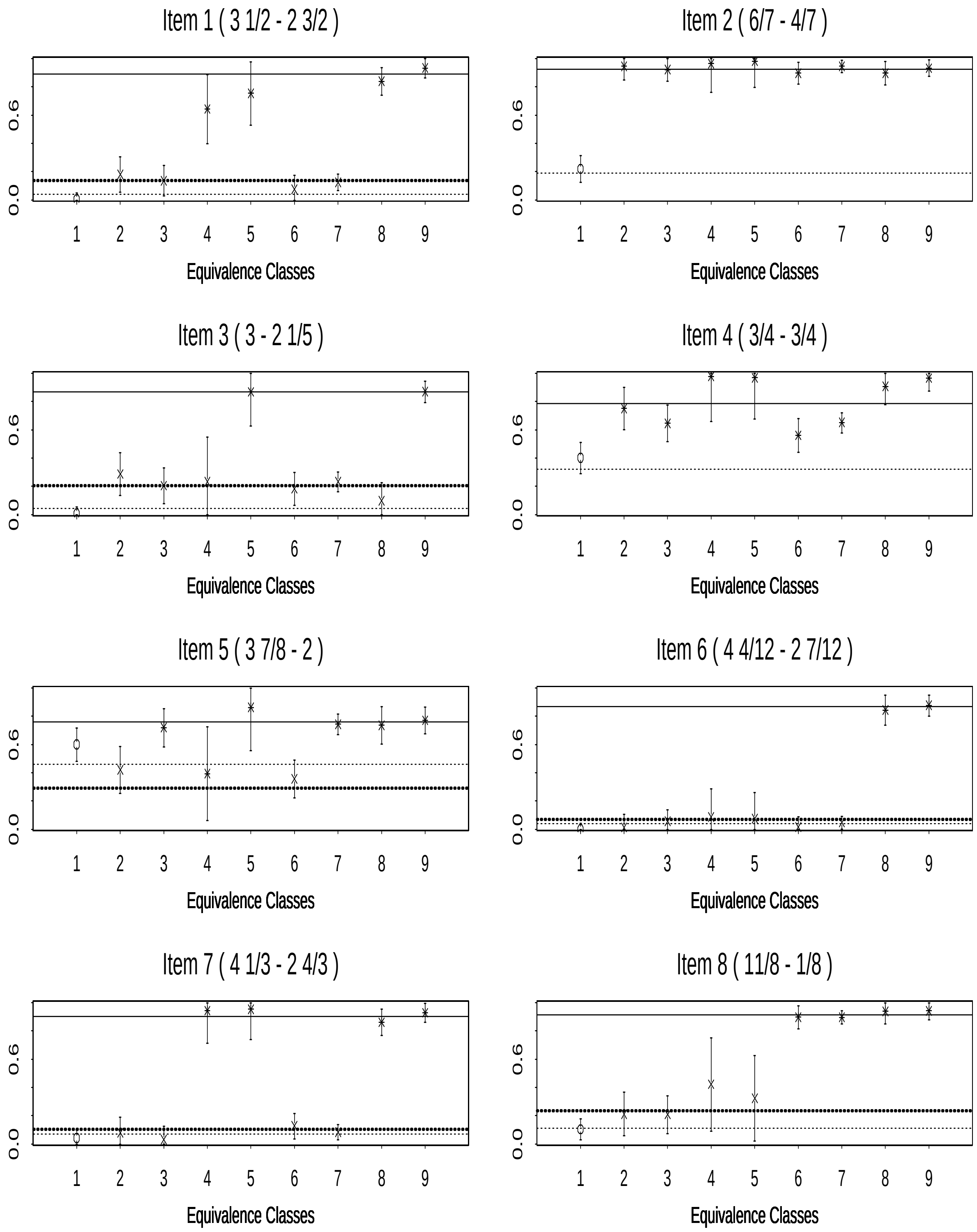


Figure **9. Item fit plots for Items 1–8 in the 3LC model.**

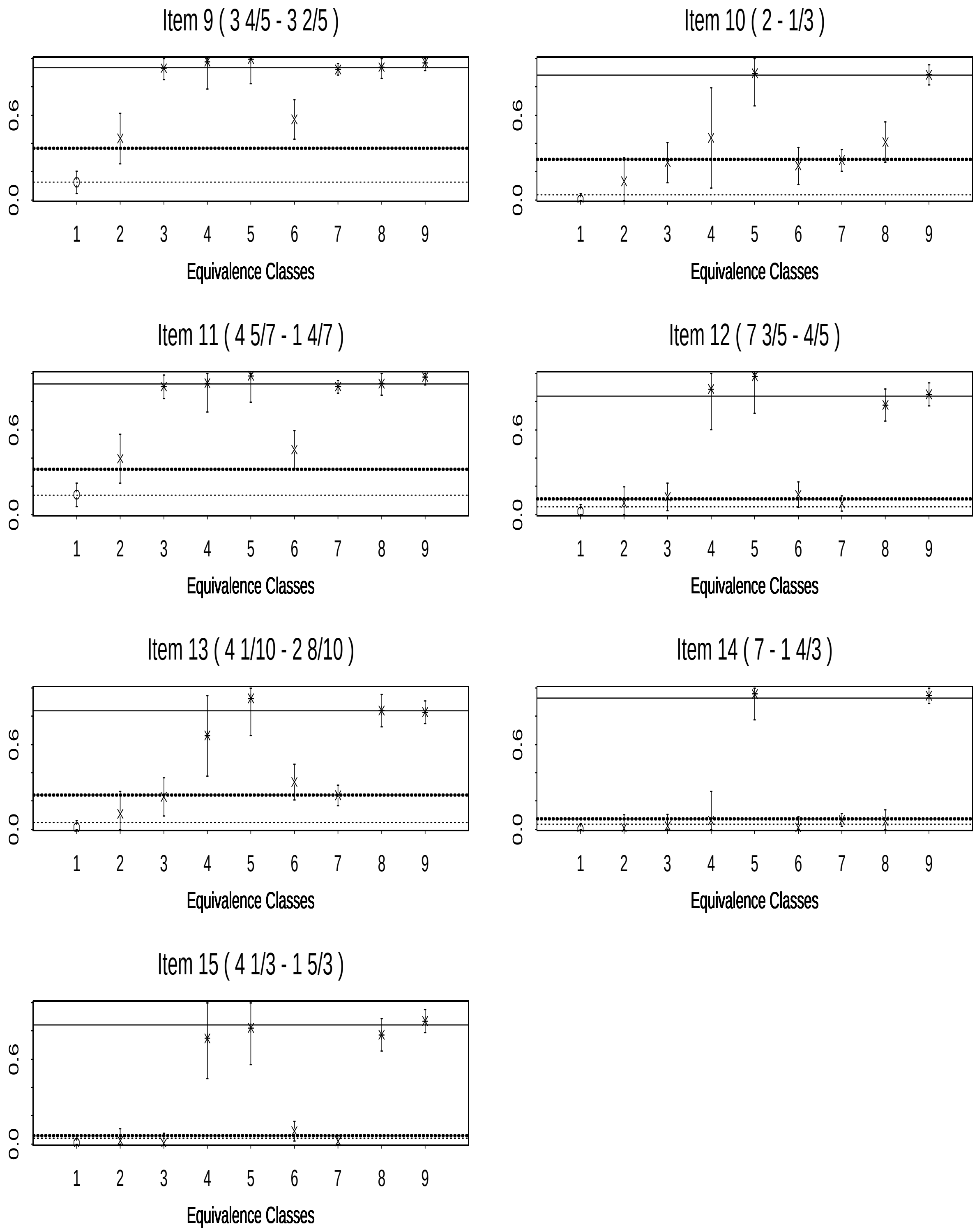


Figure **10. Item fit plots for Items 9–15 in the 3LC model.**

A look at the two sets of plots suggests that the 3LC and 4LC models do not have the same problem of bias at the ends of the proficiency scale as the 2LC model does. For these two models, the posterior distributions of the residuals have roughly the same concentration on an average above and below the 0-line for both low-proficiency (low estimated θ-score/predicted score) and high-proficiency examinees. The "2LC+η" model, even with all the extra parameters (which results in a long run-time of the BUGS program fitting this model), does not appear to do a very good job. It still has the same problem of over-prediction (like the 2LC model) at the low end of proficiency. Also, there seems to be a linear pattern with a negative slope in the residual plot of this model.

Item Fit Plots

We create item fit plots for the 3LC model (shown in Figures 9 and 10) in the same way as we created the plots for the 2LC model. The computation of the observed proportion corrects and the rough 95% confidence interval attached to them remains exactly the same as that for the 2LC model. The predicted proportion corrects that we compare these observed proportion corrects to are computed in a similar way as with the 2LC model except that now there are three πs for each item. There are three horizontal lines in the plot for each item now, one for each π for that item. The line for π_{j0} is dashed, that for π_{j1} is dashed and bold, and the line for π_{j2} is solid. A hollow circle, O, in the middle of an interval (marked by vertical lines) indicates that the members in that equivalence class do not have Skill 1 (this interval should contain the horizontal line for π_{j0}). An x indicates that the members in that equivalence class have Skill 1, but not all the necessary skills (this interval should contain the horizontal line for π_{j1}). An asterisk, *, means the members in that equivalence class have the necessary skills to solve that item (this interval should contain the horizontal line for π_{j2}). Note that Items 2 and 4 have only two πs each because they require only Skill 1 so $\pi_{j0} = \pi_{j1}$.

Figures 11 and 12 show the item fit plots for the 4LC model. Here, a hollow circle, $\circ$, an x, and an asterisk, *, in the middle of an interval mean the same as in the 3LC model. Additionally, here we use a solid circle, $\bullet$, to imply that the members in that equivalence class have all the Skills 1–5. The four horizontal lines for π_{j0}, π_{j1}, π_{j2}, and π_{j3} are dashed, solid dashed, solid, and solid bold respectively. Again Items 2 and 4 require Skill 1 only and have three πs each ($\pi_{j0} = \pi_{j1}$).

For the 3LC model, there is a discrepancy between the observed and predicted counts for Item 4 (Equivalence Classes 6, 7, and 9) and Item 5 (Equivalence Class 1). Item 9 also has a discrepancy with Equivalence Class 6, but as with the 2LC model, the effective sample size in Equivalence Class 6 is small, so this could be just chance fluctuation. Items 1, 3, 10, 11, and 12, for which misfit was detected for the 2LC model, seem to be consistent with the 3LC model. Hence, the 3LC model improves upon the 2LC model and seems satisfactory for the data.

Item fit plots for the 4LC model suggest that there is still some discrepancy with Items 4 and 5 (Equivalence Classes 8 and 1, respectively). For both the 3LC and 4LC model, the posterior mean of π_{j0} is higher than that of π_{j1} for these items. This is counterintuitive because π_{j0} is the success probability of the examinees who have not mastered any skills, while π_{j0} is the success probability of the examinees who have just mastered Skill 1.

Items 4 and 5 both have unusual forms that might indicate possible problems with the items. Item 4 is $\frac{3}{4} - \frac{3}{4}$. A student can solve this item without using method A or B (Section 2), by observing that the two quantities are the same, and hence their difference should be 0; Item 5 is $3\frac{7}{8} - 2$. This can also be solved without using method A or B, by using knowledge of integer subtraction and by noticing that an integer plus $\clubsuit$ minus another integer is the difference between the integers plus $\clubsuit$. As both of these items admit to solutions without using the mixed number subtraction algorithms being taught, they may be inappropriate for this assessment.

Test Statistics

Table 5 shows the values of the item fit statistics, given by (7), for the 2LC, 3LC, and 4LC models. The reference distribution for the 3LC model is the χ^2 with six degrees of freedom (as three parameters are estimated for each item) and five degrees of freedom for the 4LC model (as four parameters are estimated for each item).

While the 2LC model has four items flagged as problematic, only Item 4 ($\frac{3}{4} - \frac{3}{4}$) is flagged for the 3LC and 4LC models. Note that the problematic Item 5 is not flagged for

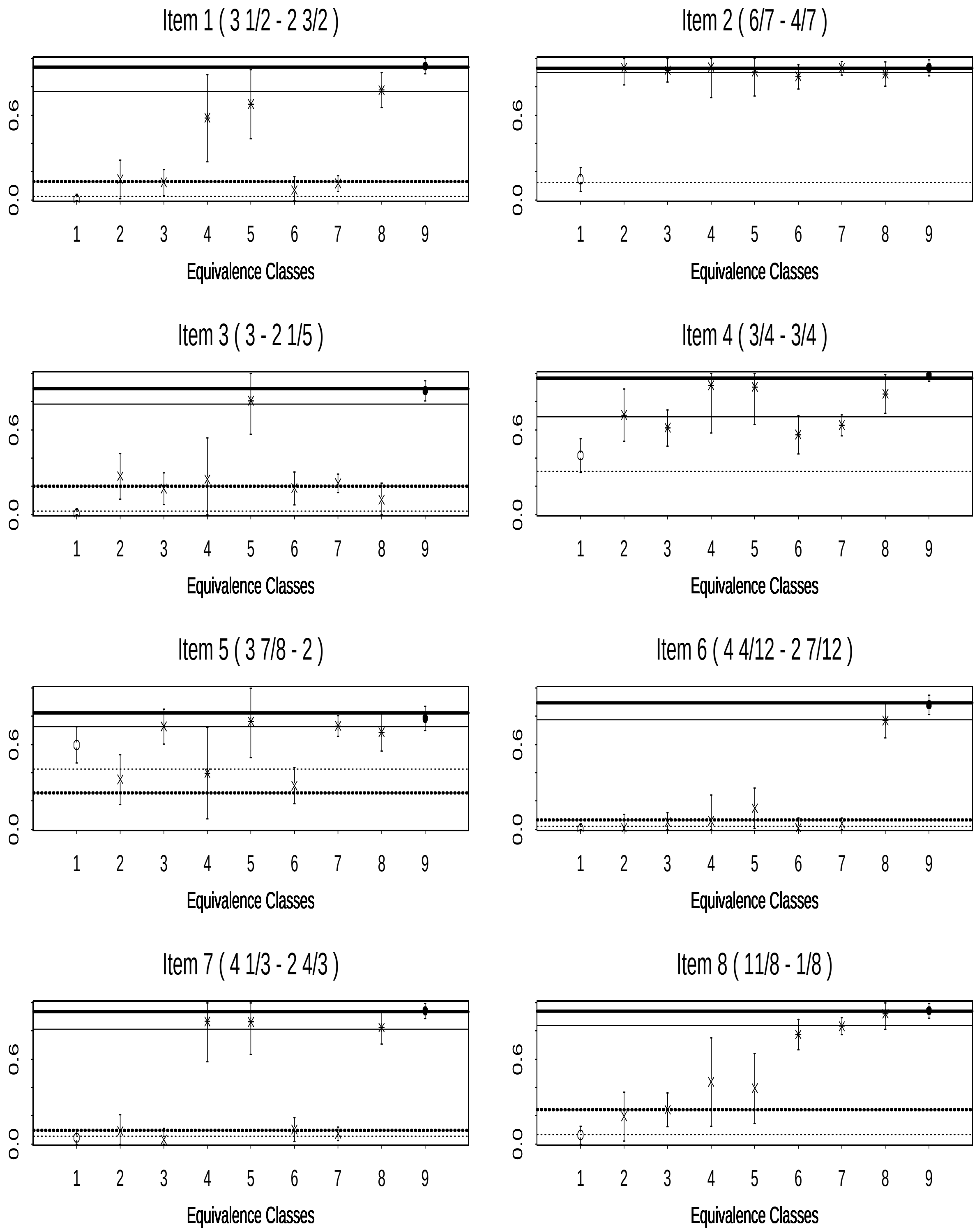
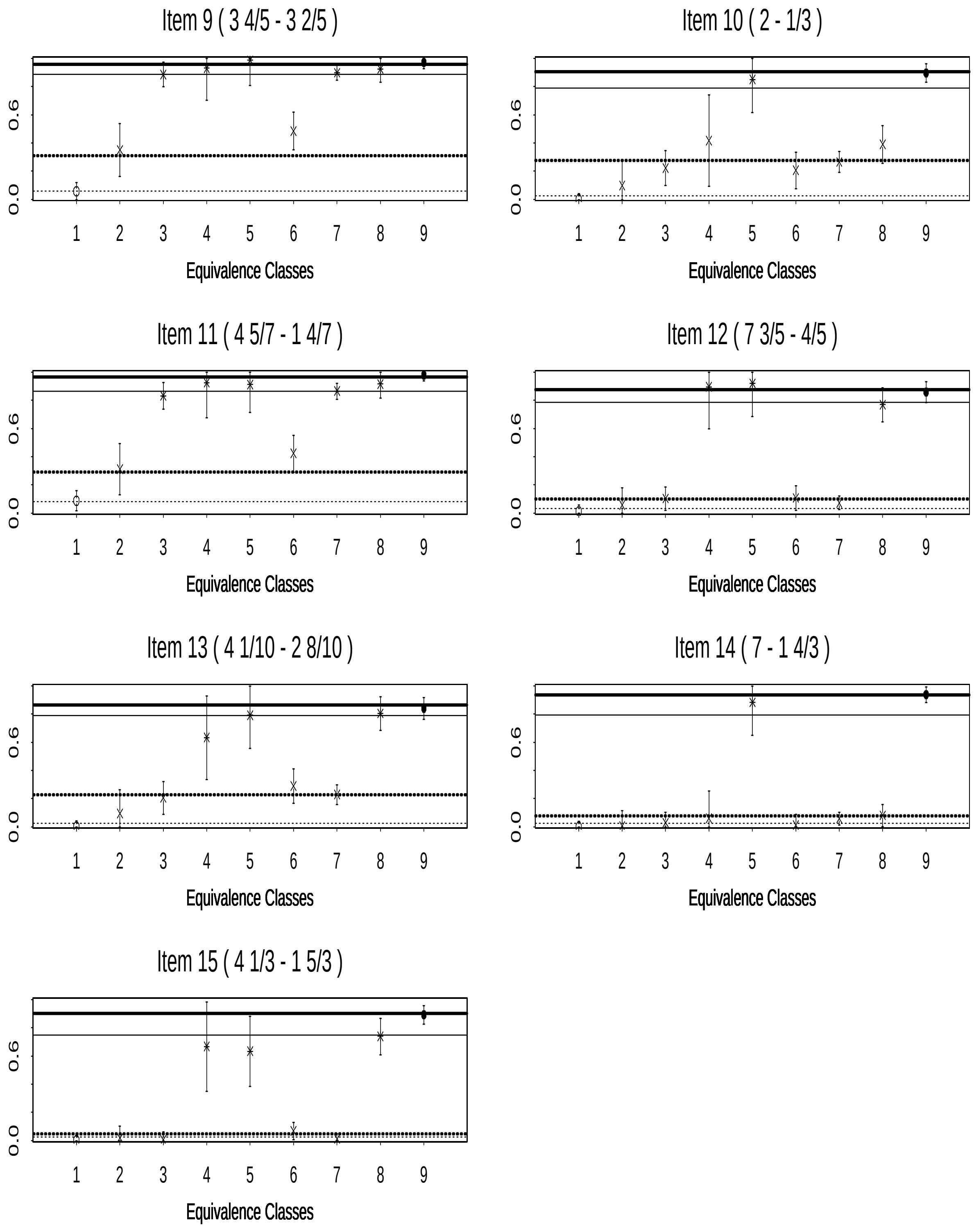


***Figure* 11. Item fit plots for Items 1–8 in the 4LC model.**



Item 9 (3 4/5 - 3 2/5)

Item 10 (2 - 1/3)

Item 11 (4 5/7 - 1 4/7)

Item 12 (7 3/5 - 4/5)

Item 13 (4 1/10 - 2 8/10)

Item 14 (7 - 1 4/3)

Item 15 (4 1/3 - 1 5/3)

Figure **12. Item fit plots for Items 9–15 in the 4LC model.**

Table 5.

Values of χ^2 Item fit Statistics for the 2LC, 3LC, and 4LC Models

Item	2LC model	3LC model	4LC model
no.	χ^2_j	χ^2_j	χ^2_j
1	10.7	7.6	2.5
2	2.2	2.0	1.9
3	15.0[a]	5.2	3.9
4	39.5[a]	39.3[b]	13.1[c]
5	7.6	8.6	8.8
6	3.3	5.1	4.3
7	2.7	3.4	1.4
8	3.7	2.3	3.3
9	8.5	5.6	3.5
10	24.1[a]	6.1	5.3
11	10.6	5.2	2.8
12	2.8	4.3	2.7
13	14.4[a]	4.1	2.5
14	4.0	5.1	3.6
15	3.9	7.9	5.6

[a] Larger than 95 percentile of $\chi^2_7 \approx 14$.

[b] Larger than 95 percentile of $\chi^2_6 \approx 12.5$.

[c] Larger than 95 percentile of $\chi^2_5 \approx 11$.

these models.

Posterior Predictive Model Checking

Application of the posterior predictive model checking methods to the 3LC and 4LC models yields similar results to the 2LC model. The measures do not indicate any item misfits or overall model misfit, but flag a number of person for possible misfits. However, the number of persons flagged is about half of that with the 2LC model, implying that these two models perform much better than the 2LC models in describing the response patterns of the examinees. The examinees whose responses are still not fit well are 36, 40, 42, 44, 45, 75, 113, 137, 148, 161, 179, 288, 315, and 319. Note that this list includes the three examinees (with unusual responses) discussed earlier in Section 4.

Choosing a Model

The statistical diagnostic tools considered here indicate that the 3LC model and the 4LC model seem to explain the data adequately. They are significant improvements over the 2LC model in explaining the overall fit and contain fewer items that do not fit. Between these two models, the 3LC model is preferable because it uses fewer parameters than the 4LC model without a noticeable lack of overall fit. The introduction of the parameter η in the "2LC+η" model does not seem to offer a significant improvement.

6. Evaluation of the Diagnostics

The set of diagnostic measures we proposed allowed us to uncover a problem with the 2LC model, characterize the nature of that problem, and come up with an alternative model, the 3LC model, which seems to fit the data better. Even though some practical issues remain (in particular, determining appropriate reference distributions for test statistics), the diagnostics still have some value in practical applications.

The Bayesian residual plots are simple, but may be powerful tools for detecting model misfit. They clearly detect the misfit of the 2LC model and even help us characterize the problem; specifically, the plots suggest misfit at the two ends of the proficiency scale. This knowledge leads to the suggestion of three possible improved models. Two of these three models (3LC model and 4LC model) seem to correct the problem. A refined and more powerful version of Bayesian residual plot is also suggested, but not pursued in this work. Although there is a recent surge of Bayesian statistical analysis in psychometrics, there has not been many applications of Bayesian residual analysis in the field, and this paper addresses that issue partially.

The item fit plots seem quite promising as well in detecting item misfits and overall model misfits. They provide more details about the lack of fit than do the residual plots. In particular, they help isolate the problem to Equivalence Class 1, which suggests the 3LC model as a good remedy. When applied to the 3LC model and the 4LC model, these plots suggest flagging two items that cannot be explained well by the cognitive theory ("$\frac{3}{4} - \frac{3}{4}$" and "$3\frac{7}{8} - 2$"). These items admit alternate solution paths and hence may be inappropriate for this examination.

The χ^2 test statistic corresponding to the item fit plots seems to have some power to detect problematic items. Consequently, some of the graphical tests could be automated, looking at the item fit plots only for items with high χ^2 statistics. Unfortunately, the reference distribution for these statistics is still unknown. Although the χ^2 statistic seems to provide a good heuristic, the true level and power of the test is still unknown. Perhaps a simulation study would provide a more useful reference value (Williamson et al. 2000, use simulations with other diagnostic measures in the context of these models).

The posterior predictive model checking method does not indicate any lack of fit of the overall model or the items. This lack of power possibly indicates that we have not yet stumbled upon a powerful discrepancy measure to use with this method. The posterior predictive model checking method does flag a number of examinees, whose response patterns indeed appear unusual in the context of the problem.

Although the residual plots and item fit plots proved useful in this problem, they are still conservative diagnostics because both of them use "observed values," which depend on parameters estimated from the data. Consequently, although they will detect extreme model misfit, they may miss subtler problems. Furthermore, the lack of reference distribution for the χ^2-type statistic means that its power and level are still unknown.

However, the most powerful argument for this method is that it can point to possible improvements in the underlying cognitive theory. The weakness in the 2LC model was a result of the strict conjunctive model for the applications of skills to items. The data clearly show us that individuals who lacked all of the requisite skills behaved differently from those who only lacked some. This finding may lead to better understanding of how to structure learning situations for students learning mixed number subtraction. Applying the diagnostics to similar situations can provide us the information to consider revising the cognitive theory, the statistical model, or the way data is collected and interpreted.

References

Albert, J., & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika, 82*(4), 747-59.

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

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

Chaloner, K., & Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika, 75,* 651-659.

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

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.

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 tests* (pp. 359–384). Hillsdale, NJ: Erlbaum.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 143–200). New York: Macmillan.

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

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology, 26,* 195-211.

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

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*(2), 171-189.

Johnson, M. S., Cohen, W., & Junker, B. W. (1999). *Measuring appropriability in research and development with item response models* (Technical Report No. 690). Pittsburgh, PA: Carnegie Melon University, Department of Statistics.

Junker, B., & Sijtsma, K. (2000). *Cognitive assessment models with few assumptions, and connections with nonparametric IRT.* Retrieved January 28, 2004, from http://www.stat.cmu.edu/%7Ebrian/apm/np-cog4.pdf

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: University of Illinois, Computer-based Education Research Laboratory.

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.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Madigan, D., & York, J. C. (1991). *Strategies of Bayesian updating of graphical models.* Paper presented at 23rd Interface Conference, Seattle, WA.

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 K. B. Laskey & H. Parde

(Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 437-446). San Francisco, CA: Morgan Kaufmann.

Mislevy, R. J., & Gitomer, D. (1996). The role of probability-based inference in an intelligent tutoring system. *User-modeling and User-adapted Interaction, 5,* 253-282.

Mislevy, R. J., Senturk, D., et al. (2001). *Modeling conditional probabilities in complex educational assessments.* Paper presented at the International Meeting of the Psychometric Society, Osaka, Japan.

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.

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.

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

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., & Stern, H. S. (2003). Variance component testing in generalized linear mixed models (ETS RR-03-14). 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, UK: Cambridge University, 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.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20,* 221–230.

Williamson, D. M., Almond, R. G., & Mislevy, R. J. (2000). Model criticism of Bayesian networks with latent variables. In C. Boutilier & M. Goldszmidt (Eds.), *UAI '00: Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence* (pp. 634–643). San Francisco: Morgan Kaufmann.

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

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245–262.