The Fusion Model for Skills Diagnosis: Blending Theory With Practicality

Sarah Hartz

Louis Roussos

October 2008

ETS RR-08-71

The Fusion Model for Skills Diagnosis: Blending Theory With Practicality

Sarah Hartz

Washington University, St. Louis, Missouri

Louis Roussos

Measured Progress, Dover, New Hampshire

December 2008

Abstract

This paper presents the development of the fusion model skills diagnosis system (fusion model system), which can help integrate standardized testing into the learning process with both skills-level examinee parameters for modeling examinee skill mastery and skills-level item parameters, giving information about the diagnostic power of the test. The development of the fusion model system involves advancements in modeling, parameter estimation, model-fitting methods, and model-fit evaluation procedures, which are described in detail in the paper. To document the accuracy of the estimation procedure and the effectiveness of the model-fitting and model-fit evaluation procedures, this paper also presents a series of simulation studies. Special attention is given to evaluating the robustness of the fusion model system to violations of various modeling assumptions. The results demonstrate that the fusion model system is a promising tool for skills diagnosis that merits further research and development.

Key words: Formative assessment, skills diagnosis, Markov chain Monte Carlo methods, fusion model, model fit, stepwise algorithm, item response theory, simulation, robustness, blocking, Q matrix

1 Introduction

Education and training, which constitute 9% ($722 billion) of the U.S. gross national product, require two kinds of test-influenced decision making. On one hand, high-stakes dichotomous decision making, including high school graduation, National Merit college scholarship awards, and admission to the college of one's choice, requires that the individual's performance scores be judged to be above a variety of different application-specific thresholds on a continuous unidimensional scale. As a result, test theory is dominated by the challenge of locating individuals on such unidimensional scales.

On the other hand, the day-to-day affairs of teaching and learning are predominantly characterized by relatively low-stakes decision making that classifies individuals based on dichotomous traits or attributes, often called *skills*. For example, a good teacher uses skill profiling in the classroom. After grading tests and noticing the classroom performance of each student on different relevant skills, the teacher has a sense of the specific skills for which each student needs targeted remediation. The focus of this paper is on advancing such skills-based diagnostic techniques.

The necessity for such categorical skills analysis has been most recently highlighted by the U.S. government's No Child Left Behind Act of 2001 (NCLB), which calls for classifying examinees into the categorical standards-based achievement levels of *below basic, basic, proficient*, or *above proficient* (U.S. House of Representatives, 2001). Further, the U.S. Department of Education draft regulations for NCLB proposes mandatory skills diagnosis in response to the Act itself (U.S. Department of Education, 2002):

> §200.8 Assessment reports.
>
> a) Student reports. A State's academic assessment system must produce individual student interpretive, descriptive, and diagnostic reports that ...
>
> **(1)(ii)** Help parents, teachers, and principals to understand and address the specific academic needs of students; and
>
> **(2)** Are provided to parents, teachers, and principals.

Skills diagnosis, sometimes referred to as *skills assessment* or *skills profiling*, is a relatively new area of psychometrics developed to statistically rigorize the process of evaluating (a) each examinee on the basis of his or her level of competence on an array of skills and (b) the test by assessing the strength of the relationship between the individual skills being assessed and the individual test items. Such examinee evaluations, often administered periodically throughout the teaching-and-learning process, is understood to be relatively low stakes and to have a valuable influence on the teaching and learning process. It is frequently referred to as *formative* assessment, in contrast to *summative* assessment, which takes place at the end of the teaching-and-learning process. A student's final grade in a class is typically heavily dependent on some weighted average of such summative assessments. Moreover, in the United States, all 50 states have embraced the use of summative assessments in the form of single-score-scale standardized tests to determine the proportion of students who have achieved state-designated academic standards at the end of each school year.

Instead of assigning a single ability estimate to each examinee, as in typical item response theory model-based summative assessments, skills diagnosis model-based formative assessments partition the latent space into more fine-grained, often discrete or dichotomous, skills (or other latent attributes), and evaluate examinees with respect to their level of competence on each skill. For example, suppose designers of an algebra test are interested in a standard set of algebra skills: factoring, using the laws of exponents, solving quadratic equations, and so on. A skills-diagnosis-based analysis attempts to evaluate each examinee with respect to each skill, whereas a standard summative psychometric analysis typically evaluates each examinee with respect to an overall scaled score on the algebra exam.

2 Toward More Effective Skills Diagnosis

The focus of published skills-diagnosis modeling methodologies has been primarily on the selection of the skills-diagnosis model itself. However, developing a model is only the first step in developing a skills-diagnosis system. A complete skills-diagnosis system must not only have (a) a well-selected model but also (b) a reliable, accurate, and efficient statistical estimation method and (c) effective implementation procedures for real data analyses.

To accomplish this first step, we reviewed the many skills-diagnosis models that have been presented in the literature. Of these models, the unified model of DiBello, Stout, and Roussos (1995) was chosen as the foundation for the model developed in this paper, which we call the *fusion model.* The unified model was chosen as our starting point because it includes a greater variety of components than other models, thus giving us more flexibility for shaping our own model. However, the unified model's large number of parameters also required some reparameterization and simplification due to nonidentifiability of some of the unified model parameters. This is discussed in more detail below.

To accomplish the second step, selecting the estimation method and probability framework, we considered a number of approaches, such as the EM algorithm and Bayesian networks, and finally settled on using a Bayesian probability framework in conjunction with a Markov chain Monte Carlo (MCMC) estimation algorithm.

Although estimation procedures have been developed for many skills-diagnosis models, it does not appear that any of these modeling approaches go beyond estimation to develop dynamic implementation procedures, facilitating skills diagnosis of real data by evaluating the fit of the model to the data, and manipulating the model structure as a result of this evaluation. In contrast, the methodology presented in the current paper represents a complete skills-diagnosis system:

1. An identifiable and interpretable model called the fusion model, an enhancement of the unified model (DiBello, Stout, & Roussos, 1995),

2. An effective and efficient parameter-estimation method operationalized in a software program referred to as Arpeggio, which employs a Markov Chain Monte Carlo algorithm within a Bayesian framework,

3. A model-fitting procedure referred to as the *stepwise algorithm* that is used to eliminate noninformative item parameters from the model, and

4. Practical model-fit evaluation procedures.

After briefly summarizing some of the past research on skills-diagnosis modeling (to place the fusion model system in its historical and technical context in relation to past work), this paper presents detailed descriptions of the fusion model item response function, the Bayesian framework for the estimation method, the stepwise algorithm model-fitting procedure, and the model-fit evaluation procedures. This methodology is then tested on simulated data, using data generated from the model as well as data generated with a limited number of serious, but realistic, violations of model assumptions.

3 Diagnostic Modeling in Education Testing

Skills-diagnosis modeling has two major practical foci: (a) to determine the mastery/nonmastery skill profiles of examinees taking a test and (b) to evaluate the test and its items in terms of their effectiveness in measuring the individual skills. Psychometric techniques addressing these specific issues were created as early as 1973, when Fischer proposed the linear logistic trait model (LLTM) to decompose item-difficulty parameters from a logistic model into discrete skills-based components. Rather than evaluating examinees with respect to the individual skills as well, he maintained a unidimensional ability parameter in the LLTM. Conversely, Tatsuoka and Tatsuoka (1982) proposed the statistical rule space approach to decompose examinee abilities into skill components. Although the rule space approach estimates skill profiles of each examinee taking the exam, it does not statistically evaluate how well the test and its items measure the skills. While both the LLTM and rule space approaches have been applied to test data, they have not been widely implemented in practical education assessment settings. This may be because practical skills diagnosis cannot go far without modeling both the relationship between the items and the skills and the relationship between the examinees and the skills.

Many models for skills diagnosis have been developed subsequently to improve on either LLTM (Fischer, 1973) or rule space (Tatsuoka & Tasuoka, 1982). Notable examples include the multicomponent latent trait model (MLTM) of Whitley (name now changed to Embretson) (1980); the general component latent trait model (GLTM) of Embretson (1984); the restricted latent class model of Haertel (1989); the HYBRID model of Gitomer and Yamomoto (1991); the unified model of DiBello, Stout, and Roussos (1995); the Bayesian networks of Mislevy (1994) and of Yan, Mislevy, and Almond (2003); the tree-based approach of Sheehan (1997); the discrete mixture Rasch model of Bolt (1999); the conjunctive, disjunctive, and compensatory models of Maris (1999); and the dichotomization of the MLTM by Junker (2000). These advancements in skills-diagnosis modeling have fluctuated between complex models representing cognitive views of problem solving that are statistically intractable when applied to test data and simpler models (still complex in comparison to standard unidimensional models) that are more reliably applied to test data.

Despite a great need for practical skills diagnosis, for the models that have estimation algorithms, we are unaware of any further development of implementation procedures for fitting the models to real data and evaluating the subsequent fit to the data. Thus, to further advance the field of skills diagnosis, this paper develops the fusion model skills diagnosis system (fusion model system), which not only provides a model that includes both examinee- and item-diagnostic parameters but also provides fitting methods and model-fit evaluation procedures. First, we describe the fusion model item response function (a reparameterization of the unified model item parameters) and its Bayesian framework, and then we describe the estimation algorithm and the implementation methodology (model-fitting and model-fit evaluation methods). Establishing an effective estimation algorithm and effective implementation procedures provides a critical link for bridging the gap between skills-diagnosis model development and the practical use of skills-diagnosis methods in mainstream test evaluation. For more detailed reviews of skills-diagnosis models, see DiBello, Roussos, and Stout (2007), Embretson (1999), Junker (1999, 2000), and Hartz (2002).

4 The Fusion Model

Like all other IRT models, IRT-based skills-diagnosis models define the probability of observing examinee j response to item i given examinee ability parameters and item parameters. Symbolically, this probability is represented as $P(X_{ij} = x \mid \underline{\vartheta}_j, \underline{\beta}_i)$, where $X_{ij} = x$ is the response of examinee j to item i (with $x = 1$ indicating a correct response and $x = 0$ an incorrect response), $\underline{\vartheta}_j$ is a vector of examinee j ability parameters, and $\underline{\beta}_i$ is a vector of item i parameters. The fundamental assumption of IRT modeling is that, conditioned on the examinee ability parameters, examinee response to any item i is independent of examinee response to any other item i'. The distinguishing feature of skills-diagnosis models from other IRT models is that the items $i = 1, ..., I$ relate to a set of cognitive skills $k = 1, \ldots, K$ in a particular manner. Although Fischer (1973) specified this relationship as f_{ik} (the "weight" of skill k in item i), the weights are usually either 1 or 0, in which case they reduce to what is now known as the Q *matrix*, namely, $Q = \{q_{ik}\}$, where $q_{ik} = 1$ when skill k is required by item i and $q_{ik} = 0$ when skill k is not required by item i. Although the concept had been used previously in several different models, the Q matrix notation was first introduced by Tatsuoka (1990). Further, Tatsuoka's work emphasized the importance of the Q matrix to skills diagnosis.

The unified model (DiBello, Stout, & Roussos, 1995) features both skills-based item parameters and skills-based examinee parameters. Furthermore, the unified model includes additional parameters to improve the fit of the model to the data. As discussed by Samejima (1994) in her competency space theory, let the examinee parameter $\vartheta = \{\underline{\alpha}_Q, \underline{\alpha}_b\}$ (examinee subscript j suppressed) denote the complete latent space of all relevant skills. Let $\underline{\alpha}_Q$ be the vector of cognitive skills specified by the Q matrix. The remaining latent space, $\underline{\alpha}_b$, includes the relevant skills outside those specified by the Q matrix. Samejima (1995) referred to $\underline{\alpha}_b$ as skills associated with "higher order processing," and suggested that these skills may be more substantively important than $\underline{\alpha}_Q$. From the unified model perspective, however, $\underline{\alpha}_b$ does not need to be interpreted as higher order processing; it is simply a representation of the latent skills outside the Q matrix. The unified model is the first skills-diagnosis model to incorporate $\underline{\alpha}_b$ into the model by defining a single unidimensional ability parameter η_j as a unidimensional projection of examinee j's $\underline{\alpha}_b$ ability. The inclusion of this concept in the unified model will be shown below to be connected to a type of item parameter that can be used to diagnose whether a test item is well modeled by the Q matrix skills that have been assigned to it. The explicit acknowledgment that the Q matrix is not necessarily a complete representation of all the skill requirements for every item on the test differentiates the unified model from the other skills-diagnosis models.

Define $\pi_{ik} = P(Y_{ikj} = 1 \mid \alpha_{jk} = 1)$ and $r_{ik} = P(Y_{ikj} = 1 \mid \alpha_{jk} = 0)$, where $Y_{ikj} = 1$ refers to the event that examinee j correctly applies skill k to item i, $\alpha_{jk} = 1$ indicates that examinee j has mastered skill k, and $\alpha_{jk} = 0$ indicates that examinee j has not mastered skill k. The item response function (IRF) for the unified model is given in equation 1:

$$P(X_i = 1 \mid \underset{\sim}{\alpha}_j, \eta_j) = d_i \prod_{k=1}^{K} \pi_{ik}^{\alpha_{jk} \cdot q_{ik}} r_{ik}^{(1-\alpha_{jk}) \cdot q_{ik}} P_{c_i}(\eta_j) + (1 - d_i) P_{b_i}(\eta_j), \tag{1}$$

where $P_h(\eta) = \frac{1}{1+exp\{-1.7[\eta_j-(-h)]\}}$, a Rasch model with difficulty parameter $-h$.

The product term in the model indicates the assumption of conditional independence of applying the skills, provided the Q-based strategy is used. By further assuming local independence of the item responses, Equation 1 can be used to model the probability of any given response pattern, $\underset{\sim}{x}$. For each item i on the test, there are $2k_i + 3$ (k_i = number of skills required by item i) unified model item parameters: π_{ik} and r_{ik}, two IRT Rasch model parameters c_i and b_i, and the final parameter d_i, the probability of selecting the Q based strategy over all other strategies.

In addition to building an IRF based on the Q matrix, the unified model allows the predicted Q-based response to be influenced by non-Q skills with the term $P_{c_i}(\eta_j)$, and allows for alternate non-Q strategies with the term $P_{b_i}(\eta_j)$. As with the models of Maris (1999) and the GLTM of Embretson (1984), the unified model has cognitively interpretable parameters; but unfortunately, not all the parameters are statistically estimable.

The flexibility and interpretability of the unified model parameters led it to be chosen as the foundation for the fusion model skills-diagnosis system developed in this paper. However, because nonidentifiable parameters existed in the original unified model (Jiang, 1996), a reduction in the parameter space was required before its parameters could be estimated. The initial attempts to reparameterize the model included retaining all item parameters except for $k_i - 1$ π_{ik}'s for each item, where k_i refers to the number of skills required for item i (DiBello, Stout, & Jiang, 1998; Jiang, 1996). Specifically, Jiang set $\pi_{ik} = 1$ and $i = 1, ..., k_i - 1$. However, when $k_i - 1$ π_{ik}'s are fixed at 1, the interpretation of all the other item parameters is distorted, and the highly useful capacity to interpret the strength of a skill in modeling item response correctness is lost. Instead, Hartz (2002) reparameterized the unified model to be identifiable in a way that retains interpretability of the parameters. To further reduce the complexity of the parameter space and to enhance the estimability of the parameters, the modeling of the possibility of alternate strategies has been dropped for the work reported here by setting $d_i = 1$ for $i = 1, ..., I$. The reduced model (reparameterized unified model) now has $2 + k_i$ parameters per item, compared to the $2k_i + 3$ parameters per item in the original unified model. The reduced model maintains the unified model's flexible capacity to fit diagnostic test datasets as compared to other skills-diagnosis models, retaining the most substantively important components, like the capacity for skill discrimination to vary from item to item and the residual ability parameter η, an additional and potentially important component of the original unified model that is missing from all the other models. Equation 2 presents the resulting fusion model IRF, often referred to as the *reparameterized unified model.* It is based on the same examinee parameters, $\underline{\alpha}_j$ and η_j, that are used in the original unified model. The $P_{c_i}(\eta_j)$ term again refers to the Rasch model with difficulty parameter $-c_i$ (the lower the value of c_i, the lower the value of $P_{c_i}(\eta_j)$):

$$P(X_{ij} = 1 \mid \underline{\alpha}_j, \eta_j) \quad = \quad \pi_i^* \prod_{k=1}^{K} r_{ik}^{*\,(1-\alpha_{jk})\times q_{ik}} P_{c_i}(\eta_j). \tag{2}$$

It is important for understanding and applying the fusion model that the interpretation of these parameters be clearly understood. Here,

$$\begin{aligned} \pi_i^* \quad &= \quad P(\text{correctly applying all item } i \text{ required skills given} \\ &\quad\quad \alpha_{jk} = 1 \text{ for all item } i \text{ required skills}) \\ &= \quad \prod_{k=1}^{K} \pi_{ik}^{q_{ik}} (\text{under the assumption of conditional independence of} \\ &\quad\quad \text{individual skill application}) \end{aligned}$$

$$
\begin{aligned}
r^*_{ik} &= \frac{P(Y_{ijk} = 1 \mid \alpha_{jk} = 0)}{P(Y_{ijk} = 1 \mid \alpha_{jk} = 1)} \\
&= \frac{r_{ik}}{\pi_{ik}}, \text{ and} \\
c_i &= \text{the value of } \eta_j \text{ for which } P_{c_i}(\eta_j) = 0.5,
\end{aligned}
$$

where, $0 \leq \pi^*_i \leq 1$, $0 \leq r^*_{ik} \leq 1$, and $0 \leq c_i \leq 3$. (The bounds of 0 and 3 on the c parameter were chosen for convenience rather than because of any strict theoretical or logical constraint.) The fusion model reparameterization replaces π_{ik} and r_{ik} in the original unified model with π^*_i and r^*_{ik}. In addition to producing an identifiable parameter set, the new parameters are conceptually interpretable in a particularly appropriate way from the applications perspective. The parameter π^*_i is the probability that an examinee having mastered all the Q required skills for item i will correctly apply all the skills when solving item i. The correct item response probability for an examinee who has not mastered a required skill k_0 is proportional to $r^*_{ik_0}$. The more strongly the item depends on mastery of this skill, the lower the item response probability for a nonmaster of the skill, which translates to a lower r^*. Thus, r^*_{ik} is like a reverse indicator of the strength of evidence provided by item i about mastery of skill k. The closer r^*_{ik} is to zero, the more discriminating item i is for skill k.

The distinctiveness of the π^*s and r^*s in comparison to parameters in other models is important to note. Other models have indeed included components similar to π^*_i and r^*_{ik}. The models in Maris (1999) have the π_{ik} and r_{ik} parameters of the unified model (DiBello et al., 1995), which are nonidentifiable. Conversely, the discrete MLTM of Junker (2000) has skill-based item parameters that are identifiable, but not item specific, so the influence of the skill on each individual item response probability is lost. This is especially important from the perspective of skills-based test design, where one wishes to know for each skill which items are most effectively discriminating between examinee possession and nonpossession of that skill.

The $P_{c_i}(\eta_j)$ component is an important unique component retained from the unified model because it acknowledges the fact that the Q matrix does not necessarily contain all relevant cognitive skills for all the items. Interestingly, it is not present in any other skills-diagnosis model. In this component, c_i indicates the reliance of the IRF on skills other than those assigned to that item by the Q matrix. As an approximation, these other skills are modeled, on average over all the items, by a unidimensional ability parameter, η_j. When c_i is 3 or more, the IRF is practically uninfluenced by η_j, because $P_{c_i}(\eta_j)$ will be very close to 1 for most values of η_j. When c_i is near 0, η_j variation will have increased influence on the item response probability, even with $\underline{\alpha}_j$ fixed. Thus, the estimate of c_i can provide valuable diagnostic information about whether a skill is missing from the Q matrix or whether a skill already in the Q matrix needs to be added to an item's list of measured skills. Indeed, one of our robustness studies below provides a demonstration of the effectiveness of c as this kind of a diagnostic.

In summary, the fusion model enables the estimation of the most critical examinee parameters from the original unified model while reparameterizing the unified model's item parameters so that they are not only estimable but also retain their skills-based interpretability, a feature that makes such models more attractive to users of educational tests in comparison to traditional unidimensional psychometric models.

5 Estimating the Model Parameters

After reparameterizing the unified model, a variety of possible methods for estimating the fusion model parameters were explored. In order to incorporate flexibility in fusion model parameter relationships and to simplify estimation procedures, we decided to use a Bayesian approach to estimating the parameters. While using Bayesian networks is one type of Bayesian approach that could be adopted, we found that the probability structure of interest could be combined with relationships between the skills more easily by using a hierarchical Bayesian modeling approach instead. This, in effect, enhances the reparameterized unified model of Equation 2 by adding further parameters and their priors (typically vague or estimated from the data).

To frame a complex non-Bayesian model into a reasonable Bayesian model, the priors, hyperparameters (parameters of the prior distributions that in turn have completely specified prior distributions), and the priors of the hyperparameters must be constructed so that the estimated values of the model parameters will be determined predominantly by the data, not by the priors (assuming, as is usual, that little information about likely parameter values is known a priori and that the data contain parameter estimation information). Thus, the goal in building the Bayesian framework for our model was to incorporate hypothesized structural relationships between variables that are justified by the scientific theory governing the setting being modeled, and to construct noninformative priors for the unknown relationships or distributions so that the data can reveal detailed aspects of the relationships between the variables. This approach was believed to have the potential for greater inferential power than a non-Bayesian approach because it exploits aspects that are already known about parameter relationships while allowing the data to provide information about the unknown aspects of the relationships.

The most obvious example of the type of relationship that would be more difficult to account for in a non-Bayesian approach is the matrix of positive correlations that exists among the pairs of examinee skills. Incorporating this relationship is particularly important in skills-diagnosis models for education. More specifically, mastery of one skill is not statistically independent of mastery of another. Additionally, an examinee who has mastered many skills among $\underline{\alpha}$ can be expected to have a higher η. Such correlations have been observed in education testing even for abilities that are seen as highly distinct, like mathematics ability and reading ability. Thus, this assumption would certainly hold for a diagnostic test where the skills are designed to be dimensionally close to one another.

Although typical hierarchical Bayesian models would simply specify a prior distribution for the dichotomous α_{jk}s directly (for example, see the work of Yan et al., 2003), incorporating multidimensional correlations of the dichotomously parameterized skills (mastery versus nonmastery) is a very difficult task. In particular, raw correlations between the dichotomous skills are highly dependent on the differing proportions of masters for each skill. To deal with this problem, we used a modified version of a well-known latent variable technique — tetrachoric correlations. Tetrachoric correlations were developed to model the relationship between two dichotomously measured variables where it is assumed that a normally distributed latent variable generates each observed dichotomous variable (see, for example, Hambleton & Swaminathan, 1985).

In the fusion model application, tetrachoric correlations are used to model the relationship between two dichotomous mastery variables. Thus, the Bayesian framework of the fusion model incorporates continuous $\tilde{\alpha}_{jk}$, $k = 1, \ldots, K$ with standard normal priors. Once the $\tilde{\alpha}_{jk}$ variables are generated, they are converted to the dichotomous α_{jk} variables by comparing their values to κ_k cutoff values, which are estimated as hyperparameters. Specifically, $\alpha_{jk} = 1$ when $\tilde{\alpha}_{jk} > \kappa_k$. That is, the examinee is considered to have mastered the skill because the examinee latent ability is greater than the mastery cutoff value. Likewise, η_j is given a standard normal prior. Since it is assumed that the examinee skills have positive correlations, $(\underline{\tilde{\alpha}}_j, \eta_j) \sim N(\mathbf{0}, \Sigma)$ where $\Sigma = \{\rho_{mn}\}$ has 1s on the diagonal for the marginal variances of the skills and the non-negative pairwise correlations between $(\underline{\tilde{\alpha}}_j, \eta_j)$ as the off-diagonal elements. These correlations are estimated as hyperparameters and are given a Uniform prior over some interval: $\rho_{m,n} \sim \text{Unif}(a, b)$, where the variables a and b were set to 0.01 and 0.99, respectively, to prevent boundary problems. The hierarchical Bayesian model for the examinee parameters is seen in Figure 1.

Since the values of the item parameters π_i^*, r_{ik}^*, and c_i vary greatly for different datasets (indeed, they often vary greatly within a single dataset), the distribution functions for the item parameter priors were chosen to be beta distributions, allowing maximum flexibility of the shape the priors can take. For some datasets, one of the item parameters may vary little across items, in which case a highly peaked prior distribution would be expected to yield the best results. For other datasets, an item parameter may vary so greatly across items that a uniform distribution would be the most appropriate choice. By careful choice of its parameters, the shape of a beta distribution can be made to resemble either the highly peaked distribution or the uniform distribution, or nearly any distribution in between.

This valuable flexibility of the beta distribution, however, cannot be taken advantage of unless we already know the distribution of the item parameters. This seemingly insurmountable problem has a solution: The parameters of the beta priors can themselves be estimated by assigning them priors and estimating the corresponding hyperparameters. Thus, by estimating the parameters of the beta distributions used for the item parameter priors, we allow the data to intelligently inform the shape of the item parameter priors that are most appropriate for the given situation.

Using the above Bayesian framework to produce the fusion model, a Markov chain Monte Carlo (MCMC) estimation software program called Arpeggio was written in Fortran using the Metropolis-Hastings within Gibbs algorithm (see, for example, Patz & Junker, 1999a, 1999b) to simultaneously estimate examinee and item parameters. Since MCMC is used simply as an estimation tool in this perspective, the algorithm is not discussed in detail here. The reader is referred to Hartz (2002) for a detailed discussion of our application of the MCMC methodology.

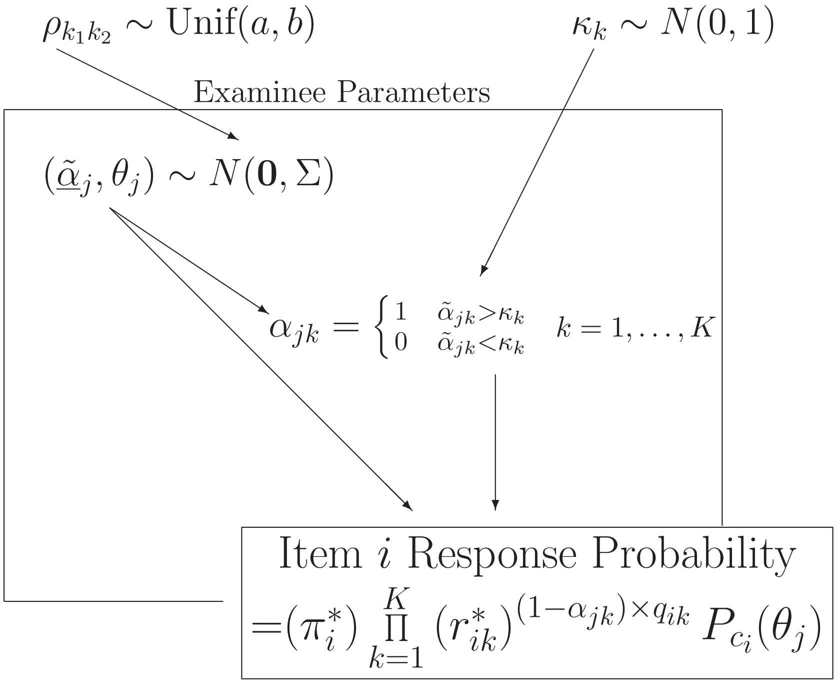
$\rho_{k_1 k_2} \sim \text{Unif}(a, b)$
$\kappa_k \sim N(0, 1)$
Examinee Parameters
$(\underline{\tilde{\alpha}}_j, \theta_j) \sim N(\mathbf{0}, \Sigma)$
$\alpha_{jk} = \begin{cases} 1 & \tilde{\alpha}_{jk} > \kappa_k \\ 0 & \tilde{\alpha}_{jk} < \kappa_k \end{cases} \quad k = 1, \ldots, K$
Item i Response Probability
$=(\pi_i^*) \prod_{k=1}^{K} (r_{ik}^*)^{(1-\alpha_{jk}) \times q_{ik}} P_{c_i}(\theta_j)$

Figure 1. **Fusion model hierarchical Bayesian formulation (examinee parameters).**

6 Stepwise Algorithm Model-Fitting Procedure

To conduct effective skills diagnosis, it is not enough to have merely a good model and a good parameter estimation algorithm. One also needs accompanying procedures for ensuring model fit. To this end, we have developed a stepwise algorithm to help eliminate statistically non-influential item parameters, specifically, noninfluential r^*s or cs, as described in more detail below.

Even when a Q matrix is very carefully developed for estimation of the fusion model with real data, or even when one generates simulated data and uses the known Q matrix in one's estimation model, there may be item parameters that, for the given data, are statistically insignificant.

For example, in a real data setting, a skill may be assigned by the Q matrix to an item, but it may be that examinees do not actually need or use the skill in correctly responding to the item (for example, the item may require only a very elementary application of the skill and thus play an almost nonexistent role statistically, even though it is required for the item), or a skill may be needed to a much lesser degree than other skills that also are delineated by the Q matrix. (This real data setting can be emulated in simulation studies by using a Q matrix in the estimation model that does not match the one used in simulating the data; see below.)

Even when simulated data are analyzed using the Q matrix that was used in generating the data and all the item parameters are identifiable, the structure of the data may not contain sufficient information to well estimate all the parameters used to generate the data.

An example of this situation is when data are simulated with low r^*, high π^* (referred to as *high cognitive structure*), and moderate c values. In this case, the c parameters may not be estimable because the item responses are largely determined by whether examinees have or have not mastered the skills and are very little affected by their proficiency on non-Q skills. Additionally, if an item measures both a hard skill and an easy skill, and the r^* for the harder skill is very low, the r^* for the easier skill will have little influence in the item response function. Thus, eliminating such noninformative parameters is considered a critical component of the fusion model system because it helps the model to concentrate its statistical power where there is diagnostic information to be found.

To identify noninfluential parameters, the stepwise algorithm estimates the influence of each parameter, either r^* or c, on the IRF that uses that parameter, employing a common probability scale for measuring this influence. Thus, the algorithm uses the same statistical decision rule for both r^* and c in determining whether or not the estimated influence is large enough to warrant keeping the parameter in the model. In our simulation studies, these decisions are based solely on a statistical hypothesis-testing framework in which the null hypothesis is that the item parameter under investigation has negligible influence on the parameter's item response function. It is important to note, however, that in practice, such decision making would typically be based on an interaction of both statistical and substantive input. Because a Q matrix is often developed with strong theoretical arguments to back it up, eliminating Q matrix entries (which is what happens when an r^* is eliminated) in practice requires not only strong statistical evidence but strong substantive arguments as well.

To estimate the influence of a particular item parameter, three different item response probabilities are calculated for each examinee from its fusion model IRF. With all the other parameters fixed at their estimated means, three IRF probabilities are calculated: (a) with the parameter fixed at its null hypothesis (no influence) value (an r^* would be fixed at 1.0, and a c would be fixed at 10.0), (b) with the parameter set to its estimated mean minus its estimated standard deviation, and (c) with the parameter set to its estimated mean plus its estimated standard deviation. For an r^*_{ik} parameter, this calculation is done for all examinees who are estimated as non-masters of skill k (these are the only examinees for whom the r^*_{ik} would appear in their item response function). For a c parameter, the calculation is done using all the examinees. When on average over all these examinees, the item response probability (a) is close (as defined below) to that for either (b) or (c) (or both, depending on the preference of the practitioner), the parameter is said to be non-influential.

Determining whether a difference in average probabilities is close is ultimately a subjective decision, but still there does exist experience from other IRT domains of interest that can help guide decision making here. The most common examples of such decision making that we are familiar with are in the context of differential item functioning (DIF) analyses. In DIF analyses, an average difference between IRFs of less than 0.01 is certainly considered a negligible amount, while a difference between 0.05 and 0.10 is considered moderate, and a difference greater than 0.10 is considered large (see, for example, Dorans, 1989). In the simulation studies conducted in this paper, we found that a relatively liberal approach (using cutoff values that are on the low side within the range of the typical DIF cutoffs) worked best for these particular studies. In particular, in our simulation studies we declare that a c or r^* is noninfluential for an item (and, thus, set the parameter to its null hypothesis value) when the average of (a) is within 0.03 of the average of (b) or within 0.01 of the average of (c).

Practitioners may, of course, choose other reasonable criteria and may choose to use only the difference between (a) and (b) or only the distance between (a) and (c). The former corresponds to a situation where a practitioner is philosophically choosing a null hypothesis that the parameter belongs in the model and will only take it out if the data provide strong evidence otherwise. The latter corresponds to a situation where the null hypothesis is that the parameter does not belong in the model and will only be included if the data provide strong evidence otherwise. Both philosophies are statistically valid, as is also the combined use of both types of criteria.

Finally, at any one step in the algorithm, we restrict ourselves to dropping a maximum of only one parameter per item. If more than one parameter appears noninfluential at the same time, then it is possible that the dropping of one parameter could cause another parameter to become influential. If both an r^* and a c parameter are found to be noninfluential at the same step, the c parameter is preferentially dropped to favor the possibility of retaining skill mastery estimation in the item over the possibility of retaining estimation of η since skills diagnosis is the primary purpose of the analysis. Also, if two or more r^*s for a particular item are found to be influential, the parameter with the smallest average difference between the null-hypothesis IRF probability (a) and the upper bound IRF probability (c) is dropped.

7 Model-Fit Evaluation Procedures

In addition to the stepwise parameter reduction algorithm to help ensure model fit by eliminating unnecessary item parameters, we have also developed three procedures for evaluating model fit after the estimation process is completed. These procedures are referred to as IMstats, EMstats, and FusionStats.

IMstats stands for *item mastery statistics*, and it describes on an item-by-item basis and on average over all the items how well the Arpeggio MCMC estimates of examinee mastery of each skill correspond to the actual observed performance of the examinees on each of the items. Specifically, for each item on the test, the examinees are separated into three groups according to how many of the item's skills each examinee has been estimated as having mastered. Examinees who have mastered all the skills for item i are called the "item i masters." Examinees who are nonmasters on at least one, but not more than half, of the skills required for item i are called the "item i high nonmasters." Examinees who lack mastery on more than half the skills required for item i are called the "item i low nonmasters." This use of the terms "master" and "nonmaster" is a little different from our previous usage. Here mastery is discussed relative to *all* the skills required for an item, not mastery relative to a single skill. Then, on an item-by-item basis, IMstats computes the proportion-correct score on each item for the examinees falling into each of these three groups for that item. To check if the Arpeggio MCMC estimation procedure is working well in conjunction with the Q matrix, the results of IMstats are examined to see whether the item masters have performed decidedly better than item nonmasters and, similarly, whether the item high nonmasters have performed substantially better than the item low nonmasters. We consider this decision making to be fairly subjective in that it depends on what one considers to be a negligible difference in these proportions. We do not recommend a hypothesis-testing approach here because these proportions are calculated over a large number of examinees, and their standard errors would be expected to be very small, making even nonconsequential small differences statistically significant.

EMstats stands for "examinee mastery statistics," and it is used to search for misfitting examinees who perform either unexpectedly poorly on the items for which they have mastered all the required skills or unexpectedly well on the items for which they have not mastered all the required skills. EMstats produces evaluation statistics on an examinee-by-examinee basis as well as summary statistics over all examinees. Specifically, for each examinee, the items on the test are divided into three groups (similar to the IMstats examinee groups) based on the proportion of each item's skills the examinee has been estimated as having mastered:

Mastered items. Items for which the examinee has been estimated as having mastered all the Q skills

High nonmastered items. Items for which the examinee has been estimated as having nonmastered at least one, but not more than half, of the skills required by the item

Low nonmastered items. Items for which the examinee has been estimated as having nonmastered over half of the skills required by the item

The last two categories are also combined to form a fourth general category for all the nonmastered items. For each examinee, EMstats calculates the number of mastered, high nonmastered, and low nonmastered items for that examinee and the examinee's proportion-correct score on each of these sets of items.

EMstats then compares the examinee's observed proportion-correct score for the examinee's mastered items with a criterion value to see if this score is unusually low. Similarly, the scores on the examinee's high nonmastered, low nonmastered items, and all the nonmastered items together are also compared to criteria to see if these scores are unusually high. Because these observed proportion-correct scores are frequently based on fairly small numbers of items, we decided that a hypothesis-testing approach to the decision making was needed to avoid high Type 1 error rates. Thus, for each examinee EMstats performs a one-tailed hypothesis test for each set of items. The hypothesis tests are one-tailed because, for the mastered items, we only care whether the examinee had an especially low proportion-correct score, and for any of the sets of nonmastered items we only care whether the examinee had an especially high proportion-correct score. The criterion values are either set by the user or are based on the observed item proportion-correct scores for masters, high nonmasters, and low nonmasters from IMstats. The latter set of criteria are used in the current study.

The hypothesis tests are conducted in the following manner. First, the examinee's observed proportion-correct score is subtracted from the appropriate criterion value. Then, under the assumption that the null hypothesis holds, the standard error is computed as $\sqrt{\sum P_{ci}(1 - P_{ci}) \div n}$, where the summation is over the n items of interest and P_{ci} is the criterion value for item i. Next, a simple z statistic is formed by dividing the difference by the standard error. Assuming that this z statistic is approximately normally distributed, the hypothesis test is performed by determining if the calculated z statistic is less than -1.645 when the focus of the hypothesis test is on the mastered items or is greater than 1.645 when the focus is on the nonmastered items.

To help ensure the validity of the assumption of approximate normality, we require the number of items in a category to be above some minimum level. Specifically, the minimum level is equal to the number of items needed for the hypothesis test to reject an observed difference of 0.2 or more at level 0.05.

FusionStats stands for "fusion model statistics," and it compares two sets of model predicted statistics with the values of the statistics that are actually observed in the data. The two statistics are the item proportion-correct scores and the examinee proportion-correct scores. Thus, for every examinee, FusionStats prints out the examinee's observed proportion-correct score on the test and the proportion-correct score predicted by the model using all the estimated item parameters and the examinee's estimated ability parameters (the dichotomous skill mastery estimates and η). Similarly, for each item, FusionStats prints out the observed proportion-correct score for the item and the proportion-correct score predicted using all the examinee estimated ability parameters and the the estimated item parameters for the item. These observed and predicted statistics can then be examined to see if there are any large differences that may indicate lack of fit, where large is an admittedly subjective value decided on by the user. Again, as with IMstats, we do not recommend a hypothesis-testing approach because these proportions are based on large numbers of items or examinees and would have very small standard errors, making even nonconsequential differences statistically significant.

8 Simulation Studies

Before the fusion model system can be applied to real test data, it must be verified that the MCMC estimation method, the stepwise algorithm model-fitting procedure, and the model-fit evaluation procedure are behaving appropriately. Such evaluation requires a series of realistic and informative simulation studies.

Although there are many ways to structure a skills-diagnosis simulation study, we decided to focus our initial study on the case of high cognitive structure, where the data are simulated to be highly informative about skill mastery. We felt it important that we first verify that our procedures work well in situations where they are intended to perform well.

To maintain manageability of the size of our study while still varying some critical factors, we decided to focus our simulations on a 40-item, 7-skill test with randomly generated item parameters and 1,500 examinees. In addition, for all the studies presented here, the simulated proportions of masters for the skills were held constant across all the models studied, as were the distributions of the underlying continuous skill variables (the means and standard deviations were fixed; the correlations between the $\tilde{\alpha}$'s and η were randomly generated for each dataset).

The different simulation studies use data that either vary the complexity of the generating Q matrix or introduce a variety of carefully selected misspecifications in the Q matrix used in the estimation algorithm. By misspecifications, we mean that the Q matrix used to estimate the algorithm differs from the Q matrix used to generate the data.

The generation of the datasets, including the selection of the Q matrix, item parameters, examinee parameters, and item responses, is explained in section 8.1.

In particular, our simulation study may be thought of as a three-layered study: (a) evaluation of fusion model estimation under ideal low-complexity conditions, (b) evaluation of fusion model estimation with a more cognitively complex (thus, less ideal) Q matrix, and (c) evaluation of fusion model estimation robustness when the Q matrix used in the fusion model parameter estimation differs from the Q matrix used to generate the data (using the low-complexity Q). Using this structure as the framework for the simulation studies, the following specific research questions and subquestions were then answered:

1. A detailed evaluation of the overall performance of fusion model parameter estimation and, in particular, the fusion model MCMC estimation procedure itself was conducted under ideal test conditions. By ideal test conditions, we mean that (a) the Q matrix that generates the simulated dataset is the same as that used in the MCMC fusion model parameter estimation procedure, (b) the data are generated using items whose responses are heavily influenced by examinee mastery versus nonmastery of the Q required skills (high cognitive structure), and (c) the number of skills per item is moderate (low complexity). The model used here will be referred to as the *ideal low complexity* model. The questions addressed in this case were:

    1. How well does the Markov chain of the MCMC algorithm converge to the appropriate posterior distribution (section 9.1)?

    2. Are the item parameters well estimated (section 9.2)?

    3. Are the examinee parameters (especially the αs) well estimated (section 9.3)?

    4. How well does the fusion model fit the data (section 9.4)?

Using the performance of the fusion model system for the ideal low complexity model as a baseline, the above performance criteria were then evaluated for fusion model parameter estimation with datasets analyzed under less ideal conditions:

1. We first extended the above evaluation to the case of a more complex Q matrix. As in the study above, the correct Q matrix was given to the MCMC estimation algorithm, but, in this case, the Q matrix had a greater number of skills per item. The model used in this case will be referred to as the *high complexity model.* Thus, in addition to addressing the above questions, this study also asked, What happens to the parameter estimation and model fit when the number of skills per item in the Q matrix is increased from an average of two skills per item (the ideal low complexity model), to an average of three skills per item (section 10.1)?

Besides modifying the complexity of the Q matrix used in generating the data, it is important to acknowledge that the user-supplied Q matrix is unlikely to be identical to the Q matrix that generated the data.

2. The second type of nonidealness introduced is the situation where the user-supplied estimate of the Q matrix differs from the Q matrix that was used to generate the data, which is done using the ideal low complexity model. Thus, in addition to evaluating the same four questions from the first simulation study, this study also looked at the following robustness questions:

    - What happens when an extraneous skill is included in the Q matrix used in the MCMC fusion model parameter estimation algorithm; i.e. a skill is asserted to be influential, in many items, via the Q matrix used in the MCMC fusion model estimation algorithm but in fact plays no role in influencing examinee item performance in the model used to generate the data (section 10.2.1)?
    - What happens when a skill is left out of the Q matrix used in the MCMC fusion model estimation algorithm, but it in fact plays a major role (for many items) in influencing examinee item responses (i.e., the skill was used for the data generation; section 10.2.2)?
    - What happens when the skills specified to be required by one particular item are severely incorrect in the Q matrix used in the fusion model estimation algorithm for parameter estimation (section 10.2.3)?

8.1 Generating Data for the Simulation Studies

This section provides a detailed discussion of the generation of the Q matrix, examinee parameters, and item parameters for the datasets corresponding to the ideal low complexity model. The resulting simulated dataset is subsequently used to address Research Questions 1 (ideal test conditions) and 3 (incorrect specification of the Q matrix). The dataset corresponding to Research Question 2 (high complexity model) is generated in a similar manner, except that a different Q matrix is used, as described in section 10.1.

A Q matrix was randomly generated to require two skills per item, on average: 12 items required one skill, 18 items required two skills, 8 items required three skills, and two items required four skills. This translates to an average of 11.4 items for each skill. The randomly generated Q matrix entries are given in Table 1.

Since some skills are easy to master and others are difficult, tests typically require skills of varying mastery prevalence in the population. Thus the simulated examinee proportion of skill mastery was spread out over the range of 0.30 to 0.65 in the following manner: 0.30 for Skill 1, 0.40 for Skill 2, 0.45 for Skill 3, 0.50 for Skill 4, 0.55 for Skill 5, 0.60 for Skill 6, and 0.65 for Skill 7.

Item parameters were generated to maximize cognitive structure while retaining estimation power for each individual r^* (i.e., by reducing the effects of blocking; see discussion in section 9.3), important in cases where multiple skills are required per item. The generated values are given in the Appendix in Table A1. The examinee parameters were generated using the Bayesian structure given in Figure 1, using randomly generated, positive correlations of continuous multivariate normal latent abilities $(\underset{\sim}{\tilde{\alpha}}, \eta)$ between 0.13 and 0.85 (with a mean of 0.42 and a standard deviation of 0.22). We purposely chose to include low correlations in addition to high ones in order to stress the model estimation, although the lower correlations are much less likely to occur in practice.

Item responses were generated using the given reparameterized unified model parameters for 1,500 examinees. The proportions of simulated examinees that correctly responded to each item are given in Table 2. Notice that the item parameters in combination

Table 1

Randomly Generated Q Matrix for a 40-Item, 7-Skill Exam; Ideal Low Complexity Model

	Skill							
	1	2	3	4	5	6	7	Total
Item 1	0	0	1	0	1	0	0	2
Item 2	1	0	0	0	0	0	0	1
Item 3	1	0	0	0	1	0	0	2
Item 4	0	0	0	0	0	0	1	1
Item 5	1	0	0	1	0	0	0	2
Item 6	0	1	0	0	1	0	1	3
Item 7	0	0	1	1	0	0	1	3
Item 8	0	0	1	0	1	0	0	2
Item 9	0	0	1	0	1	0	0	2
Item 10	0	0	0	1	0	0	0	1
Item 11	0	0	0	0	0	1	0	1
Item 12	0	0	0	0	0	0	1	1
Item 13	1	0	0	1	0	0	1	3
Item 14	0	0	0	1	0	0	0	1
Item 15	0	1	1	0	0	0	0	2
Item 16	0	0	0	0	1	0	1	2
Item 17	0	0	0	0	1	0	1	2
Item 18	1	0	1	0	0	1	1	4
Item 19	1	0	0	0	0	1	0	2
Item 20	1	0	0	0	0	0	0	1
Item 21	0	1	0	0	0	0	0	1
Item 22	0	0	0	0	0	1	0	1
Item 23	0	0	0	1	0	0	0	1
Item 24	0	0	0	1	0	1	0	2
Item 25	0	1	0	0	0	1	0	2
Item 26	0	0	0	0	1	0	0	1
Item 27	0	0	0	0	0	1	1	2
Item 28	0	1	1	1	0	0	0	3
Item 29	0	1	1	0	1	0	1	4
Item 30	0	0	1	0	0	0	1	2
Item 31	1	0	0	1	0	1	0	3
Item 32	0	1	0	1	0	0	0	2
Item 33	0	0	0	1	0	0	0	1
Item 34	0	1	1	0	0	1	0	3
Item 35	1	1	1	0	0	0	0	3
Item 36	0	1	0	0	1	0	0	2
Item 37	1	0	0	1	0	0	0	2
Item 38	0	0	0	0	1	1	0	2
Item 39	1	0	0	0	0	1	1	3
Item 40	0	0	0	0	0	1	1	2
Total	11	10	11	12	11	12	13	

Table 2
Item Proportion Correct for the Generated Data Using the Ideal Low Complexity Model

	Proportion correct		Proportion correct		Proportion correct		Proportion correct
Item 1	0.34	Item 11	0.38	Item 21	0.44	Item 31	0.21
Item 2	0.22	Item 12	0.49	Item 22	0.37	Item 32	0.33
Item 3	0.24	Item 13	0.30	Item 23	0.35	Item 33	0.39
Item 4	0.54	Item 14	0.44	Item 24	0.39	Item 34	0.24
Item 5	0.27	Item 15	0.22	Item 25	0.28	Item 35	0.21
Item 6	0.39	Item 16	0.36	Item 26	0.54	Item 36	0.40
Item 7	0.31	Item 17	0.39	Item 27	0.56	Item 37	0.32
Item 8	0.28	Item 18	0.17	Item 28	0.29	Item 38	0.32
Item 9	0.26	Item 19	0.29	Item 29	0.24	Item 39	0.20
Item 10	0.39	Item 20	0.39	Item 30	0.43	Item 40	0.45

with the proportion of masters for each item resulted in low proportion-correct scores for the majority of items (i.e., the simulated test is difficult). Not surprisingly, the item with the highest proportion correct, Item 27, required only the easiest skills (i.e., high proportion of masters), Skills 6 and 7.

9 Results for Ideal Low Complexity Model

9.1 Evaluating Markov Chain Convergence

Although much has been written in the literature regarding the convergence of Markov chains in MCMC estimation, there is no simple statistic that reliably evaluates whether the Markov chain for each model parameter has converged. In this application, we measured convergence in two ways. First, for each of the parameters, the estimated posterior distributions from several independent chains were compared to each other by visually inspecting the estimated densities and by calculating the Gelman and Rubin $\hat{R}$ (the ratio of between-chain variance plus within-chain variance to within-chain variance; Gelman, Carlin, Stern, & Rubin, 1995). These procedures were used to confirm that the independent chains were estimating the same posterior distribution. Second, the time-series plots and autocorrelations of the Markov chains for individual parameters were evaluated.

Rather than display voluminous tables and graphs, we now summarize our results for determining convergence.

In evaluating convergence in terms of $\hat{R}$, we compared each of the calculated $\hat{R}$ values with the standard upper-limit cutoff value of 1.2 recommended by Gelman et al. (1995). Our results showed that $\hat{R}$ values for all the item parameters and for η were all below 1.2, indicating that convergence with respect to $\hat{R}$ had occurred.

In evaluating convergence in terms of the autocorrelation function for a parameter being estimated by MCMC, the function should ideally decrease to zero quickly. As would be expected, the decrease is slower for many of the parameters of a complex model, such as the fusion model. Partial evidence for nonconvergence of a fusion model parameter was thus defined as an autocorrelation of above 0.2 after lag 200.

A few of the parameters were indeed observed to have autocorrelations of 0.3 at lag 200, despite the fact that the standard deviations of the proposal distributions were optimally manipulated to produce relatively quick convergence (see, for example, Raftery & Lewis, 1996). Although this is consistent with nonconvergence, it is also consistent with convergence and a multidimensional likelihood surface that is difficult to maximize (flatness or ridges, for example). This latter explanation is the more plausible one given that long chains (five of length 6,000, each with 1,000 burn-in) were used, that the posterior distributions from independent chains were very similar (as determined by $\hat{R}$ and visual inspection), and that the parameter estimation in the simulation results was observed to be very accurate. In particular, parameters that had little influence on the item response (like a high c_i) would be expected to have a particularly flat likelihood surface, resulting in high autocorrelations. This is a common problem with complicated models that are likely to have nonsignificant parameters, parameters whose values for a given dataset are not significantly different from the values one would assume under an appropriate null hypothesis of no parameter influence (for example, see Carlin, Xia, Devine, Tollbert & Mulholland, 1999). This phenomenon is not the same as the nonidentifiability of overparameterized models. In such cases, the likelihood is relatively flat for all datasets.

9.2 Evaluating the Quality of Item Parameter Estimation

Although the main goal of a skills diagnosis is to correctly classify examinees in terms of the skills they have or have not mastered, mastery classification will not be effective unless the item parameters are reasonably well estimated. Moreover, reasonably accurate estimation of the item parameters is essential for evaluating the quality of the items for effectively measuring skill mastery versus nonmastery and for improving the Q matrix as specified based on prior substantive considerations (as shown in the simulations of section 10.2.1). Further, reasonably accurate estimation of item parameters is essential if these parameters are to be used for skills-level test design from a pool of items calibrated with the fusion model.

Each item i has one c_i, one π_i^*, and an r_{ik}^* for each required skill k. For the 40-item exam generated for the ideal low complexity model case, there are 40 c_is, 40 π_i^*s, and, as determined by the Q matrix, 84 r_{ik}^*s.

When substantive interpretation of the item parameters is essential (and hence their accurate estimation is needed) and there is concern that noninfluential parameters may confound the values of the remaining parameters, the MCMC performance may be improved by dropping these parameters. This was executed in the current study by employing the stepwise parameter-reduction algorithm. To fit the fusion model to the ideal low complexity dataset, the stepwise algorithm was iterated four times, which resulted in the dropping of noninfluential c parameters at each step. The resulting model had only 27 of the original 40 c parameters. This dropping of a large number of c parameters occurred because the low r^*s were dominating the item responses, thus decreasing the influence of η and, hence, the estimability of the c parameters. A companion consequence of this is that the examinee η values were not well estimated in this dataset. This result, although seemingly worrisome, is actually exactly what one would expect in analyzing a dataset that is generated to rely heavily on the examinee αs and little on the examinee ηs.

Although these dropped c parameters did not seem to strongly influence the item responses in this dataset, these c parameters were indeed used to generate the data. Thus, dropping them could introduce a small amount of bias in the estimated π^*s and r^*s when a model without the c parameter for those items is applied to the data. In particular, π^* for an item would be expected to be slightly underestimated when c is dropped from the model because such estimation bias compensates for the lack of the logistic portion of the model which supplies a factor slightly < 1 in the item response function of the model.

Figure 2 shows the mean ± the estimated standard deviation (from the estimated posterior distribution) plotted against the generating parameter values for the three types of item parameters. The plotted diagonal lines show where a data point would lie when the estimate is exactly equal to the true value.

Because our plotted confidence intervals are only a single standard deviation (sd) wide on each side, we expect to to see approximately 32% of the estimated values fall outside these confidence intervals. Thus, the 15 cases of a true π^* lying outside the mean ± sd bounds in Figure 2 is a little high; but it turns out that 12 of these cases occur for items for which the c parameter was dropped, thus explaining the observed estimation bias. No bias was evident for the r^* parameter estimation, as the number of estimates outside the approximate 68% confidence intervals (26) is actually less than one would expect for 84 hypothesis tests. However, it is still interesting to note that when an estimate falls outside the approximate 68% confidence interval, it is more likely to occur for an item whose c parameter was dropped.

The results of this section have demonstrated that the fusion model system's MCMC estimation method with the stepwise algorithm does result in reasonably accurate recovery of the r^* and π^* item parameters for the ideal case of low-complexity high cognitive structure. In particular, the stepwise algorithm, as expected, dropped a large proportion of the c parameters from the model because they were simulated to have little influence in the generated item responses. Also, this dropping introduced only slight bias (as expected) into the estimation of the π^* parameters, while introducing no noticeable bias in the r^* estimation.

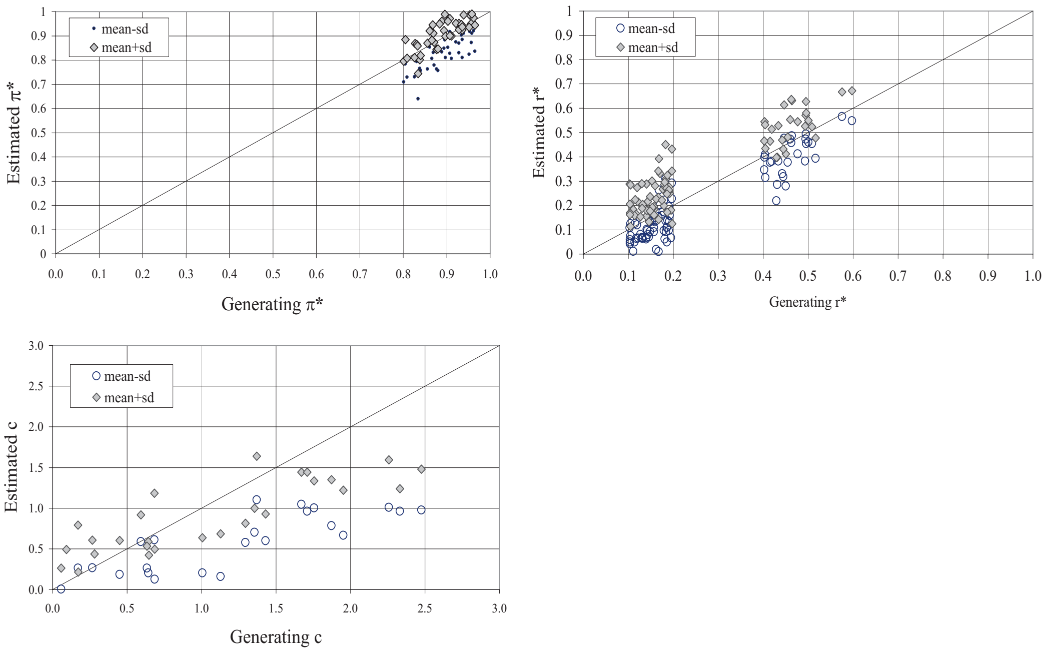

Figure 2. **Estimated 68% confidence intervals versus generated item parameters. The line is $y = x$.**

9.3 Evaluating the Quality of Examinee Parameter Estimation and Examinee Classification

Recall the two examinee parameters in the fusion model are $\underline{\alpha}$, the vector of indicators of skill mastery versus nonmastery, and η, the continuous ability that summarizes skills not specified by the Q matrix. First we evaluate the estimation of η. Since η is a continuous parameter, its estimation can be evaluated in the same manner that the estimation of the item parameters was evaluated. The proportion of times the true η falls within one standard deviation of the posterior mean is 57%, somewhat below the expected 68%. The correlation between the estimated and true η values is 0.80. This degree of accuracy was actually higher than we expected given the weak reliance of the item responses on the ηs.

Next, we evaluate the estimation of $\underline{\alpha}$. Because $\underline{\alpha}$ is a vector of dichotomous variables, the MCMC algorithm produces the estimated probabilities of mastery for an examinee on each α_k. These probabilities can then be used to estimate each α_k, that is, to classify an examinee as either a master ($\alpha_k = 1$) or a nonmaster ($\alpha_k = 0$). There are many

Table 3
Correct Classification Rates (CCRs) for Individual Skills; Ideal Low Complexity Model

Skill	p_k	$\hat{p_k}$	CCR for masters	CCR for nonmasters	Mean CCR	Chance CCR	Cohen's kappa
1	.30	.24	.78	.99	.93	.60	.83
2	.40	.37	.88	.97	.93	.53	.85
3	.45	.36	.73	.94	.85	.51	.69
4	.50	.50	.96	.97	.97	.50	.94
5	.55	.57	.96	.91	.94	.51	.88
6	.60	.59	.93	.92	.93	.52	.85
7	.65	.73	.99	.77	.91	.57	.79

ways to accomplish this. One obvious way is simply to round the posterior probability of mastery to 0 or 1: An examinee is labeled as a master on a skill if his or her posterior probability of mastery on that skill is greater than (or equal to) 0.5; otherwise the examinee is labeled as a nonmaster. Examinees were classified in this manner, and the number of times misclassification occurred was tabulated (there are 7 skills × 1,500 examinees = 10,500 classifications). Of these 10,500 classifications, the examinees were correctly classified 92.2% of the time. Moreover, 62% of the examinees were classified without error on all seven skills; 87% of the examinees had, at most, only one misclassified skill; and 98% had, at most, only two misclassified skills. The individual skill mastery correct classification rates (CCRs), as well as the results for the estimation of p_k, the population proportion of masters for the k^{th} skill, are given in Table 3. The reported p_k estimates were obtained by simply calculating the proportion of examinees on each skill who were estimated to have $\alpha_k = 1$ based on their posterior probabilities of mastery.

The mean absolute difference (MAD) between the p_k estimates and parameters is 0.04, and it is clear that the estimates and parameters are highly correlated with each other. The "chance CCR" given in Table 3 tells the mean CCR that would be expected to occur by chance based on the given p_k parameter and p_k estimate for each skill, specifically: chance CCR $= p_k\hat{p}_k + (1 - p_k)(1 - \hat{p}_k)$.

The "Cohen's kappa" in Table 3 tells the proportion of the distance between chance CCR and perfect CCR (1.00) that is represented by the mean CCR for a skill. The formula for Cohen's kappa is given by: Cohen's kappa $= \frac{\text{CCR}-\text{chance CCR}}{1-\text{chance CCR}}$.

The results show that the mean CCR is 85% or more in every case and, most importantly, is well above the rates that would be expected based on chance classification. Five out of seven of the values of Cohen's kappa are over 80%. In the worst cases, Cohen's kappa was 0.79 for Skill 7 and 0.69 for Skill 3, which was the only skill that had a value substantially lower than 0.80.

Although there is little variation in the number of items that require each skill (see Table 1), and the item parameters were randomly generated, one explanation for the relatively poor classification in Skill 3 seems plausible. When many of the items that require skill a also require a second harder skill b (there may or may not be other skills involved), and the r^*_{ib}s for those items are small (i.e., possessing skill b is very important for correctly answering those items), it is very difficult to determine whether examinees have mastered skill a when they have not mastered skill b. This is because skill b nonmasters have a very low probability of correctly answering any of the items that require Skill a, regardless of their mastery of Skill a. This effect, termed *blocking* by L. DiBello (personal communication, 2005), was first explicated by Samejima (1995). In this case, although the Q matrix was randomly generated to minimize the effect of blocking, 5 of the 11 items that require Skill 3 also require Skill 2, the most overlap that exists in this Q matrix. Since Skill 2 is a harder skill, many examinees are incorrectly responding to those items because they lack Skill 2, thus making it difficult to determine the contribution in the item responses from Skill 3. In fact, three of the five corresponding r^*s for Skill 2 are estimated to be significantly higher than their true value. Because of this blocking, the r^* values for the blocked skill on the problematic items will be estimated with higher values than they should have. This may explain why the Skill 3 CCR is lower than the CCRs for the other skills.

Rather than classifying the examinees by simply rounding the posterior probabilities, another way of classifying examinees is to acknowledge that posterior probabilities near 0.5 are not informative and should not be used for classifying. Instead, examinees are classified

Table 4
CCRs for Individual Skills; Ideal Low Complexity Model

Skill	p_k	Drop rate	CCR for masters	CCR for nonmasters	Mean CCR	Chance CCR	Cohen's kappa
1	.30	.02	.79	.99	.94	.60	.85
2	.40	.04	.90	.98	.95	.53	.89
3	.45	.09	.77	.97	.88	.51	.75
4	.50	.02	.97	.98	.97	.50	.94
5	.55	.06	.97	.94	.96	.51	.92
6	.60	.04	.95	.94	.94	.52	.88
7	.65	.05	.99	.82	.94	.57	.86

Note. Classification based on posterior probability outside of (0.4, 0.6).

as masters if their posterior probability lies above 0.6 and as nonmasters if their posterior probability lies below 0.4. Examinees are not classified at all if their posterior probability of mastery lies between 0.4 and 0.6 (0.4 and 0.6 are arbitrarily specified), indicating there is insufficient information to make a classification. Under this scenario, the overall correct classification rate increases to 94.0% with a 4.6% overall drop rate (the rate at which examinee/skill combinations are left unclassified). Moreover, there is a 22% reduction in the overall rate of incorrect classifications (7.7% to 6.0%). The detailed skill breakdowns are given in Table 4. Note that the two lowest values of CCR in Table 3 for masters and nonmasters (0.73 and 0.77 for Skills 3 and 7, respectively), were substantially increased to 0.77 and 0.82. These results are useful because achieving high classification accuracy is often more important than classifying all examinees. Although the correct classification rate has increased, the problems with estimating Skill 3 are still striking particularly in comparison to other skills; estimation of Skill 3, however, is still quite accurate.

The results presented in this section indicate that, although there is room for improvement, the MCMC estimation method with the stepwise algorithm yields reasonably accurate mastery classification in the case of a low-complexity Q matrix and high cognitive structure, and thus provides a baseline for further evaluation under less ideal conditions.

9.4 Evaluating Model Fit

The many-parameter multidimensional structure introduces the possibility of overfitting the data, while it also enhances the difficulty of evaluating model fit. Thus, it is as important to evaluate the appropriateness of using the fusion model on a dataset from the model-fit perspective as it is to evaluate the estimation of the examinee and item parameters. In order to obtain a better understanding of the fit of the fusion model to a particular dataset, several data-based goodness-of-fit statistics were developed and described above. We now apply these statistics to the simulated data and fitted model of this section of the paper.

We first applied IMstats. Recall that IMstats calculates, for each item, the proportion-correct score for three types of examinees: examinees classified by the MCMC estimation of the fusion model as masters on all the item's required skills (called the "item i masters"), examinees classified as nonmasters on at least one but less than half the item's required skills (called the "item i high nonmasters"), and examinees classified as nonmasters on at least half the item's required skills (called the "item i low nonmasters"). If the fusion model model fit was good for the current data, a strong difference in performance between item masters, item high nonmasters, and item low nonmasters would be expected.

The differences in performance for item masters, item high nonmasters, and item low nonmasters can be discerned in Figure 3 for each item. Because the data are simulated to have strong dependence on skill mastery, it is not surprising that there is a dramatic difference between the item masters and item nonmasters. Over all items, the item masters have an average proportion-correct score of 0.81, the item high nonmasters have an average proportion-correct score of 0.24 and the item low nonmasters have an average proportion-correct score of 0.03. Additionally, for each item individually, the proportion-correct score for the item high nonmasters is many times higher than the proportion-correct score for the low nonmasters.

We next applied EMstats to our simulated data. Recall that for a fixed examinee, j, EMstats first separates the items into four groups, based on the proportion of each item's required skills the examinee has been estimated as having mastered. An examinee's

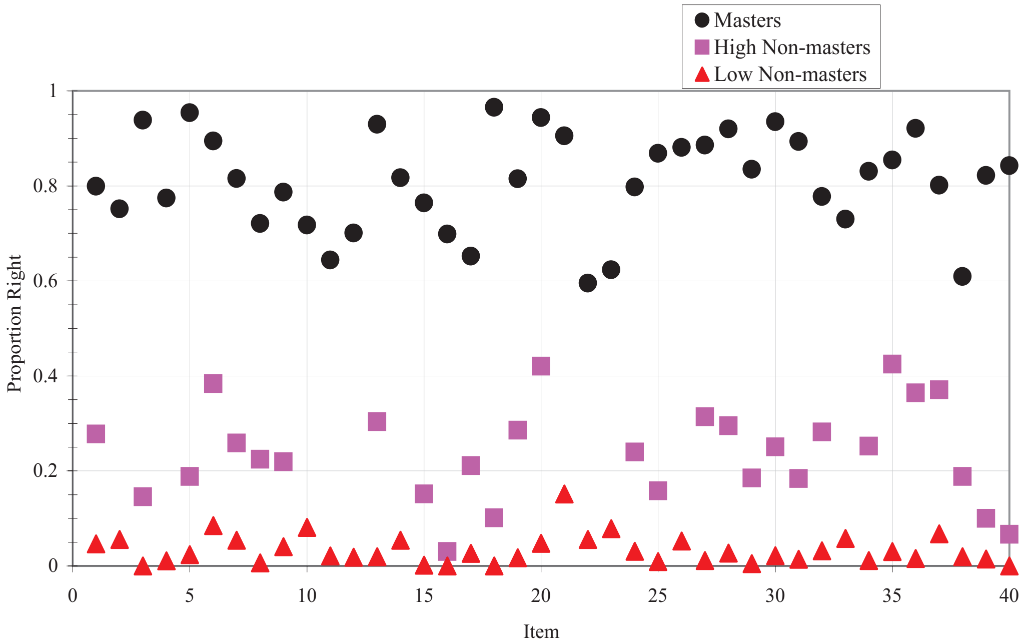

Figure 3. **Item statistics for three groups of examinees: masters, high nonmasters, and low nonmasters.**

mastered items are those for which the examinee has been estimated as having mastered all the Q skills for that item. An examinee's high nonmastered items are those for which the examinee has been estimated as having nonmastered at least one, but not more than half, of the skills required by the item. An examinee's low nonmastered items are those for which the examinee has been estimated as having nonmastered more than half of the skills required by the item. And an examinee's nonmastered items are simply the union of the high nonmastered and low nonmastered items for that examinee. EMstats then calculates the examinee's proportion-correct score on each of these sets of items.

Next, EMstats compares the examinee's observed proportion-correct score for each set of items with a criterion value to see if this score is either unusually low for the set of mastered items or unusually high for any of the sets of nonmastered items. These comparisons are based on hypothesis tests that are conducted only when the number of relevant items in a set is large enough to result in statistical rejection for a true difference of 0.2 or more between the examinee j's observed proportion correct and the criterion score.

Table 5
Output From the Examinee Mastery Statistics Program

		H_0 rejected	
	No. examinees in group	No. examinees	Proportion
Masters	466	28	0.060
Nonmasters	1090	69	0.063
High nonmasters	71	2	0.028
Low nonmasters	833	12	0.014

Note. Examinee j qualifies for a particular category if he or she has enough items to perform the hypothesis test for that category.

Table 5 shows the results of EMstats for the fusion model fit to the simulated data. The observed 6% rejection rate for both masters and nonmasters is close to the 5% Type I error rate, as expected. When the nonmasters are stratified into high nonmasters and low nonmasters, however, the rejection rate decreases to approximately 1%, which is lower than expected.

The results of the IMstats and EMstats analyses provide evidence that the mastery classifications from the Arpeggio MCMC estimation algorithm correspond to large observed score differences between item masters and item nonmasters, and the presence of either unusually low-scoring masters or high-scoring nonmasters appears to be no more prevalent than would be expected by chance. Thus, IMstats and EMstats appear to be performing as expected in the case of the ideal low complexity simulation model.

Finally, we applied FusionStats to our simulated data. The first results presented are a comparison of the observed and estimated examinee scores, with the estimates calculated by using the point estimates for the fusion model examinee and item parameters. Table 6 shows, for each of five categories of observed score range, the mean and standard deviation that result from subtracting the observed examinee proportion-correct scores from their corresponding estimated values. Although there is a tendency toward negative bias (underestimation), the magnitude of the bias is generally small. Indeed, two-thirds of the estimated examinee proportion-correct scores are within 0.04 of the observed values, and 95% of the estimates are within 0.09 of the observed values, with all of the largest errors occuring in the very highest score range. The mean of all the differences between the

Table 6
Comparison of Observed and Estimated Examinee Proportion-Correct Scores

Range of observed score	No. examinees in score range	Estimated − Observed *M*	Estimated − Observed *SD*
0%–19%	551	0.005	0.027
20%–39%	384	−0.002	0.046
40%–59%	237	−0.016	0.059
60%–79%	214	−0.022	0.051
80%–100%	114	−0.045	0.040

estimated and observed scores was -0.007, and the standard deviation of all the differences was 0.045.

The observed bias is believed to be primarily due to the dropped c parameters. The estimation algorithm compensated for the dropped c parameters by underestimating the π^* values (recall that there was no bias in the r^* estimation). It seems that this compensation worked well for examinees with low scores but not as well for examinees with high scores. This makes sense because the dropped $P_{c_i}(\eta_j)$ term is smaller for low-ability examinees and is, thus, well approximated by adjusting π^* downward. By contrast, high-ability examinees have larger $P_{c_i}(\eta_j)$ terms, so dropping that term should be compensated by a much smaller reduction in π^*, which the model cannot do since it must use a single π^* estimate for all examinees. Thus, as ability increases, the estimated scores for high-ability examinees tend to display an increasingly negative bias relative to the corresponding observed values.

Additionally, FusionStats provides a comparison of the observed item proportion correct-scores with the estimated item proportion-correct scores, calculated by using the estimated model parameters. Figure 4 shows a plot of the difference between the estimated and observed item proportion-correct scores. The estimated proportion-correct scores show a consistent underestimation bias, but the size of the bias seems to be very small (the largest difference between observed and predicted values is only about 0.02). Note that the items whose proportion-correct scores are represented by the character × are items whose c parameter was dropped from the analysis. As explained in detail above, the small observed underestimation bias is believed to be due to the dropped c parameters. The bias is small

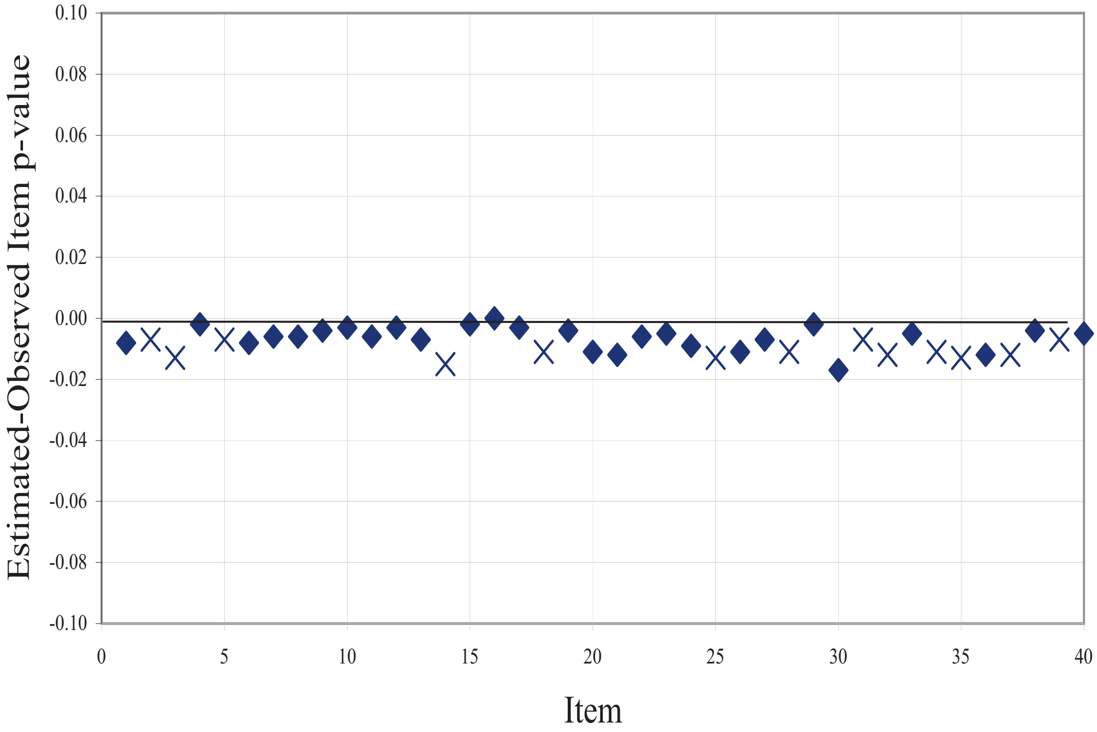

Figure 4. **The difference between the estimated item proportion correct and the observed item proportion correct. The items whose proportion correct are represented by the character × are items whose c parameter was dropped from the analysis.**

because the (η, c) portion of this model is relatively noninfluential for this dataset. Overall, the differences between the estimated and observed item proportion correct scores average -0.007 with a standard deviation of 0.004.

Taken together, the results of IMstats, EMstats, and FusionStats suggest that the model tends to fit the data reasonably well, and that what lack of fit is evident is very small. Even though the simulated data that were analyzed in this section represented the best case of a low-complexity Q matrix with high cognitive structure, these results are important. Because skills-diagnosis modeling and estimation is a complex challenge, it was important to first establish that our methods work reasonably well under the very circumstances for which they were designed to perform best.

10 Simulations Under Nonideal Conditions

It is important to extend our initial evaluation to conditions under which realistic violations of the model assumptions occur. Each of the following simulations weakens one of the assumptions from the ideal case to provide an initial evaluation of the robustness of our skills-diagnosis system.

10.1 More Complex Q Matrix

There is a statistical trade-off between the complexity of the Q matrix (the average number of items per skill, the total number of skills, etc.) and the accuracy of parameter estimation. For a fixed test length, when more skill dimensions are estimated, less power is available for estimating each dimension. Additionally, for a fixed test length and fixed number of skills, there must be a trade-off between the number of Q matrix entries and the accuracy of examinee classification. Unfortunately, the relationship between the accuracy of examinee classification and the number of Q matrix entries is inherently nonlinear. As the number of skills per item decreases (decreased complexity), the number of items per skill also decreases, which may decrease the amount of diagnostic power for the skills. At the same time, as the number of skills per item increases (increased complexity), there is less information per skill in the item, which may result in insufficient power to differentiate between the skills in an item. Thus, the relationship between the number of Q matrix entries and the estimation power of the model must be evaluated on a case-by-case basis.

The effect of a more complex Q matrix on fusion model parameter estimation is now examined. As in the ideal low complexity model, we use seven skills, 40 items, and high cognitive structure item parameters. However, the high complexity model in this section uses a randomly generated Q matrix that requires an average of three skills per item, rather than the average of two skills per item that was used in the ideal low complexity model. For this new high complexity model, the item and examinee parameters were generated using the same distributions as specified in the ideal low complexity case, which were discussed in section 8.1. The item parameters and the structure of the Q matrix are given in Table A2 (see the appendix).

The true item parameters for the high complexity model are compared to the estimated item parameters and their confidence intervals in Figure 5. Applying the stepwise parameter-reduction algorithm, 26 of the estimated cs and 9 of the r^*s did not show strong evidence of significantly influencing the item responses and were, thus, dropped from the model. The decreased influence of the c parameters for the high complexity model is believed to be due to the increased number of simulated low r^* values per item. The increased number of simulated skills per item made the items appear to be more difficult (as compared to the ideal low complexity case), which meant that only the lowest values of c would significantly influence the item responses. The increased number of dropped r^* parameters is believed to be due to increased blocking, though this effect does not seem to be very large.

Of the 111 r^*s that were estimated, 37 had true (generated) values lying outside the mean ± sd. This is very close to the 36 that are expected to lie outside the interval. Despite the increased complexity in the Q matrix, the estimation method appears to successfully recover the true values of the cognitively relevant item parameters with little, if any, bias.

The p_k estimation and CCR results for each of the skills are given in Table 7. On average, the p_k estimation results are a little worse than those for the ideal low complexity case with the MAD increasing to about 0.06 (as compared to 0.04 for the ideal low complexity case). The p_k estimates show an underestimation trend, though they are still highly correlated with the parameters. The small increase in p_k estimation error had no significant affect on the overall average of the mean CCR values and their corresponding Cohen's kappa values, which, on average over all the skills, remained the same as for the ideal low complexity case. Note, however, for the skills having p_k underestimation, that, as would be expected, the CCR for nonmasters was better than that for masters. Although the mean CCRs are good, there is some variability across the skills. This is believed to be caused by increased blocking (explained in section 9.3) in this more complex Q matrix. For instance, 11 of the 13 items requiring Skill 5 are blocked by Skill 2, a harder skill that is also required for these items. This may explain why Skill 5 had a lower mean CCR than for the ideal low complexity case.

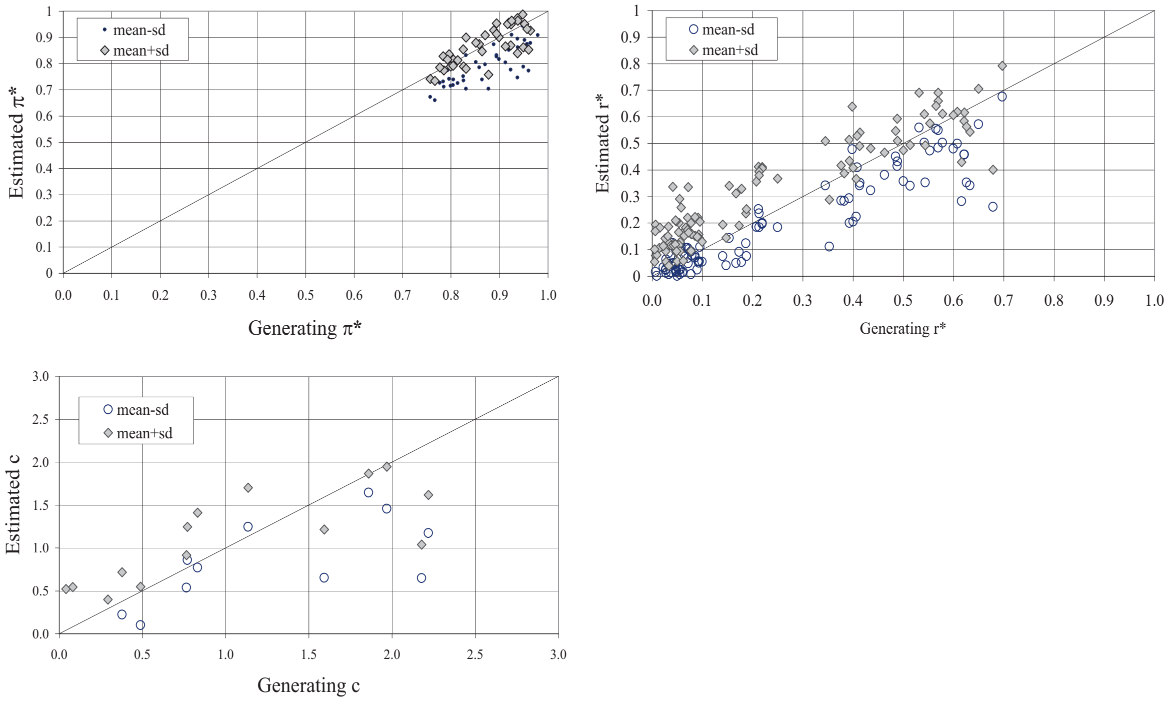

Figure 5. **Estimated 68% confidence intervals versus generated item parameters. The Q matrix requires an average of three skills per item. The line is $y = x$.**

The results in this section demonstrate that the fusion model system MCMC estimation method and stepwise reduction algorithm can be successfully applied to data generated from a more complex Q matrix with only a small proportion of the r^* parameters being dropped, but with the increased number of low r^* values per item causing the c parameters to be even less influential. The vast majority of the r^* values are still well estimated, but the increased number of parameters to be estimated causes some mild degradation of the examinee mastery classification accuracy (as compared to the ideal low complexity case), although the correct classification rates are still moderately high.

10.2 Incorrect Specification of the Q Matrix

In this section we simulate data based on the Ideal Low Complexity model, but use misspecified Q matrices in the estimation algorithm. Specifically, we evaluate the robustness of the parameter estimation algorithm to three serious, though limited in scope, Q matrix misspecifications:

Table 7
CCRs for Individual Skills; High Cognitive Structure, High Complexity Model

Skill	Items per skill	Items used	p_k	$\hat{p_k}$	Drop rate	CCR for masters	CCR for nonmasters	Mean CCR	Chance CCR	Cohen's kappa
1	20	20	.30	.30	.02	.97	.99	.99	.58	.98
2	21	21	.40	.28	.02	.70	1.00	.88	.54	.74
3	14	14	.45	.30	.03	.64	.98	.93	.52	.85
4	16	14	.50	.49	.04	.95	.98	.96	.50	.92
5	13	12	.55	.45	.09	.81	.97	.89	.50	.78
6	20	15	.60	.60	.02	.98	.96	.97	.52	.94
7	16	15	.65	.58	.05	.92	1.00	.95	.52	.89

Note. Classification based on posterior probability outside of (0.4,0.6). "Items used" is the number of Q matrix entries remaining for particular skill after dropping noninfluential r^*s.

1. The specified Q matrix asserts that an eighth skill is important for 10 items, despite the fact that the skill wasn't present in the Q matrix that generated the data.

2. The specified Q matrix is missing a skill that was required by 12 items in the Q matrix that generated the data.

3. The Q specification for a particular item on the test is wrong in the sense that none of the user-specified Q matrix skills for the item were actually used to generate its response.

For these examples, the same ideal low complexity simulated data is used as in sections 8.1–9.4 (high cognitive structure with an average of two skills per item).

10.2.1 Unnecessary skill in the specified Q matrix. An eighth skill was appended to the Q matrix by randomly choosing 10 items (out of 40) to be incorrectly specified as requiring this eighth skill. This eighth skill is specified in addition to the other skills that those 10 items were correctly specified to require in the Q matrix used to generate the simulated data. The remaining 30 items had correctly specified Q matrix entries. From the fusion model equation, one would expect the parameter estimation for the 10 misspecified items to be affected by the added superfluous skill in two ways. First, and most importantly, it is hypothesized that relatively high r^*_{i8}s would be estimated for each item i specified as requiring the unnecessary Skill 8. Such high r^* estimates would be expected to be eliminated by the stepwise reduction algorithm. Secondly, the estimation may compensate for the extra skill by decreasing its reliance on η (i.e., increasing c_i for each item i requiring Skill 8).

The stepwise parameter reduction algorithm was then sequentially applied and resulted in the elimination of all the r^*_{i8}s. This is the ideal result. The parameter estimation of the remaining parameters was nearly identical to the estimation in the correctly specified Q matrix case (see sections 9.2 and 9.3). Besides Skill 8, no other r^*_{ik} parameters were dropped from the analysis. In the earlier ideal low complexity analysis, 13 c parameters were dropped, and in this analysis, 15 c parameters were dropped.

The p_k estimation and CCR results for each of the skills are given in Table 8. The MAD between the p_k estimates and their parameters is about 0.06, as compared to the MAD of 0.04 that occurred for the ideal low complexity case. Interestingly, if the poorer p_k estimation for Skill 6 is ignored, the MAD values for the two cases become essentially the same.

As expected, when a p_k parameter is underestimated, CCR for nonmasters is favored over that for masters; and when a p_k is overestimated, CCR for masters is favored over that of nonmasters. This causes some differences in this case in comparison to the case of ideal low complexity for a few of the skills. However, the mean CCRs for this case, in comparison to the ideal low complexity case, are nearly identical.

This analysis suggests that the fusion model parameter estimation algorithm exhibits robustness to the inclusion of an extra skill in the Q matrix. Neither item parameter estimation nor examinee skill mastery classification was adversely affected. The extra skill itself was labeled as noninfluential and dropped from the analysis.

10.2.2 Necessary skill missing from the specified Q matrix. The second type of Q matrix modeling misspecification simulated is when a skill that is present in the latent structure (in this case, a skill used to generate the data) is entirely missing from the Q matrix. When the Q matrix indicates an item does not depend on a skill that is actually required, one might expect the estimates of the two remaining types of item parameters

Table 8
CCRs for Individual Skills; Ideal Low Complexity Model

			Drop	CCR for	CCR for	Mean	Chance	Cohen's
Skill	p_k	$\hat{p_k}$	rate	masters	nonmasters	CCR	CCR	kappa
1	.30	.24	.03	.82	.99	.94	.60	.85
2	.40	.38	.04	.92	.97	.95	.52	.89
3	.45	.39	.07	.77	.95	.87	.51	.73
4	.50	.51	.02	.98	.97	.98	.50	.96
5	.55	.63	.06	.98	.88	.96	.51	.92
6	.60	.74	.04	.97	.89	.94	.53	.87
7	.65	.58	.05	1.00	.77	.96	.57	.91

Note. Q matrix used in model estimation required an unnecessary eighth skill. Classification based on posterior probability outside of (0.4, 0.6).

(π^* and c) on such an item to decrease as they "soak up" the extra variation that has no correct parameter to absorb it. Thus, one would hypothesize that the corresponding π^*s and cs would be significantly underestimated. In particular, the π^*s may be underestimated because the masters of all the specified required skills may include nonmasters of the missing skill (especially if the missing skill has a lower p_k than the other skills for that item). Likewise, the cs may be underestimated because without the information from the missing skill, η "absorbs" the missing skill and thus becomes more important in influencing item response. Indeed, η is intentionally included in the model to absorb the influence of missing skills.

To test these hypotheses, Skill 5 was arbitrarily chosen to be eliminated from the Q matrix used in fusion model parameter estimation. The one noticeable effect of dropping Skill 5 in the item parameter estimation was on the c parameter estimation. Specifically, all items that were generated using Skill 5 retained their c parameter when Skill 5 was left out of the estimation model, all of which were significantly underestimated. Estimation of the π^*s and r^*s for the items requiring Skill 5 is essentially uninfluenced by the missing skill. This result indicates that it is mainly the c parameter that compensates for the missing skill.

Interestingly, lacking a skill in the specified Q matrix did not decrease the ability to estimate mastery on the other skills that were in the Q matrix. The examinee classification rates are given in Table 9. There are no substantial differences in the classifications on the

Table 9
CCRs for Individual Skills; Ideal Low Complexity Model

			Drop	CCR for	CCR for	Mean	Chance	Cohen's
Skill	p_k	$\hat{p_k}$	rate	masters	nonmasters	CCR	CCR	kappa
1	.30	.23	.02	.78	.99	.94	.61	.85
2	.40	.38	.04	.91	.98	.95	.52	.89
3	.45	.34	.06	.70	.97	.85	.52	.69
4	.50	.49	.02	.96	.98	.97	.50	.94
5	.55	–	–	–	–	–	–	–
6	.60	.61	.07	.96	.92	.94	.52	.87
7	.65	.74	.04	.99	.75	.92	.57	.81

Note. The Q matrix used in the estimation of the model parameters did not include Skill 5. Classification based on posterior probability outside of (0.4, 0.6).

remaining six skills when the model parameters are estimated without the presence of Skill 5, as compared to when the model parameters are estimated with the true Q matrix.

The increased reliance on the c parameter brings up the question of whether the examinee ηs are improved. Autocorrelations for the η estimates using this Q matrix (missing Skill 5) were compared to the corresponding autocorrelations using the full Q matrix. Rather than many autocorrelations remaining above 0.4 at lag 50, as was the case for η when the full Q matrix was used in estimation, the autocorrelations for the η parameters when Skill 5 was missing from the model were predominately at or below 0.2 at lag 50. This constitutes empirical evidence for the conjecture that high autocorrelations do not necessarily imply MCMC non-convergence, but, alternatively, may simply be indicating a relatively flat likelihood surface.

In the previous subsection, it was shown that the fusion model parameter estimation algorithm is robust when an unnecessary skill is added to the Q matrix. This was true because the stepwise parameter reduction algorithm correctly identified the noninfluential item parameters, resulting in accurate estimation of the remaining item and examinee parameters. In this section, the robustness of the fusion model parameter estimation algorithm was evaluated in the case where an important latent skill was not present in the Q matrix specified in the MCMC estimation algorithm. Robustness was exhibited in this case because the inclusion of η in the model allowed the flexibility needed to maintain good fit with a Q matrix missing an important skill.

10.2.3 Incorrectly specified item. The final Q matrix modeling misspecification evaluated in this robustness study is the case in which the hypothesized relationship between an item and its required cognitive skills is seriously incorrectly specified by the Q matrix. In particular, the case was investigated where the item responses for one item were generated assuming that one set of latent skills influences its examinee item responses, and were fit using a completely different set of latent skills (as specified by the Q matrix used in the estimation procedure) for that item. In practice, this corresponds to making a mistake when constructing the row in the Q matrix corresponding to that particular item (the error could be either accidental or based on an imperfect scientific theory). There was no working hypothesis for what would happen in this case.

For simplicity, Item 1 was chosen to be incorrectly specified. In the Q matrix used to estimate the fusion model parameters, it was specified that Item 1 required Skills 1, 6, and 7 rather than Skills 3 and 5, the skills actually used for Item 1 when the data were generated. When the fusion model was estimated for these data, the stepwise reduction algorithm resulted in the dropping of $r^*_{1,1}$ (r^* for Item 1, Skill 1) from the model. The two other incorrectly specified skills for Item 1, Skills 6 and 7, were retained as important for correct item response. Thus the item was misfit. It is interesting to note that Skills 6 and 7 are more correlated with Skills 3 and 5 than is the dropped Skill 1. Skill 6 was correlated 0.17 and 0.71 with Skills 3 and 5, respectively, while the corresponding correlations for Skill 7 were 0.33 and 0.85. Skill 1, on the other hand, had correlations of only 0.13 with both Skills 3 and 5.

The results showed that the accuracy of the estimation of the item parameters for the remaining items was conserved. During the stepwise reduction of the model, 11 of the c parameters were dropped from the analysis, comparable to the 13 cs dropped in the initial analysis of the fit of the model to the data using the correct Q matrix (see section 9.2). Item parameter estimation for all items except for Item 1 was very similar to the item parameter estimation using the correct Q matrix in section 9.2. Likewise, the accuracy of examinee classification remained high. The correct classification rates for the individual skills are essentially identical to those shown earlier when the true Q matrix was used to fit these data in section 9.3. Hence the fusion model system displayed robustness to a severe mistake for one item in the Q matrix in that the item parameter estimation for the remaining items and the examinee classification rates are as accurate as when the mistake was not present.

Since two of the incorrectly specified skills required by Item 1 are retained in the final model, the fit of the model to Item 1 is of particular interest. Figure 6 gives the performance for the masters, high nonmasters, and low nonmasters of each item. The relationship between the three categories appears weaker for Item 1 (compared to the other items and compared to the performance of Item 1 in Figure 3 when the correct Q matrix was used), a result consistent with the incorrect skill specification for the item. This relationship is presented more clearly in Figure 7. This plot compares two ratios: the ratio of the performance of masters to the performance of high nonmasters, and the ratio of the performance of high nonmasters to the performance of low nonmasters. Each ratio is a measure of the discriminating power of the item. For example, the data point in the upper right corner is for Item 16, indicating that it is the best discriminating item on the test. This is primarily due to it having the lowest proportion-correct score for high nonmasters, as is evident in both Figure 3 and Figure 6. From this perspective, the discriminating power of Item 1 is the weakest of all the items. Because the skills are positively correlated with one another, however, information from Skills 6 and 7 is used in a reasonable approximation to the item response function.

In summary, despite the fact that an item was grossly misspecified in the Q matrix, correct examinee classification rates on the skills did not decrease and item parameters were well estimated. Further, the misspecified item could be identified as distinct when looking at item-level goodness-of-fit evaluations. Of course, if many items were misspecified in this manner it would undoubtedly adversely affect both examinee classification and parameter estimation. Still, it was important to demonstrate that the item parameter estimation and

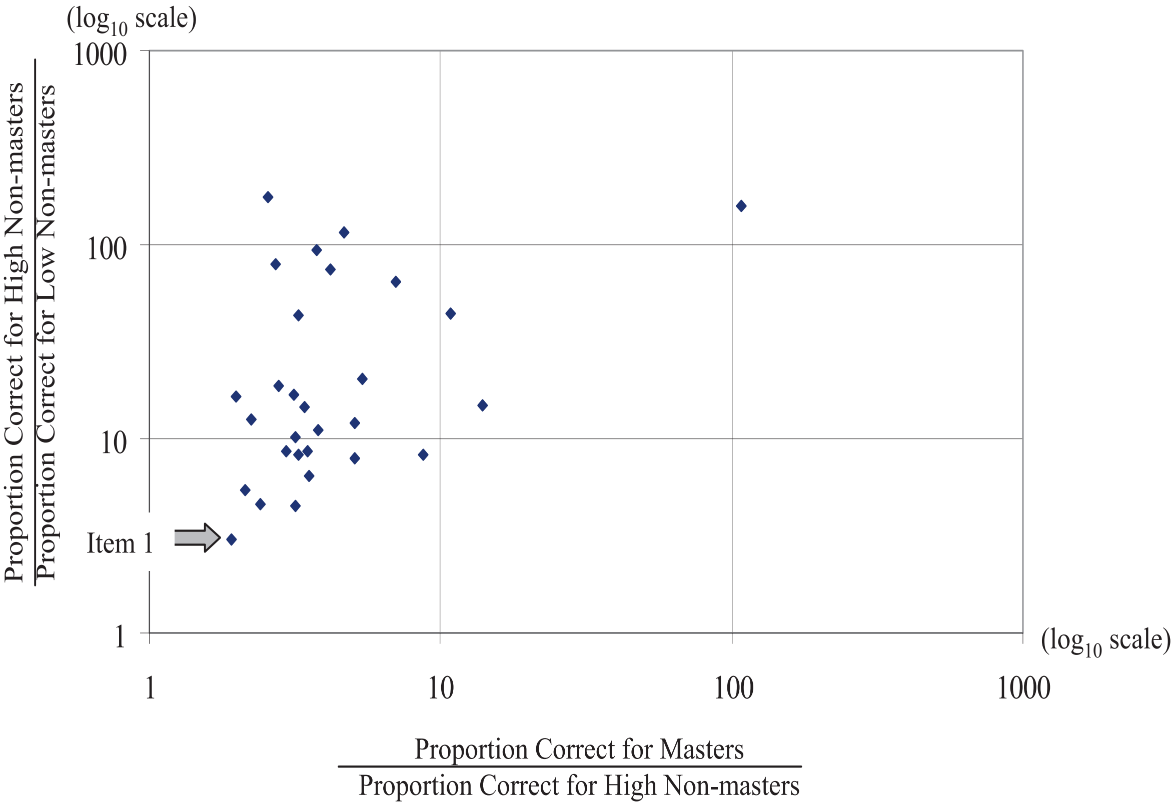

Figure 6. **Item statistics for three groups of examinees: masters, high nonmasters, and low nonmasters. The skills for Item 1 were incorrectly specified.**

examinee classification of the fusion model system is robust to more minor manifestations of misspecified items.

11 Summary and Conclusions

The purpose of the simulation studies in this paper was to evaluate the effectiveness of the fusion model skills diagnosis system under ideal (high cognitive structure and low complexity) conditions, and then to explore the effectiveness and robustness of fusion model parameter estimation and examinee classification under various realistic perturbations of these ideal conditions. The fusion model estimation algorithm was shown to be effective under ideal, high cognitive conditions in which the r^* parameters were selected to be fairly low (indicating a high reliance on mastery of the Q-matrix specified skills for correct item responses) and the Q matrix required only two skills per item, on average. Examinee

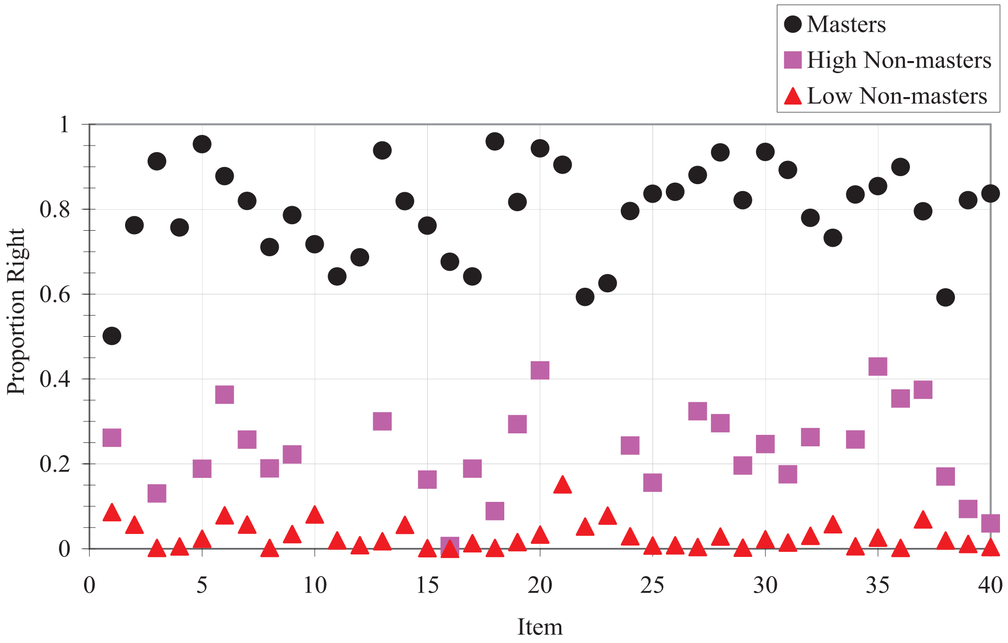

Figure 7. **Discriminating powers of the items by comparing two ratios: the ratio of the performance of masters to that of high nonmasters and the ratio of the performance of high nonmasters to that of low nonmasters. The skills for Item 1 were incorrectly specified.**

skill-mastery classification rates were high, averaging 94% with only 4.6% of the examinees dropped using the (0.4, 0.6) nonclassification interval.

Examinee classification and item parameter estimation remained reasonably accurate when the model generating the data was made less ideal in several realistic ways. First, the Q matrix was made more complex by increasing the average number of skills per item from two to three, as expected; the correct examinee classification rate was decreased to 92% due to blocking; and the drop rate was relatively unchanged at 4%. These results are consistent with prior hypotheses. Next, the robustness of the parameter estimation procedure and the stepwise algorithm was tested in a series of three studies where the Q matrix given to the estimation algorithm differed from the Q matrix used to generate the data. In the first robustness study, a superfluous skill was added to the specified Q matrix used to estimate 10 randomly chosen items. In this case, the stepwise algorithm eliminated the superfluous skill entirely from the Q matrix, and the accuracy of parameter estimation and examinee classification was maintained. In the second robustness study, a skill was missing from the Q matrix used in the fusion model estimation algorithm, although it was present in the Q matrix used to generate the data. In this case, as predicted because of the role of the η examinee parameter in the model for absorbing the influence of any missing skills, the c item parameters were significantly underestimated for the items that truly relied on the missing skill. The remaining item parameter estimates and examinee classifications were highly accurate. Note that the exhibited robustness relied on the incorporation of η and c in the model, features not present in any other skills-diagnosis model. In the final simulation study, a single item was completely misspecified in the Q matrix used in the parameter-estimation algorithm: The skills it was said to rely on were not used in data generation, and the skills used in the data generation were not said to be important in the parameter-estimation algorithm. Except for the misspecified item itself, this misspecification did not adversely influence the estimation of the remaining item parameters. In addition, it did not adversely affect examinee classification. Moreover, the misspecified item appeared as the worst performing item when using item level goodness-of-fit approaches.

Overall, the results of our simulation studies have verified that all the components of the fusion model skills diagnosis system — the MCMC estimation method, the stepwise algorithm model-fitting procedure, and the model-fit evaluation procedures — behaved appropriately for a dataset simulated under the conditions of high cognitive structure for both low-complexity and high-complexity error-free Q matrices. Furthermore, the robustness studies indicated that in the case of high cognitive structure and a low-complexity Q matrix, when a limited number of serious misspecifications are introduced in the Q matrix, the performance of the fusion model system exhibits strong robustness in terms of the accuracy of item-parameter estimation and examinee classification.

The success of this initial study establishes the fusion model system as a promising tool for skills diagnosis. The research presented in this paper has already been, and will continue to be, extended in a variety of simulation studies and real data analyses, the results of which will soon be submitted for publication. In terms of simulation studies, the fusion model system research has already been extended to include low cognitive structure models (higher values of r^* than were simulated in this paper), extended robustness studies that include both new and stronger violations of the model assumptions, development and evaluation of equating methods, a more in-depth evaluation of the effectiveness of the stepwise-reduction algorithm in improving model-fit and parameter estimation, development and evaluation of model extensions to handle polytomously scored skills and polytomously scored items, and an improved MCMC algorithm to increase parameter estimation accuracy. In terms of real data analyses, a number of applications have also already been completed, including applications to large-scale standardized tests. Indeed, overcoming the special hurdles inherent in real data analyses have required further enhancements to the fusion model system that will be discussed in forthcoming papers. Such enhancements have included the development of the theory and estimation of reliability with respect to skills diagnosis, methods for better developing Q matrices from both statistical and substantive considerations, ways to set reasonable values for the p_k parameters when such values are indeterminant from a likelihood perspective, and enhanced methods for estimating individual skill mastery (given the MCMC item-parameter estimates) as well as for estimating the distribution of skill mastery in the population. Of course, all these enhancements, which have been motivated by real data analyses, have been developed and evaluated in simulation studies.[2]

References

Bolt, D. M. (1999, April). *Applications of an IRT mixture model for cognitive diagnosis.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Carlin, B., Xia, H., Devine, O., Tollbert, P., & Mulholland, J. (1999). Spatio-temporal hierarchical models for analyzing Atlanta pediatric asthma ER visit rates. In C. Gatsonis, R. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli, & M. West (Eds.), *Case studies in Bayesian statistics, Volume IV* (pp. 303–320). New York, NY: Springer-Verlag.

DiBello, L. V., Stout, W. F., & Jiang, H. (1998). *A multidimensional IRT model for practical cognitive diagnosis.* Unpublished manuscript. ETS, Princeton, NJ.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Mahwah, NJ: Erlbaum.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2,* 217–233.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186.

Embretson, S. E. (1999), Cognitive psychology applied to testing. In F. Durso, R. Nickerson, R. Schvaneveldt, S. Dumais, D. Lindsay, & M. Chi (Eds.), *Handbook of applied cognition* (pp. 629–660). Hoboken, NJ: John Wiley & Sons.

Fischer, G. H. (1973), The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

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

Gitomer, D. H., & Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. *Journal of Educational Measurement, 28,* 173–189.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 333–352.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory.* New York, NY: Kluwer-Nijhoff.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation. Champaign, IL: University of Illinois.

Jiang, H. (1996). *Applications of computational statistics in cognitive diagnosis and IRT modeling.* Unpublished doctoral dissertation. Champaign, IL: University of Illinois.

Junker, B. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Unpublished manuscript. Pittsburgh, PA: Carnegie Mellon University. Prepared for the Committee on the Foundations of Assessment, National Research Council.

Junker, B. (2000). *Some topics in nonparametric and parametric IRT, with some thoughts about the future.* Unpublished manuscript. Pittsburgh, PA: Carnegie Mellon University.

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

Mislevy, R. J. (1994), Evidence and inference in educational assessment. *Psychometrika, 59,* 439–483.

Patz, R., & Junker, B. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

Patz, R., & Junker, B. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342–366.

Raftery, A. E., & Lewis, S. M. (1996). Implementing MCMC. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 115–130). London: Chapman & Hall, Ltd.

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 979–1030). Amsterdam: Elsevier.

Samejima, F. (1994, April). *Cognitive diagnosis using latent trait models.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Samejima, F. (1995). A cognitive diagnosis method using latent trait models: Competency space approach and its relationship with DiBello and Stout's unified cognitive-psychometric diagnosis model. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 391–410). Mahwah, NJ: Earlbaum.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and dignostic assessment. *Journal of Educational Measurement, 34,* 333–352.

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

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215–231.

U.S. Department of Education. (2002). *Draft regulations to implement Part A of Title I of the Elementary Secondary Education Act of 1965 as ammended by the No Child Left Behind Act of 2001.* Washington, DC: U.S. Government Printing Office.

U.S. House of Representatives. (2001). *Text of No Child Left Behind Act.* Washington, DC: U.S. Government Printing Office.

Whitley, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479–494.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment* (ETS Research Report No. RR-03-32). Princeton, NJ: ETS.

Notes

[1] The research reported here was completed under the auspices of the External Diagnostic Research Team, supported by ETS.

[2] The fusion model skills diagnosis system is owned by ETS. Access to the system is available through ETS.

Appendix: Q Matrices and Item Parameters

Table A.1

Randomly Generated Item Parameters for a 40-Item, 7-Skill Exam; Ideal Low Complexity Model

	π^*	r_1^*	r_2^*	r_3^*	r_4^*	r_5^*	r_6^*	r_7^*	c
Item 1	0.869			0.447		0.197			1.128
Item 2	0.834	0.146							0.156
Item 3	0.936	0.158				0.122			1.796
Item 4	0.896							0.13	1.295
Item 5	0.957	0.177			0.157				2.06
Item 6	0.889		0.494			0.442		0.184	2.476
Item 7	0.827			0.403	0.405			0.111	1.951
Item 8	0.805			0.464		0.132			0.647
Item 9	0.894			0.493	0.431	0.171			0.45
Item 10	0.871				0.153				0.684
Item 11	0.861						0.118		0.281
Item 12	0.907							0.104	0.642
Item 13	0.953	0.575			0.167		0.463	0.14	1.872
Item 14	0.838				0.105				2.164
Item 15	0.965		0.114	0.197					0.093
Item 16	0.884					0.104		0.198	0.686
Item 17	0.831					0.477		0.153	0.633
Item 18	0.952	0.142		0.43			0.103	0.162	2.406
Item 19	0.928	0.516					0.179		0.171
Item 20	0.962	0.509					0.12		1.371
Item 21	0.921		0.18						1.671
Item 22	0.911						0.191		0.057
Item 23	0.856				0.192				0.267
Item 24	0.833				0.415		0.183		1.709
Item 25	0.877		0.115			0.451	0.143		1.67
Item 26	0.934					0.148			1.755
Item 27	0.939						0.495	0.185	2.331
Item 28	0.928		0.445	0.183	0.195				1.013
Item 29	0.903		0.402	0.496		0.104		0.181	1.006
Item 30	0.958			0.419				0.159	1.429
Item 31	0.897	0.191			0.14		0.187		1.831
Item 32	0.801		0.455		0.186				1.819
Item 33	0.907				0.157				0.593
Item 34	0.84		0.141	0.404			0.168		1.108
Item 35	0.88	0.598	0.127	0.168					2.336
Item 36	0.936		0.46			0.149			2.258
Item 37	0.809	0.501			0.175				2.43
Item 38	0.866					0.434	0.167		0.172
Item 39	0.826	0.13					0.105	0.13	2.329
Item 40	0.868						0.19	0.186	1.356

Table A.2

Randomly Generated Item Parameters for a 40-Item, 7-Skill Exam; High Complexity Model

	π^*	r_1^*	r_2^*	r_3^*	r_4^*	r_5^*	r_6^*	r_7^*	c
Item 1	0.885	0.621				0.25	0.01	0.027	0.449
Item 2	0.861	0.028	0.544		0.33	0.06		0.007	1.016
Item 3	0.844						0.056		2.219
Item 4	0.919		0.047			0.058	0.167		1.169
Item 5	0.927	0.633	0.007	0.406	0.053	0.27			0.202
Item 6	0.828	0.622	0.041				0.14		0.651
Item 7	0.866		0.553			0.211			0.489
Item 8	0.895					0.055	0.147	0.005	0.041
Item 9	0.847		0.067		0.345				0.377
Item 10	0.885		0.091			0.219	0.029	0.072	1.008
Item 11	0.825				0.07				2.178
Item 12	0.822			0.463				0.053	0.888
Item 13	0.932	0.697		0.485				0.027	1.759
Item 14	0.887				0.382				0.765
Item 15	0.907								1.86
Item 16	0.954		0.569			0.219	0.07	0.093	0.081
Item 17	0.831	0.626	0.095	0.03			0.079		1.079
Item 18	0.859		0.569	0.053	0.009		0.13		0.292
Item 19	0.861	0.078	0.513	0.413	0.353		0.144	0.015	0.284
Item 20	0.932	0.679	0.085		0.4		0.052	0.026	0.025
Item 21	0.925	0.05							0.586
Item 22	0.932	0.033			0.047		0.028		2.43
Item 23	0.945	0.608	0.599			0.207		0.037	0.799
Item 24	0.873	0.055		0.031			0.006		0.648
Item 25	0.967		0.565					0.064	1.136
Item 26	0.918	0.616	0.079	0.413			0.048		0.018
Item 27	0.875		0.542		0.043				1.592
Item 28	0.932	0.014		0.077			0.178		2.158
Item 29	0.949						0.173		1.969
Item 30	0.956	0.062		0.033	0.398				1.272
Item 31	0.914			0.488			0.186		1.977
Item 32	0.973	0.094	0.5			0.213			2.163
Item 33	0.876		0.073		0.393	0.213	0.154		0.582
Item 34	0.886							0.046	0.77
Item 35	0.931	0.65			0.376		0.188	0.027	1.96
Item 36	0.812		0.578				0.049		0.831
Item 37	0.81	0.09		0.435	0.077				0.268
Item 38	0.83	0.023		0.488	0.392			0.085	0.527
Item 39	0.835	0.047	0.1		0.304	0.005		0.088	2.133
Item 40	0.89		0.531	0.408		0.041		0.022	2.275