1.0 INTRODUCTION

There is a long history of calls for combining cognitive science and psychometrics (Cronbach, 1975; Snow & Lohman, 1989). The U.S. standards movement, begun more than 20 years ago (McKnight et al., 1987; National Council of Teachers of Mathematics, 1989), sought to articulate public standards for learning that would define and promote successful performance by all students; establish a common base for curriculum development and instructional practice; and provide a foundation for measuring progress for students, teachers, and programs. The standards movement provided the first widespread call for assessment systems that directly support learning. For success, such systems must satisfy a number of conditions having to do with cognitive-science–based design, psychometrics, and implementation. This chapter focuses on the psychometric aspects of one particular system that builds on a carefully designed test and a user-selected set of relevant skills measured by that test to assess student mastery of each of the chosen skills. This type of test-based skills level assessment is called skills diagnosis. The system that the chapter describes in detail is called the Fusion Model system.

This chapter focuses on the statistical and psychometric aspects of the Fusion Model system, with skills diagnosis researchers and practitioners in mind who may be interested in working with this system. We view the statistical and psychometric aspects as situated within a comprehensive framework for diagnostic assessment test design and implementation. We use the term “skills diagnostic tests” to refer to tests that provide profiles of information on examinees: information
that is richer and more useful than a single score; information that is scientifically based and statistically sound; and information that can be effectively and readily acted on by students, parents, teachers, and others to improve teaching and learning. This work is based on the premise that periodic skills diagnostic testing, tuned to classroom needs, can greatly enhance teaching and learning, and can promote higher quality and greater accountability within educational and training programs. The Fusion Model system described here is an effective psychometric and statistical approach consisting of procedures and software for practical skills diagnostic testing.

In response to the assessment demands imposed by the standards movement, test manufacturers have redoubled their efforts to expand their test results beyond the reporting of standard global scores of student achievement in broad subject areas to more diagnostic scores indicating academic strengths and weaknesses on specific skills. Perhaps the best example in this regard is the recent development by Educational Testing Service (ETS) of a Score Report Plus\textsuperscript{TM} for each student who takes the College Board’s PSAT/NMSQT\textsuperscript{TM} (Preliminary SAT/National Merit Scholarship Qualifying Test). In addition to the broad-based score report given in the past, the Score Report Plus includes a list of up to three important skills needed to perform well on the test and for which there is strong diagnostic information of lack of examinee mastery.

To provide the testing industry with practical tools for expanding standardized testing to include effective skills diagnosis, a research program initiated by the foundational modeling work of DiBello, Stout, and Roussos (1995); the subsequent developmental work of Hartz and Roussos (Hartz, 2002; Hartz & Roussos, in press); and recent ETS-supported research on refining these developments has resulted in a comprehensive skills diagnosis system, called the Fusion Model skills diagnosis system or, simply, the Fusion Model system.

The Fusion Model system includes the following four components:

1. An identifiable and interpretable item response function model, a reparameterization of the foundational Unified Model (DiBello et al., 1995)
2. A parameter estimation method referred to as Arpeggio, which employs a Markov Chain Monte Carlo (MCMC) algorithm within a Bayesian modeling framework for model estimation, including item parameter estimation and ability distribution parameter estimation
3. A collection of model checking procedures, including statistical MCMC convergence checking, ability distribution and item parameter estimates with standard errors, model fit statistics, internal validity statistics, and reliability estimation methods.

4. Skills-level score statistics, including mastery/nonmastery estimation, subscore options for assessing mastery/nonmastery, and proficiency scaling statistics that relate test scores to skill mastery.

This chapter provides a detailed description of each component, with particular focus on what a skills diagnostic practitioner needs to know.

Skills diagnosis, sometimes referred to as skills assessment, skills profiling, profile scoring, or cognitive diagnosis, is an application of psychometric theory and methods to the statistically rigorous process of (a) evaluating each examinee on the basis of level of competence on a user-developed array of skills, and (b) evaluating the skills estimation effectiveness of a test by assessing the strength of the relationship between the individual skills profiles and the observed performance on individual test items. Such examinee evaluation, ideally periodic throughout the teaching and learning process, can provide valuable information to enhance teaching and learning. It is primarily intended for use as formative assessment, in contrast to summative assessment, in that formative skills diagnosis is intended by its design and implementation to be used directly to improve learning and teaching. In contrast, summative assessment records examinee status at some point in time, usually at the end of an instructional unit, without any direct attempt to improve the examinees status on what was assessed.

Skills diagnosis can often be considered relatively low stakes from the perspective of the individual test taker. It affects decisions that have relatively short time horizons, such as deciding which problems to assign for tonight’s homework or planning for tomorrow’s lesson, as opposed to large life events, such as admission to college or selection for employment. The skills diagnostic score results, at least in principle, are well integrated within the classroom environment and constitute just one complementary element of an extensive store of teacher, peer, assessment, and instructional information. Of course, the distinction between formative and summative assessment is not sharp, and much effort is being devoted to aligning formative and summative assessments, and developing broad-based assessment systems that include both summative and formative components (The Commission on Instructionally
Supportive Assessment, 2001). Because unidimensionally scaled assessment scores alone provide a relatively impoverished global overview of competence in any domain, such broad-based systems will likely rest on strategically conceived and well-designed skills diagnostic assessments.

There are two common types of applications of cognitive diagnostic methods. In some settings, it is important to revisit existing summative assessments and their data and apply diagnostic methods in a post-hoc attempt to extract some diagnostic information from these assessments (Birenbaum, Tatsuoka, & Yamada, 2004; Tatsuoka, Corter, & Guerrero, 2003; Tatsuoka, Corter, & Tatsuoka, 2004; Xin, 2004; Xin, Xu, & Tatsuoka, 2004). A related application is to employ diagnostic assessment models to existing assessment data in order to determine structural information about which skills underlie good performance on the assessment (DiBello et al., 2006). In contrast, the second type of application is on assessments that are specifically designed for the purpose of providing skills diagnosis. Most applications have been of the first kind, and a particularly well-done example is a study by Jang (2005) in which she applied the Fusion Model system to two forms of the ETS LanguEdge English Language Learning (ELL) assessment using data from a large-scale administration (about 1350 examinees per form) as well as data from a classroom setting (27 students). The LanguEdge was developed by ETS as a prototype for the latest version of its Test of English as a Foreign Language (TOEFL), but Jang defined a new classroom formative assessment purpose for the test and subsequently built a skills diagnosis framework for it to extract diagnostic information that was used by teachers and learners in a university-based summer ELL program. We refer to this particular example often throughout the chapter.

Instead of assigning a unidimensionally scaled ability estimate to each examinee as done in typical item response theory (IRT), model-based summative assessments, skills diagnosis model-based formative assessments partition the latent space into sufficiently fine-grained, often discrete or even dichotomous, skills and evaluate the examinee with respect to his or her level of competence on each skill. The term “skill” is used in a generic sense here and can refer to any relatively fine-grained entity associated with the examinee that influences item performance in the general subject area being tested. For example, suppose designers of an algebra test are interested in assessing student proficiencies with respect to a standard set of algebra skills (e.g., factoring, using the laws of exponents, solving quadratic equations, etc.). A skills diagnosis–based analysis attempts to evaluate each examinee with respect to each skill, whereas a standard unidimensional psychometric
The Fusion Model Skills Diagnosis System

analysis typically evaluates each examinee only with respect to an overall scaled score on the algebra exam. We also note here that the decision about whether to use a set of user-defined skills or a unidimensional scale score in a given assessment depends largely on the intended purpose of the assessment. If the purpose of the algebra assessment described previously is only to separate the class into two groups – the group of students who are lagging behind, and all other students – then the assessment and reporting of individual skill competency levels is likely a distraction and an impediment. For such an assessment purpose, the best type of assessment would likely be a unidimensionally scaled assessment that reliably ranks examinees along a broad competency scale, representing in this case overall algebra ability.

Before we describe the details of the Fusion Model system, it is helpful to place the system within the context of a more general framework for skills diagnosis. Roussos, DiBello, Henson, Jang, & Templin (in press; see also Roussos, DiBello, & Stout, in press) described a framework for the entire implementation process for diagnostic assessment, elaborating on practical aspects of cognitively based assessment design and illuminating practical issues in regard to estimation and score reporting.

In particular, they described the diagnostic assessment implementation process as involving the following six steps:

1. Description of assessment purpose
2. Description of a model for the skills space
3. Development and analysis of the assessment tasks (e.g., test items)
4. Selection of an appropriate psychometric model for linking observable performance on tasks to the latent skill variables
5. Selection of statistical methods for model estimation and checking
6. Development of systems for reporting assessment results to examinees, teachers, and others

As discussed in Roussos, DiBello, Henson, et al. (in press), the steps of a successful implementation process are necessarily nonlinear, requiring considerable interaction and feedback between the steps and demanding close collaboration between users, test designers, cognitive psychologists, and psychometricians. This chapter focuses on the Fusion Model system, which provides components for implementation steps 4 to 6. Specifically, the Fusion Model item response function is a model that can be selected in step 4 of the implementation process, the Arpeggio parameter estimation and model checking procedures could be employed in
step 5, and the Fusion Model system skills-level score statistics are useful for developing score reports in step 6 of the implementation process.

The first section of the chapter reviews the reparameterized unified model (RUM), which is the item response function (IRF) that is used in the Fusion Model system. The next section reviews the Monte Carlo Markov Chain (MCMC) estimation procedure. The following section describes the variety of model checking statistical techniques provided by the Fusion Model system. The next section describes the last component of the Fusion Model system, the statistical techniques that have been developed for use in score reporting. The last section of the chapter summarizes the system, where to obtain more information, and future areas of research and development. Throughout the chapter, our focus is on describing the practical statistical aspects of the system so someone with test data would be able to apply the system intelligently and know how to interpret the results, regardless of whether the application is a post-hoc analysis of existing summative test data or embedded within a full diagnostic framework beginning with the design of the diagnostic instrument. We emphasize that the statistically focused knowledge presented here is insufficient by itself for carrying out skills diagnosis. In general, all steps of the implementation process described previously are required for successful implementation, which includes not only statistical analyses, but also substantive analyses and the interaction of the two. Readers are referred to Roussos, DiBello, Henson, et al. (in press) and Roussos, DiBello, and Stout (in press) for more details on the implementation process as a whole.

2.0 THE FUSION MODEL

Item response theory (IRT) provides a highly successful and widely applied probabilistic modeling approach to assessment analysis and scoring that is based on an item response function (IRF) model. IRT-based cognitive diagnosis models, like all IRT models, define the probability of observing a particular response by examinee $j$ to item $i$ in terms of examinee ability parameters and item parameters. Symbolically, this probability is represented as $P(X_{ij} = x | \theta_j, \beta_i)$, where $X_{ij} = x$ is the response of examinee $j$ to item $i$, $\theta_j$ is a vector of examinee $j$ ability parameters, and $\beta_i$ is a vector of item $i$ parameters. In this chapter, for convenience, we restrict $x$ to be dichotomous, indicating a correct ($x = 1$) or incorrect ($x = 0$) response, but the Fusion Model can handle the polytomous case as well.
The Fusion Model Skills Diagnosis System

The fundamental assumption of IRT modeling is that, conditioned on the examinee ability parameters, an examinee response to any item \( i \) is independent of the examinee response to any other item \( i' \). The distinguishing feature of cognitive diagnosis models from other IRT models is that the items \( i = 1, \ldots, I \) relate to a prespecified, user-selected set of cognitive skills \( k = 1, \ldots, K \) that are of particular interest to the skills diagnosis. This relationship is referred to as the Q matrix, where \( q_{ik} = 1 \) indicates that skill \( k \) is required by item \( i \) and \( q_{ik} = 0 \) indicates that skill \( k \) is not required by item \( i \). The Q matrix notation was first introduced by Tatsuoka (1990), whose work highlighted the scientific development of the Q matrix as a critical component of skills diagnosis.

The Unified Model (DiBello et al., 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 \( \theta = (\alpha_Q, \alpha_b) \) (examinee subscript \( j \) suppressed) denote the complete latent space of all relevant skills. Let \( \alpha_Q \) be the vector of prespecified cognitive skills as denoted in the Q matrix. The remaining latent space, \( \alpha_b \), includes the relevant skills supplementary to those specified by the Q matrix. Samejima (1995) referred to \( \alpha_b \) as skills associated with “higher-order processing” and suggested that these skills may be more substantively important than \( \alpha_Q \). From the Unified Model perspective, however, \( \alpha_b \) does not need to be interpreted as higher-order processing; it is just a parametrically simple representation of the latent skills influencing examinee task performance that lie outside the user-specified Q matrix.

The Unified Model parameterizes ability for examinee \( j \) as \( (\alpha_j, \eta_j) \), where \( \alpha_j \) is a vector of skill mastery parameters corresponding one to one with \( \alpha_Q \), and \( \eta_j \) is conceptualized as a unidimensional composite of the elements of \( \alpha_b \) (a projection of \( \alpha_b \) onto a unidimensional scale). We call this \( \eta_j \) term the supplemental ability. The inclusion of this supplemental ability \( \eta_j \) in the Unified Model is 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 skill requirements for every item on the test differentiates the Unified Model from other skills diagnosis models.

For this chapter, we assume for convenience that the elements of \( \alpha_j \) are dichotomous, that is, \( \alpha_{kj} = 1 \) if examinee \( j \) has mastered skill \( k \), and \( \alpha_{kj} = 0 \) if examinee \( j \) has not mastered skill \( k \). In general, the Fusion Model
can also handle more than two categories of skill mastery. On the other hand, \( \eta_j \) is modeled as a standard normal continuous variable. The ability distribution for the \( \alpha_j \) parameters is assumed to be well modeled by its first- and second-order moments, namely, (a) the proportion of the population that has mastered each skill \( k \) – this proportion is denoted by \( p_k \), \( k = 1, \ldots, K \), and (b) the correlations between the \( K \) skill components of \( \alpha_j \). To facilitate better interpretability, tetrachoric correlations will be used instead of correlations to model the relationships between the dichotomous mastery variables.

Define \( \pi_{ik} = P(Y_{ijk} = 1 | \alpha_{kj} = 1) \) and \( r_{ik} = P(Y_{ijk} = 1 | \alpha_{kj} = 0) \), where \( Y_{ijk} = 1 \) refers to the unobserved event that examinee \( j \) correctly applies skill \( k \) to item \( i \). (We restrict ourselves here to the case of dichotomous skill mastery, even though, as stated previously, the Fusion Model is also capable of handling more than two levels of skill mastery. Use of the latent variable \( Y_{ijk} \) is also used in the latent response models discussed by Maris, 1999.) The IRF for the Unified Model is given in Equation 10.1:

\[
P(X_i = 1 | \alpha_j, \eta_j) = d_i \prod_{k=1}^{K} \pi_{ik}^{\alpha_{ik}} r_{ik}^{(1-\alpha_{ik})} + (1-d_i) P_h(\eta_j)
\]

(10.1)

where \( P_h(\eta_j) = (1 + \exp[-1.7(\eta_j + h)])^{1/2} \), a Rasch model (Rasch, 1961) with the difficulty parameter equal to the negative of \( h \), where \( h \) stands for either \( c_i \) in the first term or \( b_i \) in the second term.

The product term in the model indicates two important aspects of the model. First is the statistical assumption of conditional independence of applying the skills, provided the \( Q \)-based strategy is used. The second is the cognitive assumption that the interaction of multiple skills within a given item is conjunctive. That is, a high probability of answering the item correctly requires successfully executing all the required skills. By further assuming local independence of the item responses, Equation 10.1 can be used to model the probability of any given response pattern, \( x \). If \( k_i \) is the number of skills required for item \( i \), then for each item \( i \) on the test there are \( 2k_i + 3 \) item parameters: \( \pi_{ik} \) (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.

The Unified Model IRF not only models examinee responses as influenced by \( Q \), but also allows non-\( Q \) skills to influence examinee response probability 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
general component latent trait model of Embretson (1984), the Unified Model has cognitively interpretable parameters, but unfortunately not all parameters are identifiable or, thus, 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 described in this chapter. However, because nonidentifiable parameters existed in the original Unified Model (see Jiang, 1996), a reduction in the parameter space was required to make the model estimable. To accomplish this task, 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 ability to estimate the parameters, the modeling of the possibility of alternate strategies was dropped by setting \( d_i = 1, i = 1, \ldots, I \). The reduced model, referred to as the Reparameterized Unified Model (RUM), 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 data sets as compared to other skills diagnosis models, retaining the most substantively important components, such as the capacity for skill discrimination to vary from item to item and the residual ability parameter \( \eta \), an additional and potentially important component of the original Unified Model that is missing from all other models. Equation 10.2 presents the RUM IRF, which is based on the same examinee parameters, \( \alpha_j \) and \( \eta_j \), that are used in the original Unified Model. The \( P_c(\eta_j) \) term again refers to the Rasch model with difficulty parameter \( -c_i \) [the lower the value of \( c_i \), the lower the value of \( P_c(\eta_j) \)]

\[
P(X_{ij} = 1 \mid \alpha_j, \eta_j) = \pi^*_i \prod_{k=1}^{K} r_{ik}^{(1-\alpha_{jk})\times q_{ik}} P_c(\eta_j) \quad (10.2)
\]

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

\[
\pi^*_i = P(\text{correctly applying all item } i \text{ required skills given } \alpha_{jk} = 1 \text{ for all item } i \text{ required skills})
\]

\[
= \prod_{k=1}^{K} \pi_{ik}^{q_{ik}},
\]

(under the assumption of conditional independence of individual skill application)
\[ r_{ik}^* = \frac{P(Y_{ijk} = 1 | \alpha_{ijk} = 0)}{P(Y_{ijk} = 1 | \alpha_{ijk} = 1)} \]

\[ = \frac{r_{ik}}{\pi_{ik}} \]

and

\[ P_{c_i}(\eta_j) = P(\text{correctly applying skills associated with } \alpha_k \text{ to item } i, \text{ conditional on } n_j). \]

The \[ P_{c_i}(\eta_j) \] term refers to the Rasch model (one-parameter logistic model, Rasch, 1961) with difficulty parameter \(-c_i\). Note that in our parameterization, the lower the value of \( c_i \), the lower the value of \( P_{c_i}(\eta_j) \).

where, \( 0 < \pi^*_i < 1 \), \( 0 \leq r_{ik}^* \leq 1 \), and \( 0 \leq c_i \leq 3 \). (The bounds of 0 and 3 on the \( c_i \) parameter were chosen for convenience rather than because of any strict theoretical or logical constraint.)

The Fusion Model reparameterization replaces the \( 2k_i \) parameters \( \pi_{ik} \) and \( r_{ik} \) in the original Unified Model with \((1 + k)\) parameters \( \pi^*_i \) and \( r_{ik}^* \). It is easy to show that the reparameterized model is mathematically equivalent to the original Unified Model, and the \((1 + k)\) parameters are now identifiable (DiBello, Stout, & Hartz, 2000). 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 \( \pi^*_i \) 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 \( I \). For an examinee who has not mastered a required skill \( k_0 \), her correct item response probability is proportional to \( r_{k_0}^* \). The more strongly the item depends on mastery of this skill, as indicated by a value of \( r_{k_0}^* \) closer to 0, the lower the item response probability for a nonmaster of the skill. Thus, \( r_{ik}^* \) is like an inverse 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 \( k \). When most of the \( r_{ik}^* \) parameters for a skill are closer to zero (e.g., averaging less than 0.5), the test is said to display high cognitive structure for that skill, which is indicative of a test that is well designed for diagnosing mastery on the skill. Clearly, the \( r_{ik}^* \) parameters play an important role in evaluating the diagnostic capacity of an assessment instrument.

The distinctiveness of the \( \pi^* \) and \( r^* \) parameters in comparison to parameters in other models is important to note. Other models have indeed included components similar to \( \pi_{ik}^* \) and \( r_{ik}^* \). The models in Maris (1999) have the \( \pi_{ik} \) 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 non-possession of that skill.

The \( P_{ci}(\eta_j) \) 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 items. Interestingly, it is not present in any other skills diagnosis model. In this component, \( c_i \) indicates the reliance of the item response function 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 items, by a unidimensional ability parameter, \( \eta_j \). When \( c_i \) is 3 or more, the item response function is practically uninfluenced by \( \eta_j \) because \( P_{ci}(\eta_j) \) will be very close to 1 for most values of \( \eta_j \). When \( c_i \) is near 0, \( \eta_j \) variation will have increased influence on the item response probability, even with \( \alpha_j \) 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 a particular item’s list of measured skills.

The Fusion Model assumes local independence (LI) given \((\alpha, \eta)\). If multiple skill-by-item assignments are missing from the \( Q \) matrix, then the influence of the missing multiple skills are unidimensionally captured by \( \eta \). In this case, local independence with respect to \((\alpha, \eta)\) will only be approximate.

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 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.

### 3.0 Method of estimation for the model parameters

After reparameterizing the Unified Model, a variety of possible methods for estimating the item parameters were explored. To increase the capacity of the RUM to fit the data and to simplify and improve the estimation procedures, a Bayesian modeling augmentation of the RUM was
developed. Although using Bayesian Networks is one type of Bayesian approach that could have been adopted, we found that the probability structure of interest could be combined with relationships between the skills more directly by using a hierarchical Bayesian modeling approach instead. This Bayesian hierarchical structure, in effect, enhances the reparameterized Unified Model of Equation 10.2 by adding further parameters and their priors, where, as is common practice, the priors are either chosen to be uninformative, strongly informative, or estimated from the data (the empirical Bayes viewpoint). Thus, we describe the Bayesian modeling framework for the ability and item parameters, and then the MCMC procedure for estimating the parameters in the Bayesian framework.

3.1 Bayesian Framework for the Ability Parameters

The prior used for the $\eta_j$ parameter was simply set to the standard normal distribution. However, the Bayesian model underlying the dichotomous skill mastery parameters required a more complicated approach. The dichotomous $\alpha_{kj}$ ability parameters are modeled as Bernoulli random variables with probability of success $p_k$, the population proportion of masters for skill $k$. The prior for the dichotomous $\alpha_{kj}$ ability parameters consisted of the $p_k$ parameter for each skill and the tetrachoric correlations between all skill mastery pairs. These parameters are modeled as hyperparameters in a hierarchical Bayes model.

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 well-known latent variable tool called tetrachoric correlations. Tetrachoric correlations model the relationship between two dichotomously measured variables, where it is assumed that a normally distributed latent variable generates each observed dichotomous variable (see, e.g., Hambleton & Swaminathan, 1985). The use of tetrachoric correlations was preferred over raw correlations between the dichotomous skills because raw correlations are highly dependent on the differing proportions of masters for each skill, thus making them more difficult to interpret as measures of the strength of relationship between two skills.

As stated, the tetrachoric correlations assume that continuous normal random variables underlie the dichotomous $\alpha_{kj}$ mastery variables. It is assumed that the continuous variables have been dichotomized by cut-point parameters. These continuous variables, which we denote
symbolically as \( \tilde{\alpha}_{kj} \), are viewed as continuous levels of skill competency that, when dichotomized by “cut points,” determine mastery versus nonmastery of a skill. The cut-point parameters are denoted as \( \kappa_k \) and are related to the \( p_k \) parameters by the relation \( P(\tilde{\alpha}_{kj} > \kappa_k) = p_k \), where \( \tilde{\alpha}_{kj} \) is assumed to follow a standard normal distribution.

The correlations between \( \eta_j \) and the underlying \( \tilde{\alpha}_j \) variables are, of course, modeled as nonnegative correlations. These correlations are estimated as hyperparameters (allowing the data to help estimate their joint distribution) and are given a Uniform prior, Unif(\( a, b \)), where \( a \) and \( b \) were set to 0.01 and 0.99, respectively, to prevent boundary problems. Numerical techniques for drawing the correlations in the MCMC algorithm were used to maintain positive definiteness, as is required of the correlation matrix (see Hartz & Roussos, in press, for details).

### 3.2 Bayesian Framework for the Item Parameters

Because the values of \( \pi^*_i \), \( r^*_ik \), and \( c_i \) can vary greatly for a given data set, the priors (distribution functions) for the three types of item parameters were each chosen to be a Beta distribution, allowing maximum flexibility of the shape the priors can take. Beta distributions are often used in Bayesian analyses because they naturally model parameters that are constrained to finite intervals, as is the case with \( \pi^*_i \) and \( r^*_ik \). Even though, strictly speaking, \( c_i \) can be any real value, it makes practical sense to restrict its values for model parameter estimation to a finite interval. There are also mathematical reasons to prefer Beta distributions as priors. These reasons have to do with the Beta distributions being conjugates of the family of binomial distributions. To take advantage of the valuable flexibility of the beta distribution, the parameters of the Beta priors are themselves estimated by assigning them priors and estimating the corresponding hyperparameters (the hierarchical Bayesian approach allowing the data to influence the estimated shape of the distributions, as is desirable). In the current version of Arpeggio, these hierarchical priors are fixed to be approximately uniform distributions.

### 3.3 Brief Description of MCMC

Largely due to the influence of the pair of Patz and Junker (1999a, 1999b) articles, MCMC has become popular in many skills diagnostic applications, in particular, for the fully Bayesian Fusion model (Roussos, DiBello, & Stout, in press) and the NIDA and DINA models (de la Torre...
& Douglas, 2004). Although EM (expectation and maximization) algorithms are often much more efficient computationally than MCMC algorithms, EM algorithms are more difficult to extend to new models or model variants than are MCMC algorithms. As Patz and Junker (1999a) point out, EM algorithms are not as “straightforward” (p. 147) see also Junker, 1999, p. 70) to apply for parametrically complex models. Furthermore, one obtains a joint estimated posterior distribution of both the test’s item parameters and the examinee skills parameters, which can be persuasively argued to provide better understanding of the true standard errors involved (as Patz & Junker, 1999a, 1999b, argued in their papers on MCMC, uses in psychometrics).

Here, we provide a brief conceptual description of MCMC. First, a probabilistically based computational method is used to generate Markov chains of simulated values to estimate all parameters. Each time point (or step) in the chain corresponds to one set of simulated values. MCMC theory states that after a large enough number of steps (called the burn-in phase of the chain), the remaining simulated values will closely approximate the desired Bayesian posterior distribution of the parameters. It is important to note that obtaining this posterior distribution would almost always be impossible to compute analytically or approximate numerically, even using modern computational approximation methods because of the level of parametric complexity in our model. A solution to the analytical intractability is provide by implementing the Metropolis-Hastings procedure with Gibbs sampling that uses much simpler proposal distributions in place of the analytically complex full conditional distributions (see Gelman, Carlin, Stern, & Rubin, 1995). In the current version of Arpeggio, the proposal distributions (also called jumping distributions) take the form of moving windows (Henson, Templin, & Porch, 2004). The parameters defining these proposal distributions are already fixed by the program to appropriate values, and users should not change these values without first carefully consulting Henson, Templin, et al. (2004).

MCMC estimation is accomplished by running suitably long chains, simulating all parameters at each time point of the long chain, discarding the burn-in steps, and using the remaining steps in each chain as posterior distributions to estimate the parameter values and their standard errors. The practitioner must carefully choose the number of chains to be used, the total length of the chain, and the amount to be used for the burn-in. When a practitioner has little experience for a given analysis, we recommend the use of at least two chains of very long length
and very long burn-ins (e.g., a chain length of 50,000 and a burn-in of 40,000). The next section in this chapter provides information on how to evaluate the MCMC results. The post–burn-in chains are estimates of the posterior distributions of the parameters. The Arpeggio software uses these distributions to provide EAP (expectation a posteriori) point estimates and posterior standard deviations as measures of variability.

It is interesting to note that MCMC could be used from either a partially Bayesian or a non-Bayesian perspective, and both are useful in different circumstances. A common application is fully Bayesian (as described here for the Fusion Model), often with hierarchical structure to give it an appropriate empirical Bayes flavor. The primary inferential outcome is the full posterior distribution of the joint item and examinee parameter distributions, as represented by a set of random draws from the joint distribution. Many argue that having these distributions and, in particular, the available estimated standard errors of each parameter, is of considerable value. Many MCMC references exist, and the Patz and Junker (1999a, 1999b) articles and the Gelman et al. (1995) book are especially recommended to interested readers considering parametrically complex models such as the models surveyed here.

4.0 MODEL CHECKING PROCEDURES

In this section, we discuss a variety of model checking methods that are available with the Fusion Model system, how to interpret the results of these methods, and what actions one might take in reaction to the results.

4.1 Convergence Checking

All statistical estimation methods that involve iterating until convergence to obtain a solution require checking for convergence – either convergence to within some specified tolerance, as in an EM iterative algorithm, or convergence to the desired posterior distribution, as in the case of MCMC. Because the statistical information one obtains from MCMC estimation (a full posterior distribution) is richer than that obtained from an EM algorithm (an estimate and its standard error), the evaluation of whether convergence has occurred is more difficult in the MCMC case.

In skills diagnosis applications, MCMC convergence may be difficult to obtain, depending on the complexity of the model and how well the design and assumptions of the model correspond to the reality of
the data. In some cases, complex models may be statistically nonidentifiable. In other cases, identifiable models may be difficult to estimate well because of ill-conditioned likelihood functions. Of course, these modeling issues will cause problems for all estimation methods, not just for MCMC. In any case, the first thing that must be done after running the Arpeggio MCMC algorithm is to determine whether the chains converged to a stationary solution (presumed to be the posterior distribution of the parameters).

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. The Fusion Model system provides results for evaluating convergence in four ways: chain plots, estimated posterior distributions, autocorrelations of the chain estimates, and Gelman and Rubin $\hat{R}$ (Gelman et al., 1995), each of which is discussed later in this section.

One important tool (perhaps the most important tool) for checking convergence is the chain plot for each parameter – a graph showing the value of the parameter for each step in the chain. In Figure 10.1, we give an example of an investigation by Jang (2005; Roussos, DiBello, Henson, Jang, & Templin, in press) of the effect of chain length on the estimation of an item parameter in the context of analyzing a 37-item test taken by about 1350 examinees using a $Q$ matrix having nine skills, with an average of about two skills per item (which translates to an average of about eight items per skill). The plots show every 10th term in the chain to simplify the plots and reduce the sizes of the output. In this example, it seems evident from the chain plots that the first 1000 steps can adequately serve as the burn-in. The plots clearly indicate that the chains have settled into a stable distribution.

Also shown in Figure 10.1 are the posterior distributions. These graphs support the inference from the chain plots that convergence has occurred because there is little change in the posterior distribution after the first 1000 steps of the chain.

Figure 10.1 also shows the autocorrelation function, which can be used to determine the degree of independence between two groups of simulated data that are separated by a specified number of steps in the chain. The lower the autocorrelation is, the more information there is for a given chain length. Thus, if autocorrelation is relatively high, the chains must be run longer to estimate the posterior distribution with reasonable accuracy. Figure 10.1 indicates that there is little autocorrelation between
Figure 10.1. Effect of chain length on item parameter estimation convergence. (From Jang, 2005; Roussos, DiBello, Henson, Jang & Templin, in press.)
chain estimates that are spaced 500 or more steps apart. This example illustrates good convergence results and is no doubt due to the careful model development by Jang.

Note that high autocorrelations do not necessarily imply MCMC non-convergence but, alternatively, may simply be indicating a relatively flat likelihood surface. This is a common problem with complicated models that are likely for almost any particular data set to have at least a few nonsignificant parameters, parameters whose values for the given data set are not significantly different from the values one would expect under an appropriate null hypothesis of no parameter influence (e.g., 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 data sets.

In Figure 10.1, Jang (2005) looked at a single chain with varying chain lengths. Another important method supported by the Fusion Model system is to run Arpeggio for multiple chains and compare the results, including the chain plots, posterior distributions, and autocorrelations. If the chosen chain length has achieved convergence, then the results from different chains should be practically the same. Also, when you use multiple chains, the Gelman and Rubin $\hat{R}$ (the ratio of between-chain variance plus within-chain variance to within-chain variance; Gelman et al., 1995) will be reported in the results. Gelman et al. recommend that $\hat{R}$ values less than 1.2 are a necessary, but not sufficient, condition for convergence. In our experience, unfortunately, we have seldom seen $\hat{R}$ values greater than 1.2 in cases of nonconvergence. When we have, it has usually been an indication of a gross error we have made in one of our input or data files.

When convergence does not occur with the Fusion Model system, we have a few suggestions based on our own experiences. One frequent problem we have encountered is that when the $c_i$ parameter (and its accompanying $\eta_j$ ability parameter for the non-$Q$ skills) is included in the model, estimation of the model parameters pertaining to the $Q$ skills (i.e., the $\pi_i^*$ and $r_{ik}^*$ parameters) does not converge (or converges to a solution having most of the $p_k$ parameters being very large). The reason this has occurred so frequently in our analyses is that most of our applications have been with tests having a single dominant dimension, and this allows the continuous $\eta_j$ parameter to “soak up” most of the variance in the item responses. Conversely, we also sometimes have non-convergence for the $c_i$ parameters in cases in which many items have large numbers of skills required as specified in the $Q$ matrix. In such
cases, there is no variance left to be accounted for by the \( P_{c_i}(\eta_j) \) part of the model. If either of these situations arises, a good solution is to drop the \( c_i \) parameter from all the items in the model – this is accomplished through a user set option that automatically sets all the \( c_i \) values equal to 10, which effectively makes the \( P_{c_i}(\eta_j) \) term equal to 1 for all reasonable values of \( \eta_j \). This reduced version of the Fusion Model is the version we often have used for analyzing real data.

If the \( c_i \) parameter is not in the model and nonconvergence occurs, the first thing to check is whether the burn-in phase of the MCMC chain was long enough to reach the posterior distribution phase. This can be checked by running an extremely long chain. If the longer chain still does not result in convergence, one can probably rule out chain length as the problem. In this case, one can revisit the model-building steps, and reconsider the \( Q \) matrix and the selected model to identify where changes may be warranted. For example, if the model assumes that a skill is difficult (small \( p_k \)), yet the skill is assigned to items having a large range of difficulty (including both hard and easy items), the MCMC algorithm may not be able to converge to a single level of difficulty for the skill.

There is much more to say, in general, about MCMC convergence. The reader is referred to Cowles and Carlin (1996) and a recent article by Sinharay (2004) for an excellent and thorough discussion.

### 4.2 Interpretation of Model Parameter Estimates

Given that the model estimation procedure has converged, the estimates for the ability distribution (for the skills) and item parameters should be evaluated.

#### 4.2.1 Ability Distribution Parameters

A key issue for mastery/nonmastery diagnostic models is whether the proportion of examinees estimated as masters on each skill is relatively congruent with the user’s expectations. The Fusion Model system provides estimates of the \( p_k \) parameters and estimator standard errors for this purpose. If a skill turned out much harder or easier than expected, based on the \( p_k \) estimate and its standard error (e.g., harder or easier than desired from a formal standard-setting perspective, or relative to the difficulty of other skills), the \( Q \) matrix should be revisited and the item difficulty levels investigated for the items to which the skill has been assigned. In addition, the choice of tasks assessing that skill can be
revisited to see if other more appropriate tasks can be found to either include or replace existing tasks. For example, if the proportion of masters for a skill is too low, one could try replacing the harder tasks for that skill with easier ones, adding easier ones, or both. Ultimately, the definition of the skill may need to be adjusted, for example, by suitable modification of \( Q \) or, in a more basic way, leading to a new set of tasks. For example, in Jang (2005, 2006; Roussos, DiBello, Henson, Jang, & Templin, in press), one of the skills was found by statistical analysis to be much easier than expected, whereas the other eight skill difficulties were ordered in a manner consistent with Jang’s approximate expectations.

4.2.2 Item Parameters
The Fusion Model system also provides estimates and standard errors for the item parameters. The estimates for the \( \pi_i^*, r_{ik}^* \), and, if present, the \( c_i \) parameters, should be inspected in detail because they play a key role in determining the success of the diagnosis. (The interpretive meaning of the item parameters is discussed previously and is thus only briefly repeated, as needed, here.)

First, we should repeat something we covered in Section 4.1. If the \( c_i \) parameters are included in the model and estimation convergence is obtained, the solution should be carefully investigated to determine whether the \( P_c(\eta_j) \) part of the model dominates the fitted model. If that occurs, then most of the \( p_k \) parameters will be very large (artificially making nearly everyone a master on most of the skills) so most of the individual differences are explained by the \( \eta_j \) parameter. This has occurred frequently in our analyses because most of our applications have been with tests having a single dominant dimension, and this allowed the continuous \( \eta_j \) parameter to soak up most of the variance in the item responses. In this case, even if convergence has occurred, to obtain a diagnostically useful estimated model one needs to drop the \( c_i \) parameters from the model (see explanation in Section 4.1).

Assuming convergence has occurred and the estimates do not indicate any fundamental problem, as with the previous \( c_i \) parameters, the estimates for each type of item parameter will provide useful information on the performance of each item relative to its required skills.

The simplest item parameter to investigate is \( \pi_i^* \), which indicates the probability that examinees have correctly executed all \( Q \) skills required by an item, conditional on having mastered all required skills. We want this value to be close to unity (1.0). In our applications, we have interpreted values of \( \pi_i^* \) less than 0.6 as indicating items that are overly difficult for the skills assigned to them. Either more skills or different
(and likely more difficult) skills would need to be assigned to such items in order to have correct item response be indicative of mastery of the particular skills assigned to the item.

The remaining two types of item parameters, $r_{ik}^*$ and $c_i$, give information on how well the items discriminate on the $Q$ skills and the degree to which skills are missing in the $Q$ assignments for an item, respectively. A low value of $r_{ik}^*$ (0–0.5) is indicative of a highly discriminating item, and a low value of $c_i$ (0–1.5) is an indication that either a $Q$ skill is missing for the item or (if this occurs for many items) a skill is missing from the entire $Q$ matrix. High values of $r_{ik}^*$ and $c_i$ are indicative of possible model simplifications. If $r_{ik}^*$ is bigger than, say, 0.9, the item is not very discriminating for the corresponding skill and the $Q$ matrix entry can be eliminated (1 changed to a 0), and if $c_i$ is big, then the item is approximately “complete” and $c_i$ can be dropped (i.e., set equal to 10).

It should be noted that a “1” entry in the $Q$ matrix can be fully justified on substantive grounds – considering the definition of the skill and the coding procedures – and still show an estimated $r_{ik}^*$ value that is very near 1.0. This is analogous to the situation in which a test item for a unidimensional test looks good on substantive grounds, but the discrimination parameter is very small, resulting in a relatively flat item characteristic curve. For diagnostic assessments, the high estimated $r_{ik}^*$ value means that statistically the particular item is not contributing much information for distinguishing between masters and nonmasters of the given skill. So dropping the “1” entry from the $Q$ matrix is not necessarily indicative of a “mistake” in the item coding. Instead, the decision is a statistically strategic one that bets on the gain in parameter estimation accuracy, with fewer parameters to be estimated and the resulting improvement in classification accuracy and interpretability (see later discussion) to be a worthwhile benefit to justify dropping the “1” entry in the $Q$ matrix, even in light of the substantive rationale for the original “1” coding. Of course, if such cancellations occur for many of the items requiring a particular skill, the practitioner should interpret that as a signal that the skill or the skill coding procedure may need to be modified to bring the substantive coding and statistical behavior into better alignment.

The elimination of noninformative $r_{ik}^*$ and $c_i$ parameters is considered to be an important component of the Fusion Model system because it helps the model concentrate its statistical power where there is diagnostic information to be found. In other words, such reduction in the number of model parameters may be beneficial simply because fewer parameters can be better estimated (smaller standard errors) with the
same observed data. However, more important, dropping nonsignificant $Q$ matrix entries (which is what occurs when an $r_{ik}^*$ is eliminated) reduces the possibility of examinee confusion that can arise from examinees labeled nonmasters performing nearly as well as those labeled masters on the items corresponding to the $Q$ entries being eliminated. To this end, we have developed statistics to help identify statistically noninfluential $r_{ik}^*$ and $c_i$ item parameters.

To accomplish the goal of identifying noninfluential parameters, the Fusion Model system estimates the influence of $r_{ik}^*$ and $c_i$ on the item response function that uses that parameter. By employing a common probability scale for measuring this influence, the same statistical decision rule is used for both $r_{ik}^*$ and $c_i$ in determining whether the estimated influence is large enough to warrant keeping the parameter in the model. In our simulation studies, these decisions were 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. However, it is important to note that in practice, as discussed previously, 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, the dropping of $Q$ matrix entries needs not only strong statistical evidence, but also strong substantive arguments.

To estimate the influence of a particular item parameter, three different item response probabilities are calculated for each examinee from the Fusion Model IRF. With 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 $r_{ik}^*$ would be fixed at 1.0, and a $c_i$ 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, the three IRF probability calculations are averaged over all examinees who are estimated as nonmasters of skill $k$ (these are the only examinees for whom the $r_{ik}^*$ would appear in their IRF). For a $c_i$ parameter, the averaging is done using all examinees. When the average of item response probability (a) is close to that for either (b) or (c) (or both, depending on the preference of the practitioner), the parameter is said to be noninfluential and is dropped from the model.

The determination of whether two average probabilities are close is ultimately a subjective decision, but there still exists experience from other IRT domains of interest that can help guide decision making here.
The most common examples of such decision making with which we are familiar 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 to be a negligible amount, whereas a difference between 0.05 and 0.10 is often considered to be moderate and a difference greater than 0.10 is usually considered to be large (see, e.g., Dorans, 1989). In our simulation studies, we found that a relatively liberal approach (using cut-off values that are on the low side within the range of the typical DIF cut-offs) worked best for these particular studies. In particular, in our simulation studies, we declared that a $c_i$ or $r_{ik}^*$ is noninfluential for an item (and set the parameter to its null hypothesis value, thus dropping it from the model) when the average for (a) is within 0.03 of the average for (b) or within 0.01 of the average for (c).

Practitioners may, of course, choose other reasonable criteria and may choose to use only the difference between (a) and (b) or only the difference between (a) and (c). The former corresponds to a situation in which a practitioner is philosophically choosing a null hypothesis that the parameter belongs in the model and will only drop 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 the combined use of both types of criteria. The choice is, as it should be, up to the practitioner.

Finally, because the item parameters across skills and items influence one another, we perform the calculations iteratively rather than all at once. On each run, we restrict ourselves to dropping a maximum of one parameter per item before rerunning the entire model and rechecking these statistics again (one can conduct this analysis in a repeated manner as in a stepwise regression). 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 $r_{ik}$ and a $c_i$ parameter are found to be noninfluential at the same step, the $c_i$ parameter is preferentially dropped to favor the possibility of retaining skill mastery estimation in the item over the possibility of retaining estimation of $\eta$ because skills diagnosis is the primary purpose of the analysis. Also, if two or more $r_{ik}$ parameters for a particular item are found to be noninfluential, the parameter with the smallest average difference between the null hypothesis IRF probability (a) and the upper bound IRF probability (c) is dropped.
Even when a $Q$ matrix is carefully developed for estimation of the Fusion Model with real data, or even when you generate simulated data and use the known $Q$ matrix in your 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 (e.g., 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 lower 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.)

An example of this situation is when data are simulated with low $r^*_{ik}$ values and high $\pi^*_i$ values (referred to as “high cognitive structure”), and moderate $c_i$ values. In this case, the $c_i$ parameters may not be well estimable because the item responses are largely determined by whether examinees have mastered the skills and are minimally affected by their proficiency on non-$Q$ skills. Additionally, if an item measures both a hard skill and an easy skill, and the $r^*_{ik}$ for the harder skill is very low, the $r^*_{ik}$ for the easier skill will have little influence in the IRF.

An example of a real case in which parameter dropping was used is the analysis of Jang (2005, 2006; Roussos, DiBello, Henson, Jang, & Templin, in press). In this study, the estimated item parameters led Jang to eliminate about 9% of her $Q$ matrix entries because certain items did not discriminate well on some of the skills that had been assigned to them. Specifically, the estimated parameters indicated that the ratio of skill performance by nonmasters to that by masters was 0.9 or more for about 9% of the item-skill combinations.

4.3 Model Fit Statistics

Once the Fusion Model has been calibrated from data, model fit is evaluated by using the fitted model to predict observable summary statistics that are compared to the corresponding observed statistics. The predicted statistics are obtained by simulating data from the fitted model and calculating the statistics on the simulated data. The item parameters used in the simulation model are the item parameter estimates (EAPs) from the real data analysis. The ability parameters used in the simulation model are probabilistically generated as follows. For each examinee in the calibration data set, the fusion model system provides
a posterior probability of mastery (ppm) for each skill and an estimate of $\eta$. In the simulation, examinees are sampled with replacement from the calibration sample, and an $(a_k, k = 1, \ldots, K)$ vector is simulated by generating an independent series of Bernoulli random variables, one for each skill, with the Bernoulli probability for each skill being equal to the sampled examinee's ppm for each skill. The resampled examinee's actual estimated $\eta$ is used for the simulated value. The item parameters and ability parameters are then used in the standard way to generate simulated item responses. The resampling is done, say, 500,000 times, to obtain accurate predicted statistics. This type of approach is called posterior predictive model checking (see Gelman et al., 1995; Gelman & Meng, 1996; Sinharay, 2005; Sinharay & Johnson, 2003). The statistics that are typically calculated in the Fusion Model system are the proportion-correct scores on the items, the item-pair correlations, and the examinee raw score distribution (see Henson, Roussos, & Templin, 2004, 2005). The item pair correlations and the score distribution represent more demanding tests of data-model fit, and must be interpreted with care and in light of the diagnostic purpose of the assessment.

As a specific example, in the application of Jang (2005; Roussos, DiBello, Henson, Jang, & Templin, in press), the mean absolute difference (MAD) between predicted and observed item proportion-correct scores was 0.002, and the MAD for the correlations was 0.049, supporting the claim of good fit. One can also compare the fit between the observed and predicted score distributions. Figure 10.2 shows a comparison between the observed and the predicted score distributions from the

![Figure 10.2. Comparison of observed and model estimated score distributions. (From Jang, 2005, 2006; Roussos, DiBello, Henson, Jang, & Templin, in press.)](image-url)
real data analysis of Jang (2005, 2006; Roussos, DiBello, Henson, Jang, & Templin, in press). The misfit at the very lowest and highest parts of the distribution were expected as the mastery/nonmastery examinee model overestimated the scores of the lowest scoring examinees and underestimated the scores of the highest scoring examinees. Because the goal of the analysis was to estimate mastery/nonmastery rather than to scale examinees, this misfit actually had no effect on the mastery/nonmastery classification. Even though the lowest scoring examinees had overestimated scores, they were still all classified as nonmasters on all the skills. Similarly, although the highest scoring examinees had underestimated scores, they were still estimated as masters on all skills.

This is an example of a case in which one might discover that a simpler unidimensional model with a continuous ability parameter would fit the score distribution data better (i.e., produce predictions of total scores that look more like the observed data). But if such a model did not yield the desired mastery/nonmastery classification estimates without further work, it would not satisfy the diagnostic assessment purpose.

For a broader discussion of model diagnostic fit statistics relevant for cognitive diagnosis models, the reader is referred to Sinharay (2006). In this article, Sinharay demonstrates several model diagnostic approaches in the context of a real data analysis, and identifies several instances of poor model fit and severe nonidentifiability of model parameters. Although the Sinharay article involves a real data skills diagnosis that used a Bayes net approach, the model fit approaches he demonstrates are broadly applicable to any parametric skills diagnosis model, including the Fusion Model.

4.4 Internal Validity Checks

In general, broad conceptions of validity go beyond criterion validity in which model estimates and predictions are compared to observable data. For example, construct validity for an assessment instrument refers to the extent to which measurements of the instrument’s construct(s) have expected relationships with other constructs in a wider nomothetic structure of related constructs; and consequential validity refers to the extent to which desired and positive consequences result from the use of assessment information.

Nevertheless, criterion validity based on observable data remains a critical component of the overall test validation process. We categorize criterion validity into two broad categories: internal validity
(synonymous with internal consistency) and external validity. In the case of internal validity, the observable data come from the test data themselves. External validity statistics relate test-derived ability estimates to some other ability criterion external to the test – in particular, a criterion indicating whether the particular educational purpose of the skills diagnosis has been well served. The Fusion Model system includes internal validity checks, and users are also strongly encouraged to conduct external validity checks according to whatever criteria external to the test are most appropriate for the user’s setting.

One kind of internal validity check is to measure the differences in observed behavior between examinees that are classified differently, and this idea provides the basis for the internal validity checks we have developed. Specifically, the Fusion Model produces two types of such statistics: IMstats for item mastery statistics, and EMstats for examinee mastery statistics.

IMstats describes how well, on an item-by-item basis and on average over all items, the Arpeggio MCMC estimates of examinee mastery of each skill correspond to the actual observed performance of the examinees on each item. Specifically, for each item on the test, the examinees are divided into three groups according to how many of the particular skills required by the item each examinee has been estimated as having mastered. Examinees who have mastered all skills required by 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 different from our previous usage with respect to mastery of individual skills. Here mastery is discussed relative to all skills required for an item, not mastery relative to a single skill. Then, on an item-by-item basis, IMstats computes the observed proportion-right score on each item for the examinees falling into each of these three groups for that item. To determine whether 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 formal 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 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. After each examinee has been estimated as either mastering or not mastering each skill, the particular set of mastered skills for a given examinee provides the basis for dividing the examinee’s items into three groups. Specifically, for each examinee \( j \), 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 selected examinee has been estimated as having mastered:

- **Examinee \( j \)'s mastered items.** The set of items for which examinee \( j \) has been estimated as having mastered all \( Q \) matrix skills required for each item.
- **Examinee \( j \)'s high nonmastered items.** The set of all 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, according to the \( Q \) matrix.
- **Examinee \( j \)'s low nonmastered items.** The set of all items for which the examinee has been estimated as having nonmastered more than half of the skills required by the item, according to the \( Q \) matrix.

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

EMstats then compares the examinees observed proportion-right score for the examinee mastered items with a criterion value to determine whether this score is unusually low. Similarly, the scores on the examinees high nonmastered, low nonmastered items, and all nonmastered items together are also compared to criteria to determine whether these scores are unusually high. Because these observed proportion-right 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,
The Fusion Model Skills Diagnosis System

*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-right score, and for any of the sets of nonmastered items we only care whether the examinee had an especially high proportion-right score. The criterion values are either set by the user or are based on the observed item proportion-right scores for masters, high nonmasters, and low nonmasters from *IMstats*. The criterion values based on the latter case are used in this study.

The hypothesis tests are conducted in the following manner. First, the examinee’s observed proportion-right score is subtracted from the appropriate criterion value. Then, under the assumption that the null hypothesis holds, the standard error is computed as 
\[
\sqrt{P_c (1 - P_c) / n},
\]
where the summation is over the \( n \) items of interest and \( P_c \) 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 whether the calculated \( z \) statistic is less than \(-1.645\) when the focus of the hypothesis test is on the mastered items or 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 that the number of items in a category 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.

As an example of the use of the *IMstats* internal validity check in a real data analysis, Figure 10.3 presents the *IMstats* results from Jang (2005, 2006; Roussos, DiBello, Henson, Jang, & Templin, in press) for one of the test forms she analyzed. These results clearly indicate a high degree of internal validity for the skills diagnosis because the mean score differences between item masters and item nonmasters are quite large for the vast majority of the items. The results also clearly point to certain problematic items, which Jang investigated. She discovered that these items tended to be either extremely easy or extremely hard. As noted by Jang, the test she used had been originally intended as a norm-referenced test, with the purpose of placing examinees on a continuous scale, thus requiring items of a wide range of difficulty. Jang quite rightly notes that such a test, developed for another purpose, is not necessarily a good one for doing skills diagnosis, and the identification of these problematic items through the internal validity analysis drove home this point.
4.5 Reliability Estimation

Standard reliability coefficients, as estimated for assessments modeled with a continuous unidimensional latent trait, do not translate directly to discrete latent space modeled cognitive diagnostic tests. We note, however, that conceptions of reliability from first principles do still apply. Diagnostic attribute classification reliability can be conceptualized in terms of the twin notions of, on the one hand, the correspondence between inferred and true skills mastery state and, on the other hand, the consistency of classification if the same assessment were administered to the same examinee multiple times. It is this concept of consistency that is operationalized as the method for reliability estimation in the Fusion Model system.

Specifically, to estimate classification reliability, the calibrated model is used to generate parallel sets of simulated data and then to estimate mastery/nonmastery for each simulated examinee on each set. The simulation is carried out in the same way described previously for the model fit statistics in Section 4.3. For each skill, we calculate the proportion of times that each examinee is classified correctly on the test (estimation of correct classification rate) and the proportion of times each examinee is classified the same on the two parallel tests (estimated test–retest consistency rate). These rates are reported separately for true masters and true nonmasters of each skill, and also averaged over masters and nonmasters. The mastery/nonmastery estimation is accomplished by...
calculating the ppm for each examinee and declaring the examinee a master if $ppm > 0.5$. An option of using an indifference region is also provided, such that an examinee is declared a master if $ppm > 0.6$, nonmaster if $ppm < 0.4$, and unknown otherwise (the user can choose values other than 0.4 and 0.6 for the upper and lower boundaries of the indifference region). The use of an indifference region increases the reliability, at the cost of declaring a set of examinees to be of indeterminate mastery.

In any case, another set of statistics produced are a set of “pattern” statistics: the proportion of examinees whose estimated skill masteries are incorrect (estimated as a master when truth is nonmaster, or estimated as nonmaster when truth is master) for 0 skills, 1 skill, 2 skills, etc. This is a useful statistic because it can be used to judge how likely it is that an individual examinee has a certain number of errors in their estimated skill profile. For example, in an analysis by Henson, He, and Roussos (in press), 99% of the simulated examinees had one or no errors. Such statistics are easy for score users to interpret. When they get an estimated score profile, they can compare the results with their own records, keeping in mind that more than one incorrect skill mastery classification is extremely unlikely. Thus, if they find one skill classification that does not seem quite right, they should feel comfortable that they can disregard it without wasting precious time and effort trying to figure out the discrepancy. This seems like the kind of reliability estimation that teachers may be able to understand and use in their diagnostic score interpretations for a given skills profile – how many errors are there likely to be?

5.0 **SCORE REPORTING STATISTICS**

The last component in the Fusion Model system is the production of statistics that are useful for skills-level score reporting, which, for example, would be directed toward students, teachers, and parents in an educational diagnostic assessment setting.

Cognitive diagnostic analysis using the Fusion Model system can provide students and instructors with detailed skills-level student assessments that can be used to improve instruction and learning. Once a satisfactory calibration of a given test has been completed, then diagnostic scores can be provided in the classroom or across the Web by using either the *Arpeggio* MCMC procedure, with item and population parameters fixed to calibrated values, or the Bayesian posterior scoring program called the *Fast Classier*. The latter is the preferred method because
it computes the mathematically correct maximum likelihood or maximum a posteriori estimates of the posterior probabilities of mastery and is computationally more efficient.

5.1 \( ppm \) and \( \eta \)

The basic set of statistics reported are the ability estimates: the estimate of \( \alpha \) (\( \alpha_k \), \( k = 1, \ldots, K \)), and the estimate of \( \eta \) if the \( c_i \) parameters were included in the model.

As mentioned previously, the estimate of the \( \alpha_k \)s are in the form of \( ppm \)s. These are obtained directly from the post–burn-in section of the MCMC chain for each skill. During the MCMC estimation at each step in the chain, \( \alpha \) is proposed as a string of 1s and 0s for an examinee. After the chain has converged to the posterior distribution and has run for a large number of steps, the proportion of 1s in the chain for a particular skill is, by definition, the \( ppm \) for the skill for that examinee. An examinee is then declared a master of a skill if \( ppm \) is above some user-provided upper threshold (0.5 or more), a nonmaster if \( ppm \) is below some user-provided lower threshold (0.5 or less), and is declared “indeterminate” if \( ppm \) is between the two thresholds. One interesting choice for the indifference region is (0.2, 0.8), which results in Cohen’s kappa of at least 0.6 for all classifications, where 0.6 is regarded as a rule of thumb for a classification that is “substantial” relative to chance (on a scale of no agreement, slight, fair, moderate, substantial, and nearly perfect agreement; Landis & Koch, 1977).

From our experience, an examinee’s \( ppm \) may fall into the indeterminate region typically for one or more of the following reasons: (a) the items were not very good for discriminating mastery from nonmastery for a particular skill, and/or (b) the examinee behaved inconsistently, neither like a master nor like a nonmaster but, rather, somewhere in between these two states (recall that the mastery/nonmastery dichotomy, although necessary for decision making, is a dichotomy of a more continuous or multicategory construct). If case (a) holds, the result may be a higher than desired proportion of examinees in the indifference region. The previous reliability calculation can help determine whether case (a) is occurring for a particular skill.

In regard to \( \eta_j \), the Fusion Model system provides a Bayesian EAP estimate based on the MCMC estimated posterior distribution. In general, the estimate for \( \eta_j \) is not directly connected to the skills specified in the \( Q \) matrix and consequently is not used for diagnostic purposes.
5.2 Relating Scores to Skill Profiles: Proficiency Scaling

The ppm's mentioned in Section 5.1 are the primary (most direct) means for conveying skills-level diagnostic information to test takers and test users (e.g., students, parents, teachers, administrators). However, for test takers and users to have a high level of confidence in the ppm's, it is important that they have statistics that illuminate the relationship between skill mastery and the observed scores on the test – the scores being an entity with which they are much more familiar and can easily observe for themselves.

To this end, the Fusion Model system estimates the relationship between all possible test scores and all possible skill mastery patterns (all possible values of $\alpha$, the vector of dichotomous mastery parameters for the K skills). The estimation method is based on simulating 100,000 examinee item responses from the fitted model, as described in Section 4.3 (for more details, see Henson, Roussos, & Templin, 2004, 2005). To make these results more interpretable, two sets of summary statistics are produced.

The first set of statistics summarizes the distribution of skill mastery patterns given a test score by reporting (a) the three most probable skill mastery patterns for each fixed score (and their associated probabilities in the simulated population at that score); (b) a count of the minimum number of skill mastery patterns whose probabilities sum to at least 50%, 75%, and 90% at each fixed score; and (c) the mean or expected ppm for each skill at each fixed score.

The second set of statistics summarizes the distribution of test scores for each possible skill mastery pattern. The Fusion Model system reports for each skill mastery pattern: (a) the expected score; (b) the standard deviation of the scores; and (c) the three most likely scores, the corresponding probability of each, and the sum of these three probabilities.

For well-designed tests and carefully implemented skills diagnosis, these statistics will help give users more confidence because they will see that an expected test score generally increases as the number of mastered skills increases. Furthermore, in knowing this conditional information, quick rudimentary estimates of skill mastery can be made from an examinee's test score.

Such information can largely be useful in low-stakes situations to give teachers additional information that can help in their approach to teaching and improving the learning environment. By knowing the relationship between test scores and skill mastery patterns, teachers
can quickly evaluate each examinee’s likely skill mastery pattern and
determine what would be an appropriate focus for the examinee to
improve his or her performance. However, such summaries of the joint
distribution of test score and skill mastery pattern are not limited to
those we have presented in this section. Others could be used for specific
purposes that we have not yet addressed.

5.3 Use of Subscores for Estimation of Skill Masteries

In some test settings, practitioners may require an alternative to the ppm
(see Section 5.1) that come from the Arpeggio MCMC scoring engine or
from the Fast Classier. For example, a state testing program might man-
date that mastery decision making be based on test cut-scores, the most
easily interpretable information that can be transmitted to test users.
This section discusses reasonable uses of subscores in cases in which
we have performed a Fusion Model–based analysis of a test that has
been well designed for cognitive skills diagnosis. We note that many
commercially produced tests currently provide subscores as “diagnos-
tic” score information, without any proper psychometric foundation. In
contrast to these practices, we indicate here how we have constructed
simplified subscoring systems that are based on a Fusion Model analy-
ysis and that can be shown to be close in accuracy to using the optimal
Fusion Model–based ppm.

In some settings, we may have skills coded to items in such a way
that each item has only one skill coded to it. We use the term “simple
structure” for such Q matrices. One possible method for producing a
subscore for each skill is to compute proportion correct or total scores
over just the items coded to that skill. If the Q matrix for a given test
includes items with multiple skills (as is often the case), we can still
define a subscore for each skill in exactly the same way. Once we have a
method for computing subscores, we need one more element to provide
mastery/nonmastery classification scores. A cut-point in the proportion
correct subscore is necessary to be able to translate the subscore into
a mastery/nonmastery designation. The choice of a good cut-point is
essential for good diagnostic performance of such a subscore method
for producing diagnostic profiles (see Henson, Stout, & Templin, 2005)
and is intimately related to standards setting.

In this section, we briefly discuss variations of a Fusion Model–
based subscoring approach that uses the results of a Fusion Model
parameter calibration to calculate appropriately weighted subscores
The Fusion Model Skills Diagnosis System

and reasonable model-driven cut-points. We discuss how to achieve adequate performance, compared to optimal Fusion Model mastery/nonmastery scoring, by defining reasonable subscores (either based on a simple structure $Q$ matrix as mentioned previously or a $Q$ matrix that includes multiple skills per item) of some sort and user-selected cut-points as indicators of mastery. We note that a poor choice of cut-points or substantially suboptimal subscores (i.e., subscores that fail to incorporate a substantial portion of the information that would be provided about skills by a full Fusion Model system analysis) will likely produce substantially suboptimal skills classification accuracy (Henson, Stout, & Templin, 2005).

In the case of a nonsimple structure $Q$ matrix, we have investigated methods for using the Fusion Model information to convert the nonsimple structure $Q$ matrix into a suboptimal but reasonable simple structure $Q$ matrix. Namely, each item with multiple skills is assigned to that skill $k$ for which the estimated $r_{ik}^*$ is smallest among all the skills assigned to the item by $Q$. Then that item will contribute only to the subscore for the $k$th skill. This is called simple structure subscoring. An alternative, and reasonable, approach is to leave the original $Q$ matrix unchanged and allow items that are coded for multiple skills to contribute to multiple subscores.

We note that if the original $Q$ matrix does include items with multiple skills, then both of these subscoring methods have been shown to be suboptimal compared to Fusion Model scoring (Henson, Stout, & Templin, 2005). Under the first subscoring method, in which each item is assigned to only one skill, an item contributes “credit” (from a correct response) or “blame” (from an incorrect response) to only one of its originally coded skills, and the “credit” or “blame” for the other skills of that item are missing from their subscores. Under the second subscoring method, in which the original nonsimple structure $Q$ matrix is retained and each subscore adds up all items that are coded to it, an item that is answered incorrectly now contributes “blame” to the subscores for all the skills coded to that item. But in a conjunctive setting such as the Fusion Model, in which successful execution of all required skills is needed for correct item performance, the reason for the incorrect item response may be the lack of mastery on just one of the skills. The second subscoring method has no mechanism for incorporating the fact that a given incorrect response on a multiple-skill item may be due to nonmastery on only one skill. Note that these missing credits or blamess also hold for weighted subscoring, which we discuss in the next paragraphs. So the proper
The question is how to make use of Fusion Fodel-based item information to minimize the credit and blame problems, and when we do that, how close do we come to the optimal Fusion Model mastery/nonmastery scoring.

The first thing we do is use weighting that takes advantage of useful skill discrimination information contained in the estimated item parameters. That is, some items are better indicators of mastery versus nonmastery when compared to others. For this reason, we suggest sometimes using weighted subscores, where the item weights should quantify how strongly each item measures the skill in question. We thus have three ways of forming subscores: two producing ordinary (i.e., unweighted) subscores, and the third producing potentially more informative weighted subscores.

An intuitively reasonable choice of weights is \( \delta_{ik} = \pi^*_i(1 - r^*_ik) \). For, smaller \( r^*_ik \) values and larger \( \pi^*_i \) both indicate that the item is very informative about skill \( k \) mastery. The weight \( \delta_{ik} \) ranges from 0 to 1 and can be interpreted as an indicator of how strongly skill \( k \) influences the item response function (0 means no impact, and 1 implies that it is impossible to get the item right when lacking skill \( k \); Henson, Stout, & Templin, 2005). The reporting of these weights is useful even if ppm's instead of subscores are being used to decide mastery/nonmastery. When examinees receive a score report indicating mastery or nonmastery on a given skill and are told which items measured the skill, they will naturally examine their performance on those items and will typically notice both correct and incorrect responses. The index can be used to better and more easily interpret their item scores, for example, by listing the items for them to inspect in order of highest to lowest (see score reports developed by Jang, 2006).

Using this weighting scheme, it is now possible to appropriately define a weighted sum of items that is especially sensitive to the skill \( k \) mastery/nonmastery dichotomy. Those items for which the \( k \)th skill is most influential will have the largest weights. We define the weighted complex skill \( k \) sum-score, \( W_{jk} \), for the \( j \)th individual as a weighted sum of those items required for skill \( k(q_{ik} = 1) \) as follows:

\[
W_{jk} = \sum_{i=1}^{f} \delta_{ik}(q_{ik}x_{ij}).
\] (10.3)

Unlike simple Q-based subscores, weighted Q-based subscores are heavily influenced by those items with small \( r^*_ik \) values relative to those
with large $r_{ik}$ values. For this reason, it is expected that the weighted ($Q$-based) subscores method of classification will have higher correct classification rates when compared to the ordinary (unweighted) $Q$-based subscore method, and simulation studies demonstrate this (Henson, Stout, & Templin, 2005). Based on these simulation results, it is also hoped that the performance of weighted subscores for many settings will be roughly as good as the correct classification rates produced by a full Fusion Model system analysis.

The next critical issue when using any weighted or unweighted subscoring approach is determining the cut-point at which a person is classified as a master. Henson, Stout, and Templin (2005) show that correct classification rates will be significantly lower than optimal if the cut-point for each subscore is selected in a way that ignores the Fusion Model parameters. In addition, Henson, Stout, and Templin (2005) show via simulation studies that by using a Monte Carlo estimation approach to determine the cut-point and using the aforementioned weighted subscores, the Fusion Model system’s classification performance can be fairly well approximated.

The simulation method used in determining a cut-point for subscore-based mastery classification use is as follows. The Fusion Model system is used to obtain item parameter estimates and estimates of the joint distribution of the examinee skills space (see description in Section 4.3 or, for more details, see Henson, Stout, & Templin, 2005). Next, 100,000 simulated examinees and their simulated item responses are generated. Then, using a simple search, an appropriate cut-point for the weighted subscore is determined that maximizes correct classification rates for each skill, as determined in the simulated data set. This is possible because the true skill pattern for each of the simulated examinees is known.

In general, even if a full Fusion Model analysis is used to help design an effective skills-level test for diagnostics and then to calibrate the test, it may be unreasonable, or seen as undesirable or impractical in some settings, to do a full Fusion Model system–based classification. In such cases, the Fusion Model–based weighted subscoring approach can provide a relatively efficient, convenient, intuitively plausible, and inexpensive model-driven alternative. In these cases, although the Fusion Model is not directly used for skill pattern estimation, valuable information concerning the calibrated Fusion Model is used to determine the weights to use for the weighted subscores and the “optimal” cut-point. Henson, Stout, and Templin (2005) showed via simulation studies
that although the Fusion Model system outperformed the subscoring approaches, in all cases the weighted subscoring Fusion Model–based method described previously provides a reasonable approximation. Henson, Stout, and Templin further showed that the unweighted subscoring approaches, perhaps mandated for the test setting, are clearly suboptimal relative to weighted subscoring but nonetheless can be reasonably effective. An upcoming paper by Henson and Templin (in press) will provide a detailed discussion of the subscoring approaches and the research results concerning them.

6.0 CONCLUDING REMARKS

This chapter has presented the Fusion Model skills diagnosis system, a comprehensive and ready-to-use data analysis system for conducting skills diagnosis. A detailed description of the Fusion Model system has been given in terms of four essential components required for a fully developed system: (a) the IRF model (RUM), (b) the MCMC estimation algorithm Arpeggio and its Bayesian modeling framework, (3) the collection of model checking procedures, and (4) the skills-level statistics that can be used for mastery/nonmastery scoring. Special emphasis was given to describing the probabilistic model and statistical methods constituting the Fusion Model system, outlining the user-focused results that the system reports, and explaining how to interpret and use these results. Although in this chapter the system was described as though it were restricted to dichotomously scored items and dichotomous skills (mastery vs. nonmastery), as noted at appropriate places in the chapter, the Fusion Model system is also applicable to polytomously scored items and/or polytomous skills (more than two levels of mastery). Indeed, the authors believe that these two polytomous directions will prove especially important in the future because of their potential to improve performance of diagnostic tests.

Educators have long implored the testing industry to more fully contribute to the formative assessment of learning, in particular by providing tests that can diagnose examinee mastery on skills of interest to teachers, parents, and other test users. Extensive research is underway to identify ways in which properly conceived and implemented assessment systems can improve teaching and learning, and lead to school improvement (see Pellegrino & Chudowsky, 2003; Popham, Keller, Moulding, Pellegrino, & Sandifer, 2005; Shavelson, Black, William, & Coffey, in press; Shepard, 2000). In the volatile accountability-focused
standardized testing atmosphere that currently exists in the United States, it is especially important to understand how formative and summative assessment systems can be aligned, and how they can be mutually beneficial and supportive. The development of practical, effective psychometric and statistical models and approaches, such as the Fusion Model system, represents one essential component of this multifaceted problem. Many educational measurement specialists, including the authors of this chapter, see this formative assessment challenge providing psychometric solutions that will advance the effectiveness of formative assessment in classrooms as not only intellectually exciting from the research perspective, but also societally important from the educational perspective.

Motivated by this formative assessment challenge, this chapter presents in some detail one major attempt at providing the testing industry with a practical set of statistical tools sufficient for conducting educationally useful skills diagnosis, namely, the Fusion Model system. In particular, the Fusion Model system provides a comprehensive set of statistical and psychometric tools ready to provide effective formative assessment at a more fine-grained level than currently provided by the ubiquitous unidimensional test scoring currently used. The Fusion Model system also provides tools to help improve the design of diagnostic tests.

As noted herein, this system’s performance has been heavily studied in simulation studies and has been applied in various real data settings. Particular applications include the briefly discussed Jang TOEFL work, as well as skills-level analyses of data from the PSAT/NMSQT and from the U.S. National Assessment of Educational Progress (NAEP). From the practitioner’s perspective, the next necessary direction is the conducting of pilot studies and field studies of the Fusion Model system to investigate and develop procedures for the implementation of the system within the context of operational testing programs.

Indeed many such settings seem potentially ripe for such applied work. These include state and district K–12 standardized tests, development of periodically administered skills diagnostic tests intended for classroom use, and integration of diagnostic assessments into test preparation programs for high-stakes tests such as the undergraduate, graduate, medical, and other school or professional admissions testing. Applications in industrial and military training settings seem promising, too, especially because such training settings often have well-defined standards of what must be learned in order for students to be declared
masters of the course material. Such standards would, of course, need to be converted into sets of skills by appropriately chosen experts. Future applications within 2- and 4-year colleges also seem promising. One example of a particularly intriguing and challenging application is the development of higher education courses in which diagnostic assessments serve the dual roles of formative assessment embedded in instruction and learning (based on student performance on periodically administered quizzes) and evaluation of overall summative assessment for traditional grading purposes.

Further psychometric and statistical challenges will arise in moving forward with the developments anticipated. However, the authors of this chapter believe that the status of the Fusion Model system is advanced enough to justify a concerted focus on the very real dual challenges of (a) learning to build effective diagnostic tests that are optimally tuned for strong diagnostic performance, and (b) addressing the substantial challenges in teacher professional development and development of curricular and instructional methods and systems that can benefit from such diagnostic assessment. Although our chapter focuses on the Fusion Model system, other systems at various stages of development are also available. Depending on their capabilities and performance properties, some of these systems will likely also be important in addressing this vital diagnostic educational challenge. Success in this diagnostic enterprise will depend vitally on high quality on at least three fronts simultaneously: the cognitive modeling of skills and skills acquisition; the preparation of teachers, curricula, and instructional programs and materials for a next generation classroom that is integrated with diagnostic assessment; and the psychometric quality of models and scoring procedures.

In summary, the Fusion Model system is well enough developed that it is ready for pilot-/field-testing in preparation for operational use, both (a) in existing settings where the summative test already exists and now needs enhanced scoring to carry out formative assessments, and (b) in contributing to the design of skills diagnostic formative assessments from scratch. Numerous interdisciplinary efforts are needed that draw on cognitive psychologists, teaching and learning experts, substantive curricular experts, practicing educators, and psychometricians to develop and make operational a new engineering science of informative skills level–based standardized testing. This chapter describes progress along the psychometric dimension specifically as represented by the Fusion Model system.
We conclude that the statistics and psychometrics concerning the Fusion Model system are advanced enough for developing a coordinated approach with the cognition, teaching, and learning components to significantly impact our vast and growing educational enterprise.

References


The Fusion Model Skills Diagnosis System


