DiBello and Roussos are co-first authors.

The second half of this chapter represents a focused updating of an earlier report by Roussos (1994). In addition we acknowledge the important review of statistical approaches to cognitively relevant assessment by Junker (1999), especially for the idea of grounding a discussion of psychometric models within the broader framework of assessment design. We follow that approach here.
ABSTRACT

This chapter is divided into two main sections. The first half of the chapter focuses on the intent and practice of diagnostic assessment, providing a general organizing scheme for a diagnostic assessment implementation process, from design to scoring. The discussion includes specific concrete examples throughout, as well as summaries of data studies as appropriate.

The second half of the chapter focuses on one critical component of the implementation process – the specification of an appropriate psychometric model. It includes the presentation of a general form for the models as an interaction of knowledge structure with item structure, a review of each of a variety of selected models, separate detailed summaries of knowledge structure modeling and item structure modeling, and lastly some summarizing and concluding remarks.

To make the scope manageable, the chapter is restricted to models for dichotomously scored items. Throughout the chapter, practical advice is given about how to apply and implement the ideas and principles discussed.
1 Chapter Preliminaries

As a context for describing existing approaches to cognitive diagnostic modeling, this chapter describes an assessment implementation framework, including assessment design and subsequent development, model specification, estimation, and examinee score reporting. In the second half, the chapter presents a survey of psychometric models, focusing primarily on several current diagnostic models and approaches. The chapter promotes further research along two streams: 1) research into substantive areas required for effective diagnostic assessment design, trial, and evaluation in educational testing settings that may benefit from skills diagnostic scoring, and 2) further statistical and psychometric investigations of these models, methods and applications. The terms “skills” and “attributes” are used interchangeably throughout.

2 Implementation Framework for Diagnostic Assessment

As noted by Junker (1999), the challenges of designing a diagnostic assessment, are: “how one wants to frame inferences about students, what data one needs to see, how one arranges situations to get the pertinent data, and how one justifies reasoning from the data to inferences about the student.” This recognition of assessment design as a process of reasoning from evidence was summarized in the NRC Report, Knowing What Students Know (Pellegrino, Chudowsky, & Glaser, 2001), as the “assessment triangle” consisting of three main components: cognition, observation, and interpretation. The evidence centered design (ECD) paradigm developed by Mislevy, Almond, and their colleagues (see, e.g., Almond, Steinberg, & Mislevy, 2003; Mislevy, Steinberg, & Almond, 2003; or Mislevy, Almond, & Lukas, 2004) frames that recognition into a systematic and practical approach to assessment design. In this section we adapt and augment the ECD paradigm to describe a framework for the entire implementation process for diagnostic assessment, elaborating on practical aspects of cognitively based assessment design and also going beyond design to illuminate practical issues in regard to estimation and score reporting.

We conceptualize the diagnostic assessment implementation process as involving the following main components:

(1) Description of assessment purpose,
(2) Description of a model for the latent attributes of diagnostic interest (the skills space),
(3) Development and analysis of the assessment tasks (e.g., test items),
(4) Specification of a psychometric model linking performance to latent skills,
(5) Selection of statistical methods for model estimation and evaluating the results, and
(6) Development of systems for reporting assessment results to examinees, teachers, and others.

As we will detail next, the components of a successful implementation process are necessarily nonlinear,
requiring considerable interaction and feedback between the components and demanding close collaboration between users, test designers, cognitive psychologists, and psychometricians.

2.1 Assessment purpose.

The purpose of the assessment should be clearly delineated, and this purpose has strong implications for the description of the latent attribute space. For example, if the purpose is to diagnose human competency on a user-selected set of multiple skills, this would seem to require an explication of what it means to be competent in each of the skills and would seem to point toward a small number of discrete levels of classification on each skill, e.g., the dichotomy of competent versus not competent. However, if the purpose of the assessment is a more traditional one of ranking examinees along some broad general competency scale, then an appropriate skill space selection would seem to be a single continuous scale.

In particular, the purpose of the assessment will have significant impact on whether the targeted latent attribute or skill space will be modeled with one or more than one variable, and whether the variables will be discrete for classification purposes or continuous for scaling purposes. It is essential to note that targeted skill inferences can be supported by the user-selected tasks only if proper attention is given to skill information provided by the tasks. Hence, the determination of purpose in the first step interacts with choosing tasks that provide appropriate information about skills.

Already, one can see the nonlinear nature of diagnostic assessment implementation in that components (2) and (3) above are necessarily partly initiated at step (1). However, it is still convenient to think of the components as separate, even though successful implementation requires tight interaction and integration among the six components.

There are two kinds of practical skills diagnostic settings considered: 1) analysis of existing assessment data using more complex skills-based models in hopes of extracting richer information than provided by unidimensional scaling or other existing analyses, and 2) designing a test from the beginning for a skills-based diagnostic purpose. In designing an effective diagnostic assessment, satisfying the assessment purpose is the prime motivation, so delineating assessment purpose should be accomplished with care. However, many of the applications with complex diagnostic models that are actually found in the literature are cases in which the skills diagnosis is conducted as a post-hoc analysis, called retrofitting by some, sometimes as a demonstration of a new statistical model and/or method, and sometimes as an attempt to extract richer information than the assessment was designed for. One good example of where the consideration of assessment purpose preceded and motivated the discussion of the development of a skills diagnosis implementation procedure is that of Klein, Birenbaum, Standiford, and Tatsuoka (1981), who investigated the diagnosis of student errors in math problems involving the addition and subtraction of fractions. Another
example is the LanguEdge study of Jang (2005, 2006; Roussos, DiBello, Henson, Jang, & Templin, anticipated 2006). The Jang study is an example of an existing test in the sense that the LanguEdge English Language Learning (ELL) assessment items already existed. Jang defined a new classroom purpose for this test and subsequently built a successful skills framework to extract diagnostic information that would be useful for teachers and learners in an ELL class. We cite both of these examples (along with a few others, as needed) throughout this chapter to illustrate important points about diagnostic assessment design and implementation.

2.2 Description of the attribute space.

As mentioned above, the assessment purpose leads naturally to the question of what is to be assessed about the test takers – what proficiencies are involved and what types of inferences are desired. The second component in our implementation process then requires a detailed formulation of the skills and other attributes that will be measured in order to accomplish the test purpose. In this step, a detailed representation of the skills or attributes space is developed in light of the purpose and based on the cognitive science, educational psychology, measurement, and relevant substantive literature. Often such literature will not lead to only one particular attribute or skills representation, and in such cases multiple representations should be considered before a selection is made. Jang (2005, 2006; Roussos, et al., 2006) provides an excellent example of such a detailed literature investigation as she considered how to represent reading comprehension in second-language learning. Jang reported great controversy as to whether reading comprehension was decomposable into skills, and that, among the many researchers who did believe in a skills representation, there was still further controversy about what exactly those skills should be. Jang based her final selection of the skills space representation on careful consideration of the several possibilities given in the literature and also on task analyses of component (3). Since in this example the tasks already existed, Jang carefully analyzed the tasks before settling on a final selection of the skills representation.

In general, although substantive cognitive theory is essential for successful design and implementation, we underscore here that the arbiter of which representation is selected for the attribute space is not solely or even primarily which representation is most cognitively or psychologically correct. The substantive theory and knowledge must always be considered within the context of the purpose of the assessment. If the purpose is mainly to rank examinees along a single competency scale, a unidimensional continuous ability model may be the best solution even though most cognitive psychologists would not seriously believe that such a model captures much of any depth about real cognition. Further, the number of skills cannot be overly large and must be controlled by issues of statistical tractability.
2.3 Development and analysis of the assessment tasks.

Logically, considering assessments as systems for reasoning about mental capabilities according to evidence from tasks administered to examinees, the choice of tasks should be based primarily on a consideration of the amount of evidence needed to support desired inferences about examinee attributes. Ideally, test developers should consider a wide variety of possible tasks, choosing feasible ones that best match the purpose of the assessment. As we demonstrate throughout the survey of the psychometric models in Section 4, task developers should not avoid tasks that require combinations of multiple skills per task.

In cases where the assessment instrument already exists, the question becomes how much information can be provided about the set of desired attributes by the test. In the preferred case when new tasks are explicitly designed for a new diagnostic assessment, the task development phase should also include a detailed analysis of the tasks to understand how many and what kinds of skills or attributes are involved, at what level of difficulty, and in what form of interaction. For example, Jang (2005, 2006; Roussos, et al., 2006), in analyzing an already existing test for use as a diagnostic assessment, identified possible examinee solution strategies for each of the tasks (the test items), conducted verbal think-aloud protocols with appropriate subjects, consulted the test specification codes from the original test developers, analyzed the task textual features (such as number of words in the task stimulus, difficulty of the vocabulary in the questions asked, etc.), and conducted a nonparametric latent dimensionality analysis using data that had been previously collected from administration of the test forms. Through the use of all these analyses together, Jang was able to develop a final skills space for her assessment design that simultaneously satisfied the new teaching and learning purpose of her assessment and was theoretically defensible from the viewpoint of existing linguistic literature. In reporting her results, she stressed that it was important that the number of skills not be so large as to be statistically unsupportable. In particular, she initially identified through substantive analyses 32 possible processes or features that could have been used as skills or attributes, a number she knew was not statistically supportable. She then reduced this to 16 and finally to 9, a number she found was both statistically and substantively supportable. She noted that further reductions by combining or eliminating skills, while possibly resulting in still increased reliability, would have not been supported by substantive theory in that either skills that are too distinct would have had to have been combined or skills would have been eliminated such that some items would have had no assigned skills. In general, the number of skills should not be reduced too much by deleting or combining so as to result in the remaining skills lacking diagnostic utility or interpretability. Making this compromise between the initially proposed set of skills and a set of skills that can be successfully statistically assessed is a vital aspect of carrying out a successful skills diagnostic analysis.

In developing the skills space, it is also important to keep in mind how the skills interact with each
other both in terms of correlational relationships among skills and the nature of the interactions among the multiple skills required by individual tasks. One of the more important distinctions for within-task interaction is one that we refer to as conjunctive versus compensatory. By “conjunctive” interaction of skills on a task, we mean a task for which successful application of all the required skills seems necessary for successful performance on the task as a whole – lack of competency on any one required skill represents a severe obstacle to successful performance on the task as a whole. By “compensatory” (or fully compensatory) interaction of skills on a task, we mean that a high enough level of competence on one skill can compensate for a low level of competence on another skill to result in successful task performance, even to the extreme “disjunctive” case in which the probability of successful task performance is high so long as competency occurs on any one or more of the required skills. For example, a disjunctive model is appropriate when the various attributes really represent alternative strategies for solving the item. In that case, successful performance of the item only requires that one of the possible strategies be successfully applied.

The choice of mode of attribute interaction clearly depends on the diagnostic setting, including the purpose of the assessment, and how the skills or attributes are defined. Jang (2005, 2006; Roussos et al., 2006) noted that depending on the skill representation she chose, she could have been led to either a conjunctive, compensatory, or a mixture of the two models. She purposely chose a conjunctive skill representation for her particular setting because based on her task analyses and her reading of the literature, most of the skill interactions seemed psychologically conjunctive in nature, and the few that were not were amenable to being combined into a single skill. Understanding of this type of attribute interaction within tasks is important because it will help determine which psychometric model is most appropriate and interpretable for the intended diagnostic assessment.

Another important part of understanding skill interactions is that there may be important relationships between the skills, such as, certain skills tending to co-occur on tasks, an ordering of the difficulty of the skills, or mastery of one skill only occurring when another skill has been mastered. Also, skill pairs tend to be positively correlated so that choosing models that presume this correlational structure can be useful. Including such information increases the accuracy of a skills diagnostic analysis if the added complexity can be reasonably well fit by the data.

Still another important outcome of the task analysis process is the development of specific task-skill coding rules that are used for an initial substantive assignment of skills to tasks so that independent raters will reliably agree on these assignments. The matrix that relates skill assignment to tasks is referred to as the \( Q \) matrix – the rows are the tasks, the columns are the skills or attributes, and the entries are 1’s and 0’s indicating, respectively, whether a specific skill is intended or not intended to be measured by a particular task. One can consider a proposed \( Q \) matrix as a hypothesized linking of skills and tasks,
based on theoretical and other substantive considerations. Subsequent statistical analyses can result in the initial $Q$ matrix being modified to improve statistical performance. An example of such a $Q$ matrix modification performed by Jang (2005, 2006) is briefly discussed later in this chapter. A successful skills diagnosis critically depends on high quality $Q$ matrix development.

Once diagnostic assessment data have been analyzed, one useful internal consistency check of the hypothesized $Q$ matrix is the consideration of whether there are any mismatches between observed or logically inferred task difficulties (proportion of test takers successfully performing each task) and the difficulty of the skills or attributes that are assigned to the tasks. For the dichotomous case of mastery/non-mastery of skills, we speak of the estimated population proportion of masters of an attribute as the difficulty of that attribute. For example, in the case of a conjunctive model, a $Q$ matrix entry should not assign a purportedly difficult skill to an easy task.

### 2.4 Psychometric model specification.

The model discussions in the second half of this chapter focus on component (4), the selection of the psychometric model. We focus primarily on psychometric models that explicitly contain multiple examinee proficiency variables corresponding to the skills or attributes that are to be diagnosed. Some researchers (other than the authors) claim the use of residual methods to derive diagnostic benefits from unidimensional item response theory (IRT) models can suffice. It should be noted that Tatsuoka’s rule space method (Tatsuoka, 1983, 1984, 1990) based upon unidimensional IRT ability and person fit statistics, can be considered a residual approach that does employ a $Q$ matrix. In the interests of space, we restrict our discussion mostly to models with an explicit $Q$ matrix or equivalent. The only exceptions are the standard unidimensional IRT models, which are included as a base of comparison because of their widespread use in educational measurement.

One important distinction between the models is that the proficiency variables may be modeled as discrete or continuous depending on the purpose of the assessment as described above. Another important distinction, also mentioned above, will be whether the required skills for a specific task interact in a conjunctive or compensatory manner. The psychometric model itself is a function specifying the probability of a particular task or item response in terms of examinee skills and item characteristics (the item parameters). For a given set of skill variables and a type of skill interaction (conjunctive versus compensatory), a variety of particular models with varying simplifying assumptions may be chosen from. For example, one might assume that the item success probability is the same for all examinees who are non-masters of any one or more required skills, regardless of how many or which required skills are non-mastered. Another simplifying assumption is that items that involve the same skills have the same values for their item parameters and
hence the same success probability for each examinee. These types of assumptions reduce the number of item parameters to be estimated, thus reducing standard errors of estimation. This can be especially useful when the total number of items measuring a given attribute is small or if the number of examinees in the sample is small. But these kinds of parameter reductions may also introduce unwanted bias if the assumptions are not warranted.

2.5 Model calibration and evaluation.

In the next step of the implementation process, the selected model must be calibrated (estimation of both item parameters and examinee population parameters), the calibration must be evaluated, and then examinees classified on the skills (e.g., for the dichotomous classification case, estimation of mastery versus non-mastery of each skill for each examinee). Models with large numbers of skills and/or large average numbers of skills per item are in particular danger of being non-identifiable (see further discussion below). Moreover, even if identifiability holds, test designs must provide sufficient numbers of items per skill, and not too many skills per item in order for accurate estimation to be possible.

The reality of much diagnostic testing is that for tests of reasonable length the number of items that contribute to the measurement of each skill will be relatively small compared to the numbers of items that typically contribute to the estimation of unidimensional scale values. For this reason, it is necessary to keep models as simple as possible while satisfying the constraints imposed by the diagnostic purpose. Even the simplest models that are able to satisfy the test purpose often will exhibit significant parametric complexity, at least more than is encountered in typical unidimensional models. Careful consideration of both parameter identifiability and systematic model evaluation help achieve effective calibration.

The discussion in this section is broadly divided into two parts: (1) estimation and computation and (2) evaluation of the estimation results.

2.5.1 Commonly used estimation and computational methods.

Commonly used estimation methods. In principle, models can be non-Bayesian or enhanced in a partially Bayesian or fully Bayesian manner. That is, model parameters can be posited in one of three ways: all parameters are fixed without prior distributions, a portion of the parameters can have prior distributions imposed (such as for the latent skills mastery/non-mastery vectors), or all of the parameters can be assigned a joint prior distribution.

Related to this selection of the level of Bayesian involvement in the model, three general estimation approaches have come to be discussed, some heavily used and others having the potential to be used. These are joint maximum likelihood estimation (JMLE), which seems natural for the non-Bayesian models,
maximum marginal likelihood estimation (MMLE), which has been widely used for the partially Bayesian models that place a prior just on the population parameters for the latent examinee skills vector space, and Bayesian Expected a posteriori (Bayes EAP) for fully Bayesian models that place priors on all model parameters. Bayes maximum a posteriori (Bayes MAP) is possible also. Although conditional maximum likelihood is possible (see, for example, Junker, 1999, p. 67), we are unaware of any serious applications in skills level assessment. In many real data examples with fully Bayesian models, instead of analytically computing the Bayes EAP, Markov Chain Monte Carlo (MCMC) estimation is used to simulate the Bayesian posterior distributions of parameters. Expectation maximization (EM) computational approaches for fully or partially Bayes models are sometimes used too.

Commonly used computational methods. Because the models are usually parametrically complex and the data sets are often large (e.g., 25 or more items and 2000 or more examinees), the choice of computational method becomes an important practical issue. First, we note that a direct application of JMLE seems largely unused currently in the skills diagnostic setting, partially because of the lack of good theoretical asymptotic consistency results, which is true even for parametrically simple scaling models like the two-parameter logistic model. However, the ever increasing speed of low-cost computing seems to allow for the practical use in fitting skills diagnosis models of the JMLE exhaustive search and divide-and-conquer approach. In such an approach, one fixes item parameters while estimating examinee parameters, then proceeds vice versa, and iterates until some convergence criterion is satisfied.

As discussed above, an alternative estimation approach to JMLE is MMLE, which can be implemented by providing an examinee skills vector distribution and integrating over that distribution. The EM computational procedure can be used for this purpose. In the empirical Bayes tradition, the examinee skills distribution could be and sometimes is updated using the observed data, perhaps by using a hierarchical Bayes approach. Once the model has been thus calibrated, a straightforward likelihood approach can be used to do examinee skills classification.

MCMC is a probabilistically based computational method that is used to generate Markov chains of simulated posterior distributions for all the model parameters given the data (see Gelman, Carlin, Stern, & Rubin, 1995). 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, anticipated 2006) and the NIDA and DINA models (de la Torre & Douglas, 2004).

It is interesting to note that MCMC could be used from either a partially Bayesian or non-Bayesian perspective. However, the most common application seems to be for fully Bayesian models, often with hierarchical structure to give it an appropriate empirical Bayes flavor. The primary inferential outcome is
the joint posterior distribution of the item and examinee parameters. Many MCMC references exist, and the Patz and Junker (1999a, 1999b) papers and the Gelman et al. (1995) book are especially recommended to readers considering parametrically complex models such as the models surveyed here.

EM algorithms have also been used in estimating skills diagnostic models, for example by von Davier in calibrating his General Diagnostic Model (GDM) (von Davier, DiBello, & Yamamoto, in press; von Davier, 2005, von Davier & Yamamoto, 2004). The EM algorithms are often much more computationally efficient relative to MCMC algorithms. On the other hand, the EM algorithm tends to be more difficult to extend to new models or model variants compared to MCMC. As Patz and Junker point out, EM algorithms are not as “straightforward” (1999a, p 14; see also Junker, 1999, p. 70) to apply for parametrically complex models. Further, MCMC provides a joint estimated posterior distribution of both the test’s item parameters and the examinee skills parameters, which may provide better understanding of the true standard errors involved (as argued by Patz & Junker, 1999a, 1999b). Moreover, MCMC routines adapt easily to produce posterior predictive model diagnostics. MCMC also provides a ready capability for comparing model parameter prior and posterior distribution as a measure of parameter identifiability (Sinharay, 2006).

2.5.2 Broad-based evaluation of estimation results and connections with validity.

When a complex diagnostic model is applied to data, the results of the estimation process must be carefully scrutinized. In this section, we deliberately broaden the typical model checking process to include evaluation of statistical convergence, interpretation of parameter estimates for the items and the ability distribution, traditional model fit statistics, reliability, internal validity, and external validity. In particular, we emphasize the intertwining of psychometrically oriented validity assessment along with essential statistical model fit procedures in evaluating the appropriateness and effectiveness of the statistically calibrated model.

Evaluation of Statistical convergence. No matter which estimation method is used, all statistical estimation methods require checking for convergence – either convergence to within some specified tolerance as in a JMLE or EM algorithm or convergence to the desired posterior distribution 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.

If non-convergence occurs, one can revisit the model building steps and reconsider the Q matrix and the selected model to see where changes may be warranted. For example, if the model assumes that a skill is difficult, yet the skill is assigned to items having a large range of difficulty, the MCMC algorithm may not be able to converge to a single level of difficulty for the skill. There is much more to say about MCMC convergence. The reader is referred to Cowles and Carlin (1996), Gelman et al. (1995), and Sinharay (2004)
for excellent and thorough discussions.

Interpretation of model parameter estimates. Given that the parameter estimation procedure has converged, the estimates for the ability distribution and item parameters should be evaluated for internal consistency, reasonability, and concurrence with substantive expectations. For example, a key issue for mastery/non-mastery diagnostic models is whether the proportion of examinees estimated as masters on each skill is relatively congruent with substantive theoretical expectations. If a skill turned out much harder or easier than expected (e.g., from a standard-setting perspective), 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 for that skill can be revisited to see if other more appropriate tasks can be found (e.g., if the proportion of masters for a skill is too low, one could try replacing the harder tasks for that skill with easier ones). 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.

In Jang (2005, 2006; Roussos et al., 2006), one of the skills was found by statistical analysis to be easier than expected whereas the other eight skill difficulties were ordered as expected. In this case, Jang did not feel it necessary to conduct any further modifications for purposes of her study, but she plans future work to investigate this particular skill.

Next, the item parameters should be inspected in detail because they play a key role in determining the success of the diagnosis. For every model, the item parameters indicate how well the items performed for diagnostic purposes. The item parameters may indicate, for example, how well each item discriminates for each skills between masters and non-masters. In the analysis of Jang (2005, 2006; Roussos et al., 2006), 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. In general, such a reduction in number of parameters may be beneficial simply because fewer parameters can be better estimated with the same observed data; but, more importantly, eliminating non-significant $Q$ matrix entries reduces the very real possibility of confusion that can arise from examinees labeled non-masters performing nearly as well as those labeled masters on the items corresponding to the $Q$ entries being eliminated.

Model checking statistics. Standard definitions of model checking appropriately focus on assessing the degree of fit between the estimated model and the observed data. For skills diagnostic implementation success and, hence, for this review, we expand this focus to include some aspects of consequential validity. This broader perspective is intended to subsume, refocus, and broaden traditional statistical model-checking when specialized to skills diagnostic assessment. More precisely, as mediated by the assessment purpose, enacted by the assessment design, and controlled by the skills classification setting (particularly testing constraints), the goal is that the choice of the model and the subsequent model estimation and examinee
skills classification provide accurate enough calibration to allow the practitioner’s skills-level assessment goals to be met. Much the same argument can be made for any type of assessment. The satisfaction of appropriate validity standards is essential to achieve success. Because of the greater complexity of diagnostic assessment, it is especially critical that validity concerns are explicitly addressed from the initial phases of determination of test purpose and assessment design through the entire implementation process.

A concrete example is a study by Zhang, Puhan, DiBello, Henson, and Templin (anticipated 2006) in which four skills are measured in an English Language Learning test of 40 items. In this particular case the standard two-parameter logistic (2PL) IRT model did a better job of predicting total score (oral communication, DiBello, 2006). In part, the reason was that students who were classified as lacking all skills or having mastered all skills were not broken down as finely as they were on the estimated continuous IRT ability scale. Consequently, in a comparison of model fit between the more complex skills-diagnostics-focused model versus the unidimensional scale-focused model with fewer parameters, the unidimensional model performed better than the complex model at predicting total score. But the unidimensional model did not provide a direct basis for inferring which of the specific skills were mastered and not mastered. This particular total-score-based fit statistic was not germane to the consistency of the skills classification purpose.

The development of appropriate fit statistics that are sensitive to the skills diagnostic purpose of assessment is an important area for further research. Model fit evaluation must be sensitive to the primary diagnostic purpose of the assessment. For example, in the case of posterior predictive model checking mentioned later, the discrepancy measures investigated should be sensitive to aspects of the data consequential for diagnostic performance.

Once a model has been calibrated from data, one relatively straightforward and easily interpretable approach to evaluating model fit is to compare observed and model-predicted statistics. When MCMC has been used to produce posterior distributions for the model parameters, one can simulate from these posterior distributions and estimate the predicted distribution of a statistic or discrepancy measure based on the simulated data, and then compare the observed value of the statistic with the predicted distribution. This particular approach is called posterior predictive model checking (see Gelman & Meng, 1996; Gelman, et al., 1995; Sinharay, 2005; and especially Sinharay, 2006, which is a skills diagnosis application). From our perspective of consequential validity used to augment classical model fitting for diagnostic assessments, the challenge in applying posterior predictive model checking and other common fit techniques is selecting test statistics that reflect aspects of data that are important for the primary goal of diagnostic assessment. For example, the Bayesian RUM (Reparameterized Unified Model) approach (Henson, Roussos, & Templin, 2004 & 2005) checks both item difficulties and item pair correlations.
As a specific example, in the application of Jang (2005; Roussos et al., 2006), 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. Shown in Figure 1 below is a comparison between the observed and predicted score distributions from the real data analysis of Jang (2005, 2006; Roussos et al., 2006). The misfit at the very lowest and highest parts of the distribution were expected as the mastery/non-mastery 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/non-mastery rather than to scale examinees, this misfit actually had no effect on the mastery/non-mastery classification.

Other standard fit statistics have also been used in real data analyses. Sinharay (2006) examined several model diagnostic approaches applied to a real data analysis of the mixed-number subtraction data of Tatsuoka (1984) and Tatsuoka, Linn, & Yamamoto (1988), using a DINA-like Bayes-net model due to Mislevy (1995). Sinharay’s methods demonstrated several instances of poor model fit, as well as severe non-identifiability of model parameters, as determined by comparing prior and posterior distributions of item parameters. Although the focus in the Sinharay (2006) article was Bayes nets, the model fit diagnostics and approaches are broadly applicable to cognitive diagnosis models such as those discussed here.

Another standard fit statistic is the log-likelihood statistic that is typically used to compare models, especially nested models. This statistic with others was employed by von Davier (2005) in an analysis of ETS (Educational Testing Service) TOEFL (Test of English as a Foreign Language) data in which he compared the predicted log likelihood of the manifest distribution for a compensatory mastery/non-mastery diagnostic model and for a unidimensional 2PL model. It was interesting to note the 2PL model had the larger likelihood. This was a case in which the Q matrix was nearly simple structure (four skills with most items having only one skill), and the estimated posterior probabilities of mastery were highly correlated with 2PL scale score. In such a case as this, one could make the argument that most of the available compensatory classification information is captured by the scale score, and so the lower-reliability, less-well-fitting classification information may not add substantial new information to the scaled score. This is an example where caution should be exercised in applying a multidimensional skills diagnosis model in a situation where a unidimensional model better fits the data.

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 skills classification reliability can be conceptualized in terms of the twin notions of, on the one hand the correspondence between inferred and true attribute state, and on the other hand the consistency of classification if the same assessment were administered to the same examinee twice. The reader is referred to Henson, He, and Roussos (in press).

For example to estimate classification reliability, Henson et al. (2004) use the calibrated model to generate parallel sets of simulated data, estimate mastery/non-mastery for each simulated examinee on each set, and calculate the proportion of times that an examinee is classified correctly according to the known true attribute state (thus producing an estimate of correct classification rate for the attribute/skill) and proportion of times an examinee is classified the same for that skill on two parallel tests (thus producing an estimated test-retest consistency rate). These agreement rate calculations based on simulated data provide reasonable estimates of actual rates. The calculated rates also can be adjusted for agreement by chance, for example with the Cohen kappa statistics. Zhang, Puhan, DiBello, Henson, and Templin (2006) applied this procedure in an analysis of large scale English language assessment data (about 33 items per form, 2 forms, 4 skills, approximately 1.3 skills per item and 11 items per skill) where they typically found skills classification test-retest consistency indices of 0.8 with the Cohen’s Kappa value being typically 0.6. These levels of skills classification reliability could be judged sufficient for low-stakes uses.

Internal validity checks. Several measures of internal validity are important to evaluate. The examinee “scores” of interest are the proficiency estimates resulting from using the psychometric model. One kind of internal validity check is to measure the differences in observed behavior between examinees who are classified differently. We call these statistics “internal validity checks” because they use the test data itself to help verify the authenticity of the model. As one illustration of such a procedure that has been developed in the case of a conjunctive model (Hartz & Roussos, in press; Roussos, DiBello, & Stout, 2006), one defines an examinee to be a master of a particular item (an “item master”) if the examinee is classified as a master of each of the skills required by that item. For each item, this separates the set of examinees into two subsets: the set of all item masters of that item, and the set of all examinees who are not masters of that item. A simple comparison used in the “IMstats” procedure of Hartz and Roussos (2005) is to note the average observed proportion of examinees answering the item correctly among the item masters versus the item non-masters. If these two proportions are close, then the inferred skill classification has little effect on performance of that item. That indicates that the item, the $Q$ matrix and the skills coding of the item should be investigated. A similar measure, called EMstats, has been developed for examinee performance by Hartz and Roussos (2005).

As an example of the use of the IMstats internal validity check in a real data analysis, we present in Figure 2 the IMstats results from Jang (2005, 2006; Roussos et al., 2006) for one of the test forms she
analyzed. These results indicate a high degree of internal validity for the skills diagnosis as the mean score differences between item masters and item non-masters are quite large for the vast majority of the items. The results also point to certain problematic items. Jang investigated these items and discovered that they 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 is not necessarily the best one for doing skills diagnosis, and the identification of these problematic items through the internal validity analysis drove home this point.

External validity checks. We use “external validity” to refer to validity as it is usually thought of — statistics that relate test-derived ability estimates to some other ability criterion external to the test. One example of such research is a study in which Tatsuoka and Tatsuoka (1997) conducted skills diagnosis in a classroom setting (more than 300 students in all) performing diagnosis at a pretest stage, providing remediation instruction based on the diagnosis, and then evaluating the effectiveness of the diagnosis-based instruction with a post-test diagnosis. The results indicated that the diagnosis was very effective. For example, as shown in Table 1 below (Table 5 in Tatsuoka & Tatsuoka, 1997), for the 114 students diagnosed with serious errors at pretest, the diagnosis-based remediation resulted in 93 of them improving to having only nonserious errors, while only one of the 108 students who had nonserious errors at pretest was assessed as having a serious error at posttest. Jang (2005; Roussos, et al., 2006) also conducted skills diagnosis in a pretest-posttest setting, but on a much smaller sample (17 students). Another example is a protocol study by Gierl (1997) comparing diagnostic classification results to observed student performance on algebra items from a large scale standardized test. There is a need for more validity studies like these, but also including control groups or alternate treatment groups, which were not used the above studies.

2.6 Score reporting.

The last step in our framework for a diagnostic assessment implementation process is the development of the score reports, which, for example, would go to students, teachers, and parents in an educational diagnostic assessment setting. Such score reports must be easy to read and interpret and, most importantly, must fit into the use environment. For example the information on the score report must be useful to teachers and
students. From our perspective of consequential validity, the score reports may include action advice such as prescribed instructional or learning activities for a given classification outcome. One example of a diagnostic score report in current practice is the College Board’s “Score Report Plus™” that is sent to each student who takes the PSAT/NMSQT™ (Preliminary SAT/National Merit Scholarship Qualifying Test). This is the first nationally standardized test to give some limited diagnostic skills-based feedback (other than the use of subscores or proportions correct on subsets of items). The information provided for the PSAT/NMSQT is based on the diagnostic methodology of Tatsuoka (1983). Interested readers are strongly encouraged to visit the College Board website (www.collegeboard.com/student/testing/psat/scores/using.html) that shows what the score report looks like and gives links for further information. Jang (2005; Roussos, et al., 2006) developed score reports that give more extensive information and is the best example we know of a complete user-friendly diagnostic report.

3 Reliability, Validity, and Granularity

Determining the appropriate level of reliability and validity necessary for successful educational applications of cognitive diagnostic assessment is not well researched. Measures of success depend as much on user needs, user environments, and market acceptance as on technical aspects of reliability and validity. But it seems clear that more work is needed for determining standards of sufficient statistical quality for various types of diagnostic assessment.

Within the IRT psychometric paradigm there is a clear trade-off between increasing model complexity to more closely align the psychometric model with skills-based psychological models for test takers at the expense of increased estimation error and even possible non-identifiability, as opposed to improving reliability that can be achieved with simpler, low dimensional continuous proficiency representations, at the possible expense of low congruence with cognitive theory. This trade-off can be thought of as a version of the well-known statistician’s bias-variance trade-off. In fact the pursuit of cognitive diagnosis within the IRT probability modeling framework enforces strong feasibility constraints on modeling complexity. Because of the combined effects of non-identifiability along with the danger of having a large number of parameters, the best we can expect to do is to modestly expand model complexity from a single ability variable to a few skill proficiencies – on the order of ten skills or so for a standard length test. A 40-item mathematics test, for example, could not possibly statistically support diagnosis of 50 skills, even if there are compelling substantive rationales for the 50 skills. We believe that advancing psychometric and test development capabilities so that 40-item mathematics tests can be used to effectively diagnose, say, 5 or 6 specified skills will provide significant teaching and learning benefit.

Two strong guidelines are presented for practitioners wishing to carry out successful, psychometrically
rigorous skills diagnosis. First it is necessary for reasons of feasibility to keep the models as parametrically simple as possible within the diagnostic purpose of the test. Second, the question of how simple can a diagnostic model be and still accomplish the diagnostic purpose is not as straightforward as it may sound and requires close collaboration between statisticians and cognitive scientists. In our view, parsimony is important, provided it is correctly applied. The aim is not to achieve the simplest model that fits the data, but rather the simplest model that accomplishes the diagnostic purpose and fits the data reasonably well. Thus, one could choose a more complex, but still slightly less well fitting model, provided substantive evidence supports the model.

Let us consider how these principles apply to a few real assessment examples, moving from unidimensional assessments to multidimensional examples. One prominent application setting for skills diagnosis models is that of standardized educational achievement tests, which are designed to measure a single dominant dimension. The dominant dimension, such as math or verbal proficiency, may be decomposed into a set of skills, but they are generally highly correlated with each other. In other settings, however, such as classroom applications of specially designed instruments for diagnosing math errors as analyzed by Tatsuoka and Tatsuoka (1997) or in application of instruments developed for diagnosing the presence of psychological disorders as analyzed by Templin and Henson (in press), the assessment instruments are designed to be more strongly multidimensional.

In cases where a single dimension tends to dominate, low-dimensional or even unidimensional skills diagnosis models may be appropriate. (Here, the number of skills to be diagnosed can be much larger than the number of latent dimensions, which may even be 1. In such cases the $Q$-matrix-based skills diagnosis model might be combined with a standard unidimensional modeling assumption to enrich the unidimensional score scale with interpretable skill information, such as identifying a correspondence between specific skill masteries and thresholds on the score scale.

In the more strongly multidimensional cases, the skills diagnosis model provides still further information beyond that which a single unidimensional scale can provide. For example, Tatsuoka and Tatsuoka (1997) provide a tree diagram for nine knowledge states where the different states of mastery cannot be ordered by a unidimensional scale. Templin and Henson (in press) provide an example in a psychological testing setting where data were analyzed from an instrument used to diagnose pathological gambling based on 10 DSM (Diagnostic and Statistical Manual of the American Psychological Association) criteria. They forcefully demonstrated that the inferred latent criteria were strongly multidimensional (e.g., criteria 3, 6, and 7 were practically uncorrelated with the other seven criteria), and they emphasized that the criteria-profile they estimated for each individual has the potential to lead to better prevention and treatment programs than would result from merely estimating whether or not each individual is a pathological gambler.
It should be stressed here that careful and systematic design for cognitive diagnosis is necessary but not sufficient. That is, given a diagnostic test purpose, an identified skill space structure, and a selection of tasks that experts believe will measure the designated skills, it is still necessary through data analysis and various model evaluation steps to confirm that the intended and expected diagnostic performance did indeed occur. At the pilot data analysis stage, it may be helpful to replace one psychometric model with a simpler model to achieve improved statistical estimation and diagnostic performance, for example by imposing further constraints to reduce the number of item parameters. The key consideration from the point of view of diagnostic assessment is that the model simplifications be made within the framework of the diagnostic purpose. If a diagnostic test is desired to assess examinee skill levels, for example, on four user-defined skills, then it is necessary to verify that the proposed four-skill diagnostic instrument is statistically sound, in addition to being valid from a design perspective. Simpler four-skill models may result in improved diagnostic performance by reducing estimation variability. Even the originally proposed skills space and task designs may need to be thoroughly revisited to determine if they can be improved.

Such statistical supportability procedures are no different in spirit from the rigorous statistical analyses done to support a unidimensional test model. Finally, if there appears no reasonable way to statistically support diagnosis of four distinct skills with the test, a more fundamental redesign from first principles is necessary, starting by reviewing the purpose and proceeding through the components of the implementation process.

To engage with the test client and the design and development teams to discover why a particular four-skill design did not achieve statistical supportability will provide a valuable service toward clarifying and satisfying the ultimate diagnostic testing needs. For example, some modification of the original four skills model may indeed achieve statistical supportability and thus satisfy the user’s assessment purpose. The purpose-driven statistician’s role is essential within the implementation process to help the client find a way to satisfy their diagnostic assessment needs, provide hard scientific information when their goal has been reached, and determine the degree of statistical quality.

4 Summary of Cognitively Diagnostic Psychometric Models

In the last two decades psychometricians have proposed a number of models apropos to cognitive diagnosis. This section reviews and compares these models within an overarching framework that allows practitioners to see clearly the differences and similarities in the models. Although the current treatise is limited in scope to models for dichotomously scored items, the general model for dichotomous items can be easily generalized to polytomously scored items, and the organizing structure is equally applicable to polytomous-item models.

The framework for discussing the models consists of the following sections:
Section 4.1: General model for item-knowledge interaction.

Section 4.2: Specific models for item-knowledge interaction.

Section 4.3: Knowledge structure modeling assumptions.

Section 4.4: Item structure modeling assumptions.

Section 4.5: Summary table of the models.

Section 4.1 presents a general model that encompasses all the specific cognitive diagnosis models reviewed in this chapter. It reviews the deterministic cognitive diagnosis model underlying this general model, introduces relaxations of the deterministic model along with psychological explanations for these relaxations, and presents a generalized model equation. We also briefly discuss two other general model frameworks that readers may find useful to further investigate.

Section 4.2 then reviews each of the specific models included in the chapter describing how each represents the interaction of item structure with knowledge structure. Unidimensional item response theory (IRT) models are included even though they are usually not used as, nor intended be, cognitive diagnosis models. They are included as a basis for comparison because they are well known for modeling the performance of examinees on tests and, as such, help motivate modifications needed to produce effective psychometric models for skills diagnosis. We also include a brief description of Fischer’s (1973, 1983) LLTM even though it does not model the ability variable as a multidimensional vector of multiple cognitive skills. As we describe below, we see LLTM’s use of a $Q$ matrix as an historically pivotal development that lies mid-way between unidimensional models such as the Rasch model and fully diagnostic models such as those on which we focus here.

Table 2 presents a list of the models included in this review along with pertinent references and with the abbreviations that will be used to refer to these models.

---

We should note that Adams, Wilson, and Wang (1997) introduced a model that is actually more general than the MIRT-C model of Reckase and McKinley (1991) and includes some of the other models. Similarly, von Davier (2005; von Davier, DiBello & Yamamoto, in press; von Davier & Yamamoto, 2004) presented a generalized form of the compensatory model of Maris (1999), within a much more general modeling framework. Our chapter focuses on specific instantiations of these models that have been presented in the literature. These more general modeling approaches of von Davier and of Adams et al. are discussed in Section 4.1.3 below.

Because our summary generally focuses on models that explicitly model the specified skills or attributes of interest as delineated by a $Q$ matrix (or equivalent), we do not present general latent class models or
loglinear models (see, e.g., Hagenaars, 1993; Rost, 1990; or Kelderman & Macready, 1990) or even skill mastery models that do not involve \( Q \) in their model structure (e.g., Macready & Dayton, 1977, 1980).

After describing the above models in detail in section 4.2, sections 4.3 and 4.4 separately focus on the knowledge structure and the item structure, respectively, comparing and summarizing how the different models represent each of these components.

For simplicity in the current chapter, we will generally use the term “skills” to refer to the latent traits or, more generally, cognitive attributes or components being measured on examinees or test takers. We recognize that the latent examinee characteristics of interest do not always fit well with the use of this term (for example, in diagnosing the amount of knowledge an examinee has or in medical diagnosis). We will also use the term “ability parameter” as a general way of referring to any variable used to model the level of examinee proficiency or standing on a skill or attribute.

The goal of a model is to represent the performance of an examinee on an item based on the skills required in responding to the item and the proficiency of the examinee on these skills. The general idea behind all the models is that the higher the proficiency of an examinee on the skills required for the item, the higher the probability the examinee will get the item right.

4.1 General model for item-knowledge interaction.

The general model form that is presented here is a slight generalization of the form described in DiBello, Stout, and Roussos (1995). The psychological justification/explanation for the model is also taken from DiBello et al. (1995).

4.1.1 Deterministic cognitive diagnosis model.

To arrive at the general model form, we first start with a description of the deterministic psychological model for cognitive diagnosis. We begin, as in Tatsuoka (Tatsuoka, K. K., 1984, 1985, 1990; Tatsuoka, K. K. & Tatsuoka, M. M., 1987), with a test of \( I \) items, and a list of \( K \) skills of interest for diagnosis on the test. Let \( Q \) be the \( I \times K \) skills-by-items incidence matrix that specifies which skills must be mastered in order for an examinee to give a correct answer on each item:

\[
q_{ik} = \begin{cases} 
1 & \text{if item } i \text{ requires skill } k. \\
0 & \text{otherwise.}
\end{cases}
\]

Implicit in the selection of these skills is a particular solution strategy. An alternative solution strategy would likely result in a different list of skills, though a large degree of overlap would typically be expected.

Next, we define for the \( j^{th} \) examinee a vector \( \alpha_j = (\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{Kj}) \) denoting the state of mastery of the examinee on each of the skills. In this deterministic psychological model of cognitive diagnosis, mastery
is modeled as a dichotomous variable (though, in general, examinee mastery could be modeled with either dichotomous, polytomous or continuous variables):

\[
\alpha_{jk} = \begin{cases} 
1 & \text{if examinee } j \text{ has mastered skill } k \\
0 & \text{otherwise.} 
\end{cases}
\]

Finally, we assume that the cognitive requirement for the multiple skills within an item is conjunctive – that is, answering the item correctly requires mastery of all the skills required by that item. Then the probability an examinee correctly responds to an item is determined from the following deterministic model:

\[
P(X_{ij} = 1 | \alpha_j) = \begin{cases} 
1 & \text{if } \alpha_j \text{ contains 1s for all skills required for item } i \\
0 & \text{otherwise.} 
\end{cases}
\]

Thus, a computationally simpler form would be:

\[
P(X_{ij} = 1 | \alpha_j) = \prod_{k=1}^{K} (\alpha_{jk})^{q_{ik}}
\]

4.1.2 Relaxations of the single-\(Q\) deterministic model.

As described above in this deterministic model, knowing an examinee’s state of mastery on the skills of an item determines whether an examinee gives a correct response on the item. Though not explicitly mentioned above, this deterministic assumption holds also for application of the skills. In other words, if an examinee has mastered a skill, then the deterministic model assumes the examinee has a probability of 1 of successfully executing it on an item that requires it; and, conversely, if an examinee has not mastered a skill, then the examinee has a probability of 0 of successfully executing it on an item that requires it. In reality, however, other factors play a role in whether an examinee gets an item right and DiBello et al. list four that seem the most salient:

**STRATEGY:** The examinee may use a different strategy from that presumed by the skills listed in \(Q\).

**COMPLETENESS:** An item may require skills in addition to those listed in \(Q\). If so, then \(Q\) is said to be not complete for that item.

**POSITIVITY:** An examinee who has “mastered” a skill will sometimes fail to execute it correctly; and, conversely, an examinee who has *not* “mastered” a skill will sometimes execute it correctly. This lack of perfect positivity can be explained by how, in practice, skills are defined. From a practical point of view, skills cannot be overly atomistic (i.e., indivisible). Otherwise there would be so many fine grained skills that it would be difficult to estimate the mastery of all of them separately. Because the defined skills are divisible, any possible definition of mastery of a skill must always allow some lack of perfect positivity. Some models deal with positivity by representing skill mastery level with continuous variables and employing item response functions (IRFs) that are sensitive to the continuous scale of the skill mastery variables. Other
models maintain the representation of skill mastery as dichotomous variables and deal with positivity by relaxing the probability of 1 given mastery to allow a probability somewhat less than 1 and/or by relaxing the probability of 0 given non-mastery by allowing a probability somewhat greater than zero.\footnote{It is not difficult to extend the dichotomous mastery models to allow for multiple discrete categories.}

SLIPS: The student may commit a random error. This is a “catch-all” category for all the remaining non-systematic error in an examinee’s responses. One example of how this could be manifested is when an examinee accidentally bubbles in the wrong response category on a multiple-choice test.

Presented next is the general form of the cognitive diagnosis model – the interaction of item structure and knowledge structure that is based upon the above psychological reasoning.

First, some additional notation needs to be defined:

\( \nu_i \) = the number of strategies for item \( i \).

\( S_{li} \) = examinee applies \( l^{th} \) strategy of the \( \nu_i \) strategies for item \( i \).

\( \psi_j \) = skill mastery proficiency vector for examinee \( j \).

\( C_{li} \) = [Examinee correctly executes all the cognitive components for strategy \( l \) on item \( i \)].

\( p_s \) = probability of a random slip.

The general form of the cognitive diagnosis model is then given by:

\[
P(X_{ij} = 1 | \psi_j) = (1 - p_s) \sum_{l=1}^{\nu_i} [P(S_{li} | \psi_j) \cdot P(C_{li} | S_{li}, \psi_j) \cdot P(X_{ij} = 1 | S_{li}, C_{li}, \psi_j)].
\] (1)

The equation can be most easily understood as mimicking an abstract summary of the cognitive progression of an examinee as the examinee solves the item. The entire equation is based on the examinee not committing a random slip, so it begins with the probability of not slipping, which is \( 1 - p_s \). Given that no slipping occurs, the examinee first chooses a particular strategy \( S_{li} \). This is represented in the model by the first term in the product within the summation, the probability that an examinee with skill mastery vector \( \psi_j \) uses the \( l^{th} \) strategy for item \( i \), namely \( P(S_{li} | \psi_j) \). Then, given that particular strategy choice, the examinee must successfully execute the skills for the item that the strategy calls for. This is represented in the model by the probability that an examinee with skill mastery vector \( \psi_j \) uses the \( l^{th} \) strategy for item \( i \), namely \( P(C_{li} | S_{li}, \psi_j) \). Finally, given a particular strategy selection and given successful execution of the skills, the examinee must still execute whatever other skills are necessary to solve the item that may have been left out of \( Q \). This is represented by the last probability in the product within the summation, the probability of a correct response, given skills mastery levels, strategy selection, and successful execution of the skills, namely \( P(X_{ij} = 1 | S_{li}, C_{li}, \psi_j) \).

4.1.3 Other general diagnostic model frameworks.

GDM. Von Davier’s General Diagnostic Model (GDM) (von Davier, DiBello, & Yamamoto, in press; von Davier’s General Diagnostic Model (GDM) (von Davier, DiBello, & Yamamoto, in press; von
Davier, 2005; von Davier & Yamamoto, 2004) is defined in a very general manner. Part of the general form of the model is provided by a link function. By setting this link function in various ways, a large class of models can be generated as special cases of the GDM, including uni- and multidimensional Rasch models, 2PL model, the generalized partial credit model, the FACETS model (Linacre, 1989), and Maris’ (1999) compensatory MCLCM, as well as more general partial non-compensatory models, including conjunctive diagnostic models. Analyses of real and simulated data have been published on the compensatory discrete MIRT version of the models. The estimation software is EM-algorithm based MML and has demonstrated good computational efficiency. The reader is referred to the references for further detail.

**MRCMLM.** The random coefficient multinomial logit model (RCMLM) was proposed by Adams and Wilson (1996) and extended to a multidimensional model (MRCMLM) (Adams, Wilson & Wang, 1997). The functional form of this model is more general than our general model here. It is quite flexible and generates a wide range of unidimensional and multidimensional models, including unidimensional and multidimensional Rasch models, Fischer’s LLTM (1973, 1983), Master’s (1982) partial credit model, the FACETS model (Linacre, 1989), Wilson’s ordered partition model (Wilson, 1992), Whitely’s MLTM (1980), Andersen’s (1985) multidimensional Rasch model for repeated testing, and Embretson’s (1991) multidimensional model for learning and change. In addition to a design matrix similar to the $Q$ matrix that specifies the contribution of each skill to item difficulty, there is a second design matrix called a scoring matrix. The scoring matrix extends the flexibility of the model and provides the capability for handling testlets (called item bundles in Wilson & Adams, 1995) and even allows a conjunctive version. Examples are provided in which the $Q$ matrix is not simple structure – i.e. some items require more than one skill. The person population parameters are assumed to be random variables and used mainly for the purpose of parameter estimation. They are decomposed into a linear combination of multiple parameters. The general model allows polytomous items, linear constraints on item parameters and flexible scoring rules. Estimation methods attempted include MML estimation that was successful up to three or four dimensions, and CML estimation in the case of Rasch versions of the model. The reader is referred to the references for further detail (Adams, Wilson, Wang, 1997; Wilson & Adams, 1995; Adams & Wilson, 1996).

The next section will now review each of the specific models from Table 2, describing how each represents the interaction of the knowledge structure and item structure. The following two sections after that will separately review the knowledge structure and item structure, respectively, in general terms, summarizing key points from the detailed reviews of the individual models. The knowledge structure review will summarize the variety of ways that skill mastery level $\psi_j$ has been seen to be modeled in Section 4.2; and the item structure review will summarize the variety of approaches seen in modeling $P(C_{li}|S_{li}, \psi_j)$, $P(X_{ij} = 1|S_{li}, C_{li}, \psi_j)$, and $P(S_{li}|\psi_j)$. Thus, Section 4.4 on item structure will be divided into three sub-
sections — one describing item structures for modeling performance on the cognitive components in $Q$ ($P(C_{li}|S_{li}, \psi_j)$), another describing item structures for modeling performance on abilities missing from $Q$ ($P(X_{ij} = 1|S_{li}, C_{li}, \psi_j)$), and another describing item structures for modeling strategy selection ($P(S_{li}|\psi_j)$).

### 4.2 Specific models for item-knowledge interaction.

This detailed discussion of specific models will generally progress from the simplest models to the more complex models; however, there is no strict ordering in terms of complexity for some of the models in the middle of the spectrum. For convenience, we will mostly use the acronyms for each model without naming it. The reader is referred to the table for name-acronym pairs. The Item-Knowledge interaction modeling of the cognitive diagnosis models will now be reviewed for each model, explicitly connecting the structure of each model to the different components of the general model. It is in this section that the various modeling approaches of Table 2 are made explicit.

#### 4.2.1 Unidimensional logistic models: 1PL, 2PL, & 3PL (Rasch, 1961; Birnbaum, 1968); and LLTM (Fischer, 1973, 1983).

The IRF for the 3PL model is as follows:

$$P(X_{ij} = 1|\theta_j) = c_i + \frac{1 - c_i}{1 + exp[-1.7a_i(\theta_j - b_i)]}$$

where

- $\theta_j$ = level of proficiency of examinee $j$ on the presumed unidimensional construct that the item is measuring.
- $a_i$ = the discrimination parameter for item $i$ and is proportional to the maximum slope of the IRF. It is “the degree to which item response varies with ability level” (Lord, 1980).
- $c_i$ = the so-called “guessing parameter,” which is the probability of a correct response for an examinee of very low proficiency. The lower asymptote of the IRF.
- $b_i$ = the difficulty level of item $i$ which is the level of proficiency at which an examinee has a probability of a correct response that is halfway between 1.0 and $c_i$.

When $c_i = 0$, the 2PL model results. When $a_i = 1$ and $c_i = 0$, the 1PL model results. Since the 1PL, 2PL, and 3PL models are unidimensional, they do not have the capability to be directly connected to a set of multiple skills. So, these models are not generally regarded as directly applicable to modeling multiple skills in item performance. Usually two or more skills are conjectured to have an effect on many, if not most, of the items and these skills are usually regarded as distinct dimensions that should appear in the model. (Multidimensional IRT models are discussed below.)

In terms of the general model (Equation 1), there is only one possible strategy and the number of skills for any item is one, and indeed the same skill for all items. Successful performance on the skill is the same
as successful performance on the item, i.e., the modeled skill is assumed to capture all the skills involved in solving the item (completeness of $Q$). Lack of perfect positivity in the skill is dealt with by modeling mastery level as a continuous variable and employing a logistic IRF to determine probability of a correct response conditioned on mastery level. The probability of a random slip is assumed equal to zero. Thus,

$$p_s = 0.$$  
$$\psi_j = \theta_j.$$  
$$\nu_i = 1.$$  
$$P(S_{1i}|\psi_j) = \begin{cases} 1, & \text{if } C_{1i} \text{ and } S_{1i}, \\ 0, & \text{otherwise} \end{cases}.$$  
$$P(C_{1i}|S_{1i}, \psi_j) = \begin{cases} 1, & \text{if } X_{ij} = 1, \\ 0, & \text{otherwise} \end{cases}.$$  

We include here a brief description of LLTM (Fischer, 1973, 1983) even though it does not model the ability variable as a vector of multiple cognitive skills. As conceived by Fischer, the LLTM does not provide, and was not intended to provide, the capability for cognitive skills diagnosis of individual students. We nonetheless include LLTM in this discussion because, in our interpretation, the analysis of item difficulty into factors of difficulty, which can represent “cognitive operations” as noted by Fischer, was a key step historically in the evolution of cognitive diagnostic models from unidimensional IRT models. LLTM includes a $Q$ matrix specifying from a given list of types of “cognitive operations” for each item, the number of each such operations required by the item. Consequently, even though LLTM ability is unidimensional, we see LLTM’s use of a $Q$ matrix as an historically pivotal development that lies mid-way between unidimensional models such as the Rasch model and fully diagnostic models such as those on which we focus below.

The IRF for this model was described by Fischer as a restricted form of the Rasch or 1PL, in the sense that the linear expression for item difficulty in terms of cognitive operations can be thought of as a constraint on item difficulty. The 1PL modeling framework was extended to include “cognitive operations” within the model. These cognitive operations were viewed by Fischer as attributes of items, and not as skills of examinees. Once cognitive determinants of item difficulty have been identified, represented by the LLTM $Q$ matrix, a natural next step is to partition overall ability measured by the test into factors of ability, one for each cognitive determinant of item difficulty. As was done by Tatsuoka, each cognitive item difficulty factor was mapped in a one-to-one fashion to the student’s ability to perform on items that involve that factor. Even though LLTM does not directly diagnose examinee mastery on each skill, it is included in our discussion here because it was the first to employ a $Q$ matrix representation and was, thus, an important stepping stone toward the more complex models and, as such, provides important insights in terms of understanding the more complex models.

Specifically, the LLTM IRF has the same basic form as that for the 1PL model with the only item parameter being $b_i$, but in LLTM the $b_i$ parameter is further explicated as:

$$b_i = \sum_{k=1}^{K} q_{ik} \eta_k + c$$

where
\[ q_{ik} = \text{the frequency with which skill } k \text{ is a factor in the solution of item } i, \]
\[ \eta_k = \text{the level of difficulty attributed to skill } k, \text{ and} \]
\[ c = \text{arbitrary constant that fixes the origin of the scale.} \]

The \( q_{ik} \) variable is almost always dichotomous, indicating whether or not the skill influences item difficulty.

In terms of the general model, there is only one possible strategy. Even though the model allows for any number of skills per item, the performance on the separate skills is not modeled. By using a single latent trait, the model effectively employs a projection of all the skills mastery levels onto a unidimensional representation that can be thought of as some implicit combination of the skills. Thus, \( \theta \) can be thought of as a weighted average of the skills mastery levels of an examinee. With such an interpretation the model can be seen to be a special case of a compensatory model. In Section 4.3 there is a discussion of when this assumption is appropriate. Also note that in the case of \( q_{ik} \) being dichotomous, all items that measure the same set of skills are assumed to have the same difficulty level. Successful performance on the skills is modeled as being the same as successful performance on the item, i.e., \( Q \) is considered to be complete. Lack of perfect positivity is dealt with in the same way as with the logistic models above: overall skill mastery level is modeled as continuous and the logistic function determines a probability of a correct response for a given level of overall skill mastery level.

4.2.2 Compensatory multidimensional IRT model: MIRT-C (Reckase & McKinley, 1991).

The IRF for the MIRT-C model can be written as follows:

\[
P(X_{ij} = 1|\theta_j) = c_i + (1 - c_i)[1 + e^{-1.7\|a_i\|^2/(\sum_{k=1}^{K} a_{ik}^2 - b_i)}]^{-1}
\]

where

\( \theta_j = \text{mastery level vector for examinee } j, \text{ indicating level of mastery of examinee } j \text{ on each of the skills required by the item}, \)

\( a_i = \text{vector of discrimination parameters, } (a_{ik}, k = 1, \ldots, K), \text{ for item } i \text{ indicating the amount of discrimination that item } i \text{ has for each of the skills that the item is sensitive to,} \)

\( c_i = \text{the so-called “guessing parameter,” which is the probability of a correct response for an examinee who has very low level of mastery on all the skills required by the item,} \)

\( b_i = \text{the difficulty parameter for item } i, \text{ and} \)

\( \|a_i\| = (\sum_{k=1}^{K} a_{ik}^2)^{1/2}. \)

This model was first introduced by Reckase and McKinley (1991) and has been popularized for use in cognitive diagnosis by Adams et al. (1997), who employ a Rasch-model based version in which all the \( a_i \) parameters are set equal to 1, \( c_i \) is set to 0, and a \( Q \) matrix determines which skills correspond to which items. The compensatory nature of the item can be seen more clearly if we look at the IRF for the case of two dimensions with \( c_1 = 0, b_1 = 0, \) and \( a_1 = a_2 = 1. \) The IRF for this special case is:
\[ P(X_{ij} = 1|\theta_1, \theta_2) = \frac{1}{1 + e^{-1.7(\theta_1 + \theta_2)}} \]

If we let \( \theta_1 \) take on a very low value, say \( \theta_1 = -2 \), and let \( \theta_2 \) compensate for this low value of \( \theta_1 \) by taking on a very high value, say \( \theta_2 = +2 \), we get \( P(X_{ij} = 1|\theta_1, \theta_2) = 0.5 \), the same value we would get if we had \( \theta_1 = \theta_2 = 0 \). Clearly in this model a high level of mastery on one skill can compensate for a low level on another to give the effect of a medium level on both. For cognitive diagnosis, this model is generally an improvement over the unidimensional IRT models in that multidimensionality is allowed. (See Section 4.3 below for discussion of this issue.) However, the compensatory nature of the interaction of the skills does not match cognitive assumptions in many diagnostic settings. In particular, cognitive analysis of item performance often requires successful performance (i.e., high level of mastery) on all the required skills for an item in order to have success on the item as a whole. In these circumstances, at least from a cognitive perspective, a conjunctive model appears to be more appropriate for the cognitive assumptions.

In terms of the general model, there is only one possible strategy. The model allows for a multitude of skills to influence an item through a multidimensional examinee skill mastery parameter. Successful performance on the skills is modeled as being the same as successful performance on the item, i.e., \( Q \) is assumed to be complete. Lack of perfect positivity in the cognitive components is dealt with by modeling ability as continuous variables and employing a logistic IRF. The probability of a random slip is assumed equal to zero. Thus,

\[ p_s = 0. \]
\[ \psi_j = \bar{\theta}_j. \]
\[ \nu_i = 1. \]
\[ P(S_{ij} | \psi_j) = 1. \]
\[ P(C_{ij} | S_{ij}, \psi_j) = \text{the MIRT-C IRF}. \]
\[ P(X_{ij} = 1 | C_{ij}, S_{ij}, \psi_j) = 1. \]

### 4.2.3 Non-compensatory MIRT model: MIRT-NC (Sympson, 1978).

The item response function (IRF) for the MIRT-NC model is as follows:

\[ P(X_{ij} = 1|\theta_j) = \prod_{k=1}^{K} \frac{1}{1 + e^{-1.7a_{ik}(\theta_{jk}-b_{ik})}} \]

where

\( \theta_{jk} = \) level of mastery of examinee \( j \) on skill \( k \),
\( a_{ik} = \) discrimination parameter of item \( i \) on skill \( k \), and
\( b_{ik} = \) the difficulty level of item \( i \) on skill \( k \).

The knowledge structure of this model is multidimensional, as is generally desired; and the interaction of the skills with item performance is non-compensatory, in the sense that its multiplicative form results in
a low probability of a correct response on the item if an examinee has very low proficiency on any one skill, regardless of how high a proficiency the examinee has on any other skill. More precisely, the model is not fully compensatory – we use the term “non-compensatory” for convenience. Note that the term conjunctive is also sometimes used instead of non-compensatory. A reformulation of the model helps to clarify the conjunctive/non-compensatory nature. Note that each term of the product is a 2PL IRF. Let

\[ Y_{ikj} = \begin{cases} 
1 & \text{if examinee } j \text{ successfully executes skill } k \text{ on item } i, \\
0 & \text{otherwise}. 
\end{cases} \]

Then,

\[ P(X_{ij} = 1|\theta_j) = \prod_{k=1}^{K} P(Y_{ikj} = 1|\theta_{jk}) \]

A justification for the use of this product is the assumption that the examinee must successfully execute all the required skills and that these operations are independent of each other for a given set of skill levels. By independence, the joint probability of successfully executing all the skills is the product of the conditional probabilities of success for the individual skills.

In terms of the general model, there is only one possible strategy. The model allows for a multitude of skills to influence an item through a multidimensional examinee skill mastery parameter space. Successful performance on the skills is modeled as being the same as successful performance on the item, meaning that \( Q \) is assumed to be complete. Lack of perfect positivity is dealt with by modeling skill mastery as a vector of continuous variables and employing logistic functions for probability of success on executing each skill. The probability of a random slip is assumed equal to zero. Thus,

\[ p_s = 0. \]
\[ \psi_j = \hat{\theta}_j. \]
\[ \nu_i = 1. \]

\[ P(S_{1i}|\psi_j) = 1. \]
\[ P(C_{1i}|S_{1i}, \psi_j) = \text{the MIRT-NC IRF}. \]
\[ P(X_{ij} = 1|C_{1i}, S_{1i}, \psi_j) = 1. \]

### 4.2.4 Multicomponent latent trait model: MLTM (Embretson, 1985, 1997; Whitely, 1980).

This model can be viewed as an extension of MIRT-NC. Like MIRT-NC, it employs unidimensional IRFs to model the probability of successful examinee performance on each skill and uses the product of these IRFs to model the probability of successful examinee performance on all the skills of an item simultaneously. In addition, MLTM models the probability of successful examinee performance on the item as a whole, conditioned on whether or not the examinee successfully executed all the skills. The modeling of examinee performance on the item (given successful execution of the skills) amounts to including the modeling of examinee performance on the non-\( Q \) skills.
First, let $Y_{ikj}$ be defined as above as a dichotomous indicator of the success of examinee $j$ in executing skill $k$ on item $i$. Then, the IRF for MLTM can be written as follows:

$$P(X_{ij} = 1 | \theta_j) = a \prod_{k=1}^{K} P(Y_{ikj} = 1 | \theta_{jk}) + g[1 - \prod_{k=1}^{K} P(Y_{ikj} = 1 | \theta_{jk})]$$

where

- $\theta_{jk}$ = level of mastery of examinee $j$ on skill $k$,
- $a = P(X_{ij} = 1 | \prod_{k=1}^{K} Y_{ikj} = 1)$,
- $g = P(X_{ij} = 1 | \prod_{k=1}^{K} Y_{ikj} = 0)$,
- $P(Y_{ikj} = 1 | \theta_{jk}) = \frac{1}{1 + e^{-\psi_{jk} - b_{ik}}}$ (a one-parameter logistic function), and
- $b_{ik}$ = the difficulty level of item $i$ on skill $k$.

In terms of the general model, this model can be viewed as having two strategies with each strategy linked to examinee performance on the skills. The first strategy is that the examinee responds to the item by trying to successfully execute the required skills; and the second is that the examinee, having unsuccessfully executed one of the skills, “guesses.” The probability of the first strategy is defined to be the probability of successfully executing all the required skills for the item. The probability an examinee applies the second strategy (guessing) is simply the probability that an examinee does not successfully execute the required skills for the item.

Successful performance on the skills is not modeled as being the same as successful performance on the item, which essentially means that $Q$ is allowed to be not complete. Only when $a = 1$ is $Q$ modeled as complete for the item. If $a = 1$ and $g = 0$ this model reduces to MIRT-NC. Lack of perfect positivity in the skills is dealt with by modeling skill mastery as a vector of continuous variables and employing logistic functions to determine success probability on each skill. The probability of a random slip is assumed equal to zero. Thus, $p_s = 0$, $\psi_j = \theta_j$, $\nu_i = 2$,

$$P(S_{1i} | \psi_j) = \prod_{k=1}^{K} P(Y_{ikj} = 1 | \theta_{jk})$$
$$P(S_{2i} | \psi_j) = 1 - \prod_{k=1}^{K} P(Y_{ikj} = 1 | \theta_{jk})$$
$$P(C_{1i} | S_{1i}, \psi_j) = 1^2$$
$$P(C_{2i} | S_{2i}, \psi_j) = g.$$ (The skill in this case, guessing, is not very cognitive.),

$$P(X_{ij} = 1 | C_{1i}, S_{1i}, \psi_j) = a,$$ and

$$P(X_{ij} = 1 | C_{2i}, S_{2i}, \psi_j) = 1.$$ 

MLTM also includes a version that allows for multiple cognitive strategies, which we now describe. In this version of the model, the strategies have an order of application such that the probability that each strategy is applied is equated to the probability that the previous strategies in the order were not successfully applied.

---

2This probability is determined given $S_{1i}$ whose probability is equal to $P(C_{1i} | \psi_j)$. That is why this is 1.
applied. Additionally all examinees are assumed to use the same order of strategy application. Specifying the order of strategy execution and the skills for each strategy is crucial to the model. The skill mastery vector, $\theta_j$, now includes skills for all the strategies. Note that this model for multiple cognitive strategies fits the two-strategy MLTM case above where the probability of applying the second strategy, guessing, was equal to the probability of non-success on the first strategy, i.e., the probability of non-success in the execution of the skills for the single-strategy MLTM case.

Let

- $S_{li} = \text{examinee applies strategy } l \text{ on item } i$.
- $C_{li} = \{\text{Successful execution of the skills for strategy } l \text{ on item } i\}$.
- $K_l = \text{the set of skills for strategy } l$.

Then, the probability that strategy $L$ is applied on item $i$ is given by:

$$P(S_{li}|\theta_j) = \prod_{l=1}^{L-1} [1 - P(C_{li}|S_{li}, \theta_j)]$$

where

$$P(C_{li}|S_{li}, \theta_j) = \prod_{k \in K_l} P(Y_{ikj} = 1|\theta_{jk})$$

As in the case of a single cognitive strategy, MLTM in the case of multiple cognitive strategies does not model successful performance on the skills for a given strategy as being exactly the same as successful performance on the item. That is, $Q$ is allowed to be incomplete for each strategy. The probability of successful performance on item $i$ for strategy $l$ is given by:

$$P(X_{ij} = 1|\psi_j) = P(S_{li}|\psi_j) \cdot P(C_{li}|S_{li}, \psi_j) \cdot a_l$$

where

- $a_l = \text{the probability of successful performance on the item given successful performance on the cognitive components of strategy } l$. This probability is a constant over all examinees and items but may vary with strategy.

In terms of the general model:

- $p_s = 0$.
- $\psi_j = \theta_j$.
- $\nu_i = \text{the number of strategies for item } i$.

$$P(S_{li}|\psi_j) = \prod_{l=1}^{L-1} [1 - P(C_{li}|S_{li}, \psi_j)].$$

$$P(C_{li}|S_{li}, \psi_j) = \prod_{k \in K_l} P(Y_{ikj} = 1|\theta_{jk}).$$

$$P(X_{ij} = 1|C_{li}, S_{li}, \psi_j) = a_l.$$

### 4.2.5 General component latent trait model: GLTM (Embretson, 1985, 1997).

This model is a mixture of the MLTM for a single cognitive strategy and the LLTM.

As with MLTM, $Y_{ikj}$ is defined as before. Then the item response function (IRF), as with MLTM, is written as follows:
\[ P(X_{ij} = 1|\theta_j) = a \prod_{k=1}^{K} P(Y_{ikj} = 1|\theta_{jk}) + g[1 - \prod_{k=1}^{K} P(Y_{ikj} = 1|\theta_{jk})] \]

where, as with MLTM,

\( \theta_{jk} \) = level of mastery of examinee \( j \) on skill \( k \),

\( a = P(X_{ij} = 1|\prod_{k=1}^{K} Y_{ikj} = 1) \),

\( g = P(X_{ij} = 1|\prod_{k=1}^{K} Y_{ikj} \neq 1) \)

\[ P(Y_{ikj} = 1|\theta_{jk}) = \frac{1}{1+e^{-\theta_{jk}-b_{ik}}} \], and

\( b_{ik} \) = the difficulty level of item \( i \) on cognitive component \( k \).

But now, as with LLTM, the \( b_{ik} \) difficulty parameter is broken down as follows:

\[ b_{ik} = \sum_{f=1}^{q} \eta_{fk} q_{ikf} + \beta_k \]

where

\( q_{ikf} \) = complexity (usually 1 or 0) of difficulty factor \( f \) in skill \( k \) on item \( i \),

\( \eta_{fk} \) = weight of difficulty factor \( f \) in overall difficulty of skill \( k \), and

\( \beta_k \) = normalization constant.

In LLTM each \( \eta_{fk} \) stands for the effect on item difficulty resulting from the item response requiring the application of required skill \( k \), but in MLTM and GLTM the effect on the item response of the required skills is more directly represented through the \( \theta_{jk} \) examinee parameters. In GLTM, each \( \eta_{fk} \) is interpreted as a difficulty factor that comes from breaking down the \( k^{th} \) skill into the stimuli that are its building blocks.

In terms of the general model, GLTM has the exact same representation as the single-strategy form of MLTM (except for the decomposition of the difficulty parameter described above). Therefore, the comparison with the general model will not be repeated here for GLTM.

### 4.2.6 Restricted latent class model: RLCM (Haertel, 1984, 1990).

RLCM is termed a restricted latent class model because the number and types of deterministic latent response vectors allowed in the model are restricted by \( Q \).

With this model, unlike the models discussed above, the level of mastery of an examinee is not represented as a continuous variable on the dimensions defined by the skills. Instead, examinee ability is modeled by a \( K \)-dimensional vector \( \alpha \) of 0s and 1s. This vector indicates for each skill whether an examinee is a master (1) or non-master (0) of the skill. RLCM takes into account lack of perfect positivity by introducing:

\( \pi_i \) = Probability an examinee get item \( i \) right, given the examinee has mastered all the skills for item \( i \).

\( r_i \) = Probability an examinee gets item \( i \) right, given that the examinee has not mastered at least one of the skills for item \( i \).

The IRF for RLCM is then defined by:

\[ P(X_{ij} = 1|\alpha_j) = \pi_i \prod_{k=1}^{K} \alpha_{jk}^{\eta_{jk}} r_i^{-1} \prod_{k=1}^{K} \alpha_{jk}^{\eta_{jk}} \]
Cognitively the model simply says that if the examinee has mastered all the skills specified as required for an item, the examinee has one (likely high) probability of getting the item right. Whereas, if the examinee has not mastered all the skills of an item, the examinee has another (likely small) probability of getting the item right. Note that with RLCM if an examinee has not mastered one or more required skills for an item, it does not matter how many nor which particular ones have not been mastered. This is an example of the deliberate introduction of modeling parsimony to improve statistical tractability. This type of skill interaction is often referred to as “conjunctive” (analogous to non-compensatory) because the model intends for a high probability of correct response to require mastery of all the skills required for the item.

Historically, RLCM can be viewed as an extension of a model proposed by Macready and Dayton (1977) that used a similar IRF but assumed all items are measuring the same skills and, thus, classifies examinees into one of two categories, either having mastered all the skills or having non-mastered at least one. The RLCM model has been more recently referred to as “DINA” (Junker & Sijtsma, 2001), which stands for “Deterministic-Input, Noisy-And.” The acronym highlights two features of the model: the $\alpha$ vector by itself that can be used to deterministically determine which items an examinee will get right and wrong (see description of deterministic model in Section 4.1.1), but also the fact the item response is probabilistic and is intended to require mastery of all the skills (the “And” part) in order to result in a high probability of a correct response. For a more detailed explanation, readers should consult Junker and Sijtsma (2001).

In terms of the general model, there is only one possible strategy. The model allows for a multitude of skills to influence an item. Successful performance on the skills is modeled as being the same as successful performance on the item, which means that $Q$ is assumed to be complete. Lack of perfect positivity in the skills is dealt with by allowing $\pi_i$ to be less than 1 and allowing $r_i$ to be greater than 0. The probability of a random slip is assumed equal to zero. Thus,

\begin{align*}
p_s &= 0. \\
\psi_j &= \alpha_j = (\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{Kj}). \\
\nu_i &= 1. \end{align*}

\begin{align*}
P(S_{1i} | \psi_j) &= 1. \\
P(C_{1i} | S_{1i}, \psi_j) = \text{the RLCM IRF.} \\
P(X_{ij} = 1 | C_{1i}, S_{1i}, \psi_j) &= 1. \end{align*}

### 4.2.7 HYBRID model (Gitomer & Yamamoto, 1991).

This model, from one perspective, is very similar to RLCM, indeed RLCM is included within HYBRID. In addition to the latent classes postulated by RLCM, HYBRID adds one more class of examinees whose item responses follow a unidimensional IRT model.

Thus, based on an examinee’s observed response vector, the examinee is either classified into one of the $\alpha$ latent classes or is simply assigned an estimate of unidimensional ability from a unidimensional IRT model (such as 1PL, 2PL, or 3PL). The IRF is given by:
\[ P(X_{ij} = 1 | \text{examinee } j) = \begin{cases} 
\text{RLCM IRF} & \text{if examinee } j \text{ is estimated to belong to a latent class.} \\
\psi_i(j) & \text{if examinee } j \text{ can not be classified into any latent class.} 
\end{cases} \]

where

\[ \theta_j = \text{unidimensional overall mastery level of an examinee who cannot be classified into a latent class.} \]

\[ P_i(\theta_j) = \text{either the 1PL, 2PL, or 3PL unidimensional IRF.} \]

In terms of the general model, this model can be viewed as having two strategies, even though the originators of the model probably did not intend for their model to be viewed this way. Examinees who use the first strategy are sensitive to the varied skills that have been specified as being required for the items, i.e., they are using the Q-based strategy. For these examinees, skill mastery is modeled as membership in a latent class. Examinees who use the second strategy are not using the Q-based skills and are modeled as though they are only sensitive to a unidimensional combination of all the skills, as though the items were measuring only a single skill. For these examinees, skill mastery is modeled as a continuous unidimensional latent trait. The probability of the first strategy is simply the probability that an examinee is in one of the latent classes. The probability of the second strategy is simply \( 1 - P(\text{first strategy}) \). HYBRID assumes that the probability a randomly chosen examinee uses a particular strategy does not vary across items and that an examinee uses the same strategy for all items on the test. This latter assumption concerns the test as a whole and is, thus, not captured in our general model which is for a single item. The comparison with the general model is the same as for RLCM, except for the addition of the second strategy. Thus,

\[ p_s = 0. \]

\[ P(S_1 | \psi_j) = P(S_1) = \text{Probability an examinee is classified into one of the latent classes.} \]

\[ \psi_j = \alpha_j \text{ or } \theta_j. \]

\[ P(S_2 | \psi_j) = 1 - P(S_1). \]

\[ \nu_i = 2. \]

\[ P(C_1 | S_1, \psi_j) = \text{RLCM IRF.} \]

\[ P(C_2 | S_2, \psi_j) = P_i(\theta_j) \text{ (1PL, 2PL, or 3PL IRF).} \]

\[ P(X_{ij} = 1 | C_1, S_1, \psi_j) = 1. \]

\[ P(X_{ij} = 1 | C_2, S_2, \psi_j) = 1. \]

### 4.2.8 Deterministic-Input, Noisy-Or: DINO (Templin & Henson, in press).

This model is very similar to RLCM (also known as DINA), except instead of being conjunctive (the “And” in DINA), it is disjunctive (the “Or” in DINO). By disjunctive we mean that this model intends for a high probability of a correct response on an item to occur whenever any of the skills are mastered (the “Or”
part). Thus, the item and ability parameters are nearly the same as for RLCM, but the IRF takes on a disjunctive form.

Examinee ability is modeled the same as with RLCM using a $K$-dimensional vector $\alpha$ of 0s and 1s. DINO takes into account the lack of perfect positivity by introducing:

$\pi_i = \text{Probability examinees get item } i \text{ right, given they have mastered at least one of the skills for item } i.$

$r_i = \text{Probability examinees get item } i \text{ right, given they have not mastered any of the skills for item } i.$

The IRF for DINO is then defined by:

$$P(X_{ij} = 1 | \alpha_j) = \pi_i \left[ 1 - \prod_{k=1}^{K} (1 - \alpha_{jk})^{q_{ik}} \right] r_i \prod_{k=1}^{K} (1 - \alpha_{jk})^{q_{ik}}$$

With DINO, if an examinee has mastered one or more required skills for an item, it does not matter how many have been mastered nor which particular ones have been mastered. This is another example of the deliberate introduction of modeling parsimony to improve statistical tractability. As mentioned above, this type of skill interaction is referred to as “disjunctive” (analogous to compensatory) because the model intends for a high probability of an incorrect response on the item to require non-mastery of all the skills required for the item.

In terms of the general model, the only difference from RLCM is $P(C_{ij} | S_{1i}, \psi_j) = \text{the DINO IRF}$. 

4.2.9 Reparameterized Unified Model: RUM (DiBello, Stout, & Roussos, 1995; Hartz, 2002).

The original Unified Model of DiBello, Stout, and Roussos (1995), although conceptually attractive, had non-identifiable item parameters and, hence, was statistically inestimable without further constraints on the parameters. There was also a version that allowed for multiple strategies, which will not be discussed here. Hartz (2002) parsimoniously reparameterized the single-strategy Unified Model, and this is the model that has been extensively researched and applied, and which will be presented here. Hartz also added a Bayesian hierarchical structure for estimation purposes and called the enhanced model the Fusion Model. Our focus here is on the RUM IRF, although readers interested in the Bayesian component are advised to consult Hartz (2002) or Hartz and Roussos (2005).

Like HYBRID, RUM can be viewed as an extension of RLCM. Note that HYBRID does not change the part of RLCM that determines the classification of the examinees into cognitively based (i.e., $Q$-based) latent classes. Instead, HYBRID merely places hard-to-classify examinees under the umbrella of a unidimensional IRT model, thus moderating the knowledge structure assumptions of RLCM. RUM, on the other hand, faithfully maintains a knowledge structure based on cognitively meaningful latent classes and, instead, brings more elaborated cognitive modeling into the item structure in an attempt to improve the classification process.
The authors of RUM intended for the model to be a unifying bridge between cognitively over-simplified unidimensional IRT models and intractable cognitive-based expert systems with thousands of nodes. In addition, within the context of cognitive diagnosis models alone, RUM can be viewed as bridging the gap between the more cognitively complex MLTM (single strategy) and GLTM and the statistically more tractable but cognitively more constrained RLCM and HYBRID.

Like MLTM and GLTM, RUM models examinee performance on the individual skills. However, RUM replaces the continuous skill level parameters with discrete ones. RUM uses RLCM-like item parameters (focused on dichotomous skill mastery) instead of logistic functions (focused on continuous skill mastery) to model examinee performance on the skills. (Research has been completed with a version of RUM in which skill levels can be any finite number of ordered categories. That work will not be discussed here.)

Thus, in terms of model complexity RUM is effectively in between (MLTM, GLTM) and (RLCM, HYBRID). As described for RLCM, RUM models the ability of examinee \( j \) by a vector \( \alpha_j \) of 0s and 1s, indicating mastery (1) or non-mastery (0) on each of the \( K \) skills in \( Q \). As described for MLTM and MIRT-NC, RUM models the joint probability of successfully executing all the skills required for an item as the product of the success probabilities for the individual skills, an assumption of skill execution independence.

Thus, as in MIRT-NC and MLTM, let

\[
Y_{ikj} = \begin{cases} 
1 & \text{if examinee } j \text{ successfully executes skill } k \text{ on item } i \\
0 & \text{otherwise.} 
\end{cases}
\]

Assuming skill execution independence,

\[
P(\prod_{k=1}^{K} Y_{ikj} = 1 | \alpha_j) = \prod_{k=1}^{K} P(Y_{ikj} = 1 | \alpha_j) \]

In MLTM, positivity is dealt with by modeling \( P(Y_{ikj} = 1 | \alpha_j) \) as a logistic function with structural parameters specific to skill \( k \) on item \( i \). In RLCM, positivity is dealt with by modeling \( P(\prod_{k=1}^{K} Y_{ikj} = 1 | \alpha_j) \) as two constants, one (high) value for when \( \prod_{k=1}^{K} \alpha_{jk}^q = 1 \), and another (low) value for when \( \prod_{k=1}^{K} \alpha_{jk}^q = 0 \).

The original Unified Model maintained structural parameters specific to skill \( k \) on item \( i \), but simplified the model by using RLCM-like constants. Specifically, in the original Unified Model (UM),

\[
P(Y_{ikj} = 1 | \alpha_j) = \pi_{ik}^{\alpha_{ik}} r_{ik}^{(1-\alpha_{ik})} \]

where \( \pi_{ik} = P(Y_{ikj} = 1 | \alpha_{jk} = 1) \) and \( r_{ik} = P(Y_{ikj} = 1 | \alpha_{jk} = 0) \).

Thus, \( P(\prod_{k=1}^{K} Y_{ikj} = 1 | \alpha_j) = \prod_{k=1}^{K} \pi_{ik}^{\alpha_{ijk}} r_{ik}^{(1-\alpha_{ijk})} \).

With UM, if an examinee has not mastered a required skill for an item, the probability of a correct response depends on which skills have not been mastered. The positivity of the non-mastered skill determines how likely the examinee will successfully execute it given non-mastery on it. If all the skills for an item have
sufficiently high positivity, then the probability that an examinee who has not mastered one or more skills will successfully execute all the skills required for an item, will be approximately the same (very close to 0) value no matter how many or which skills the examinee has not mastered. If such a strong assumption is justified then this part of UM can be effectively replaced by the simpler representation used in RLCM.

The above parameterization was not identifiable, as also pointed out by Maris (1999). To ameliorate the non-identifiability, Hartz (2002) reparameterized UM as follows,

\[ P_\pi (\prod_{k=1}^{K} Y_{ikj} = 1 | \alpha_j) = \pi_i^* \prod_{k=1}^{K} r_{ik}^{(1-\alpha_j)\pi_{ik}} \]

where \( \pi_i^* = \prod_{k=1}^{K} \pi_{ik} \) and \( r_{ik}^* = \frac{r_{ik}}{\pi_{ik}} \).

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. For an examinee who has not mastered a required skill, the item response probability will be multiplicatively reduced by an \( r_{ik}^* \) for each non-mastered skill, where \( 0 < r_{ik}^* < 1 \). The more strongly the item depends on mastery of a skill, the lower the item response probability should be for a non-master of the skill, which translates to a lower \( r_{ik}^* \) for that skill on that item. Thus, \( r_{ik}^* \) functions like a reverse indicator of the strength of evidence provided by item \( i \) about master of skill \( k \). The closer \( r_{ik}^* \) is to 0, the more discriminating item \( i \) is for skill \( k \).

Like MLTM (and unlike RLCM), RUM also allows for \( Q \) to be not complete and, thus, does not equate successfully executing the \( Q \)-specified skills with successfully executing the item. RUM includes a model for the ability of the examinee to successfully execute skills that are necessary for getting the item right but that have been left out of the \( Q \)-specified skills for an item. These skills are modeled using a unidimensional IRT model with a continuous latent trait, \( \eta_j \). Thus, the full IRF for RUM can be written as follows,

\[ P(X_{ij} = 1 | \alpha_j, \eta_j) = P(\prod_{k=1}^{K} Y_{ikj} = 1 | \alpha_j) P_{ci}(\eta_j) \]

where \( P_{ci}(\eta_j) = \frac{1}{1+\exp\{-1.7[\eta_j-(c_i)]\}} \), a 1PL model with difficulty parameter \(-c_i\) and ability parameter \( \eta_j \).

The \( \eta_j \) parameter functions as a unidimensional projection or combination of the levels of skill mastery for an examinee on the non-\( Q \) skills. The \( c_i \) parameter is referred to by DiBello et al. and by Hartz as the “completeness parameter” and is interpreted as an indicator of the degree to which \( Q \) is complete in its coverage of the skills. Hartz constrained \( c_i \) to be greater than 0, with larger values indicating greater completeness for \( Q \). When \( c_i \) is 3 or more, this term will be very close to 1 for most values of \( \eta_j \), and when \( c_i \) is near 0, the \( \eta_j \) variation will have its maximum influence on the RUM IRF.

The \( P_{ci}(\eta_j) \) part of the model simply recognizes that it is usually neither desirable nor possible to list in \( Q \) every possible skill that goes into solving an item for a given strategy. Indeed, this part of the model might sometimes be appropriately viewed as representing higher-order skills, such as self-monitoring behavior or the ability to assimilate the separate cognitive components into a coherent whole, which are less amenable to being modeled as discrete cognitive components (as suggested by Samejima, 1995).
If all examinees with the same latent skill class $\alpha$ have the same $\eta$, then the completeness term for incompleteness would be eliminated by simply reducing $\pi_i^*$ by an appropriate multiplicative constant. Also, as described above, if the positivity of all the cognitive components for an item were sufficiently high, then this part of RUM could further reduce to the corresponding part of RLCM.

In terms of the general model,

$$p_s = 0.$$  \hspace{1cm}  

$$\psi_j = (\alpha_j, \eta_j).$$  \hspace{1cm}  

$$\nu_i = 1.$$  \hspace{1cm}  

$$P(S_{1i}|\psi_j) = 1.$$  \hspace{1cm}  

$$P(C_{1i}|S_{1i}, \psi_j) = \pi_i^* \prod_{k=1}^K r_{ik}^{\alpha_{jk} \eta_{jk}}.$$  \hspace{1cm}  

$$P(X_{ij} = 1|C_{1i}, S_{1i}, \psi_j) = P_{c_i}(\eta_j).$$  \hspace{1cm}  

4.2.10 Disjunctive MCLCM (Maris, 1999).

Maris (1999) proposed several Multiple Classification Latent Class Models, including a disjunctive model which we refer to here as MCLCM-D. Let $Y_{ijk}$ be the same as defined above as for MLTM, MIRT-NC, and RUM, all of which assumed skill execution independence such that,

$$P(\prod_{k=1}^K Y_{ikj} = 1|\alpha_j) = \prod_{k=1}^K P(Y_{ikj} = 1|\alpha_j)$$

If $Q$ is complete, then

$$P(X_{ij} = 1|\alpha_j) = \prod_{k=1}^K P(Y_{ikj} = 1|\alpha_j)$$

The RUM IRF takes on this form when $Q$ is complete. Maris (1999) proposed a model of this form using the same parameterization of UM, thus rediscovering the $Q$-complete version of the original Unified Model. He called it a “conjunctive” MCLC model because the conjunction of successful skill execution on all the skills is required for success on the item.

Maris also proposed what he called a “disjunctive” model as follows,

$$P(X_{ij} = 1|\alpha_j) = 1 - \prod_{k=1}^K [1 - P(Y_{ikj} = 1|\alpha_j)]$$

where he again used the same parameterization as UM for $P(Y_{ikj} = 1|\alpha_j = 1)$ and $P(Y_{ikj} = 1|\alpha_j = 0)$, namely, $\pi_{ik} = P(Y_{ikj} = 1|\alpha_{jk} = 1)$ and $r_{ik} = P(Y_{ikj} = 1|\alpha_{jk} = 0)$.

Thus, the MCLCM-D IRF can be written as,

$$P(X_{ij} = 1|\alpha_j) = 1 - \prod_{k=1}^K (1 - \pi_{ik})^{\alpha_{jk} q_{ik}} (1 - r_{ik})^{(1-\alpha_{jk}) q_{ik}}$$

As in UM, these item parameters are not identifiable, but a reparameterization similar to that of Hartz (2002) could be implemented. The interpretation of this disjunctive model is that a high probability of
success at the item level will occur so long as examinee \( j \) has a high probability of successfully executing at least one of the required cognitive components for the item (i.e., the examinee is a master of at least one of the required components). Conversely, a low probability of success will occur only if the examinee is a non-master on all the cognitive components. This disjunctive model might be termed the maximum or extreme compensatory model, as it allows mastery on any one component to compensate for non-mastery on all the others. This model may also be viewed as a more complex version of DINO. In both DINO and MCLCM-D, a high probability of success on an item is intended to result from mastery of any of the skills; but, in MCLCM-D this probability varies depending on which skills have been mastered, whereas in DINO the same probability is the same no matter which skills or how many have been mastered (as long as at least one has been mastered).

In terms of the general model,

\[
\begin{align*}
p_s &= 0. \\
\psi_j &= \alpha_j. \\
\nu_i &= 1.
\end{align*}
\]

\[
P(S_{1i} | \psi_j) = 1. \\
P(C_{1i} | S_{1i}, \psi_j) = \text{MCLCM-D IRF}. \\
P(X_{ij} = 1 | C_{1i}, S_{1i}, \psi_j) = 1.
\]

4.2.11 Compensatory MCLCM (Maris, 1999).

Among several Multiple Classification Latent Class Models proposed by Maris (1999; see also von Davier, DiBello, & Yamamoto, in press; von Davier, 2005; von Davier & Yamamoto, 2004) was a compensatory model in the form of MIRT-C, but using the dichotomous \( \alpha_{jk}, k = 1, \ldots, K \) ability variables instead of the continuous \( \theta_j \) variables of the MIRT-C model.

Thus, the form of the IRF is as follows,

\[
P(X_{ij} = 1 | \alpha_j) = \frac{\exp[\sum_{k=1}^{K} a_{ik} \alpha_{jk} - b_i]}{1 + \exp[\sum_{k=1}^{K} a_{ik} \alpha_{jk} - b_i]},
\]

where \( a_{ik} \) is the amount of increase in probability on the log-odds scale when \( \alpha_{jk} \) goes from 0 to 1, and \( b_i \) is a threshold parameter.

This model has the attractive feature of allowing practitioners to maintain discrete (mastery, non-mastery) skills while implementing cognitive diagnosis in a situation for which a compensatory model is more appropriate than a non-compensatory model. One could imagine using this model in comparison with the complete-Q RUM to see whether a conjunctive or compensatory model is more appropriate for a given situation where the psychological reality seems to call for one over the other. It is important also to note a model sensitivity issue here. Even in cases in which a conjunctive model is better aligned with cognitive theory, a comparison of both conjunctive and compensatory discrete models can indicate if the
model outcomes are sensitive to the model differences. Research is currently being conducted to investigate this issue.

In terms of the general model,

\[ p_s = 0. \]
\[ \psi_j = \alpha_j. \]
\[ \nu_i = 1. \]

\[ P(S_{1i}|\psi_j) = 1. \]
\[ P(C_{1i}|S_{1i}, \psi_j) = \text{MCLCM-C IRF}. \]
\[ P(X_{ij} = 1|C_{1i}, S_{1i}, \psi_j) = 1. \]

### 4.3 Knowledge structure modeling assumptions.

This section of the chapter, in somewhat broad strokes to aid clarity of presentation, reviews and, in particular, compares and contrasts the nature of the knowledge structures for the models detailed Section 4.2. The two main distinctions to be made are: (1) Dimensionality (unidimensional vs. multidimensional) and (2) Skill mastery scale (continuous vs. discrete). Table 3 below summarizes the knowledge structure of the models with respect to these two dimensions.

| Insert Table 3 about here. |

Readers are advised to consult Section 4.2 and Table 2 as the model abbreviations will be used throughout this section and the detailed model descriptions will sometimes be referred to.

#### 4.3.1 Dimensionality.

The knowledge structure is explicitly unidimensional for some models (LLTM, 1PL, 2PL, and 3PL) and is explicitly multidimensional for the others (MIRT-C, MIRT-NC, MCLCM-C, MCLCM-D, MLTM, GLTM, RLCM, DINO, and RUM). One model, HYBRID, is a mix of unidimensional and multidimensional modeling. For HYBRID, most of the examinees can be assigned a multidimensional vector of skill mastery levels, but examinees who do not fit the multidimensional model are assigned a unidimensional overall mastery level estimated from a unidimensional IRT model (1PL, 2PL, or 3PL).

Given that cognitive diagnosis models are meant to be applied in situations where individual items are assumed to be measuring a number of skills, standard unidimensional IRT models (1PL, 2PL, and 3PL) lack any direct link between the individual skills and their unidimensional ability variable. LLTM does include an explicit link between skills and items through the item component of the IRF, even though its ability is unidimensional.
In LLTM the unidimensional ability can be interpreted as some sort of composite of the multiple skills. Thus, according to the model, two examinees who have the same average level of skill mastery (not necessarily the same set of mastered and not-mastered skills) are expected to have the same probability of a correct response on any item. For this assumption to hold at least roughly, LLTM implicitly requires the multiple skills be highly related. This representation could hold true in some situations. For example, consider a classroom of students learning a subject for which all start off with zero knowledge. All students receive the same instruction and thus learn the multiple skills (to the best of their ability) in the same order. Since skills are usually highly related in such a setting, the degree to which a child learns one skill could be the same for all skills. And since all skills are learned in the same order, the unidimensional knowledge structure model would hold approximately true. Thus, unidimensional models can be useful in some cognitive diagnosis settings. However, in situations where students with the same average ability over the multiple skills vary significantly on their levels of mastery on the individual skills, multidimensional cognitive diagnosis models would be necessary in order to effectively assess skill mastery levels. Multidimensional models vary widely, of course. One important distinction is whether the knowledge structure is continuous or discrete, or a mixture of the two. This distinction is discussed next with respect to all the models. Beyond this distinction, models with the same knowledge structure may also differ in their item structures which will be discussed in Section 4.4.

4.3.2 Skill mastery scale.

The second distinction in knowledge structures to be highlighted is that of continuous skill mastery scale vs. discrete skill mastery scale. The models that have unidimensional knowledge structure (LLTM, 1PL, 2PL, and 3PL) all have continuous representations for knowledge via a single real-valued number of a continuous scale, which along with the item parameters completely determines the examinee’s probability of getting the item right.

Of the models that allow the required skills for an item to be multidimensional, MLTM, GLTM, MIRT-C, and MIRT-NC represent knowledge as continuous latent traits with the number of latent traits for an item being the number of skills of the knowledge structure involved in responding to the item.

On the other hand, of the remaining multidimensional models, RLCM, DINO, MCLCM-C, MCLCM-D, and RUM (with Q complete) all use a discrete representation of the knowledge structure, namely a latent class defined by a vector of dichotomous skill mastery parameters, \( \alpha_j = (\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jK})^t \). where, as described earlier in Sections 4.1 and 4.2, \( \alpha_{jk} = 1 \) indicates examinee \( j \) has mastered skill \( k \), and \( \alpha_{jk} = 0 \) indicates non-mastery.

We ignore here the question of whether the models have mathematically equivalent discrete representa-
tions. For example, 1PL is mathematically equivalent to a model for which the ability variable is discrete with $I + 1$ ordered levels, where $I$ is the number of items. For purposes of this chapter, we consider the continuity or discreteness as explicitly defined. In the examples surveyed here, the number of discrete ordered levels for ability variables is two. In published cases using more than two discrete ordered levels, the number of such levels is usually much lower than the number of items.

The remaining models that allow for multidimensional knowledge structure, employ a mixture of continuous and discrete representations, though in very different ways. The HYBRID model uses virtually the same knowledge structure as RLCM except that, in addition to the discrete latent classes, HYBRID also models some examinees to have their knowledge characterized instead by one of the unidimensional IRT models (1PL, 2PL, or 3PL), which, as mentioned above, employ a continuous representation of skill mastery level. The RUM knowledge structure representation is similar to that of RLCM and HYBRID. All three have the same latent class knowledge representation for skill mastery, a vector of dichotomous skill mastery level parameters. The major difference between RUM and the other two models is that RUM not only includes model parameters for mastery of the $Q$-specified skills, but also includes a continuous unidimensional parameter to provide an approximate model for any skills that may be needed to get the item right and which are not included in the $Q$-specified skills. Note that this use of IRT modeling to augment the latent class knowledge structure based on $Q$ is much different from the use of the unidimensional IRT model in HYBRID. In HYBRID the knowledge structure of an examinee is either one of the restricted latent classes or an IRT ability estimate. One or the other case is assumed to hold for all items. By contrast, with RUM all examinees are posited both a latent class $\alpha$ (for their $Q$-specified skills) and a continuous IRT ability estimate (summarizing their non-$Q$ skills). Of all the cognitive diagnosis models, RUM seems to be the only one that attempts to model examinee differences in mastery on skills not included in $Q$. This flexibility may turn out to be of consequence in many modeling situations, though the continuous term as currently parameterized has often been difficult to work with.

A comparison of the above knowledge structure representations with the knowledge structure representation of the purely deterministic cognitive diagnosis model and with the relaxations of knowledge structure described in Section 4.1 can provide some guidance to deciding what type of knowledge structure is preferred. In individual cases, such preference would depend on how different the modeled situation is from the deterministic situation. Specifically, it depends on the positivity of the skills, the completeness of $Q$, and the purpose of the cognitive diagnosis. If $Q$ is approximately complete or if examinees of equal mastery on the $Q$ skills do not vary significantly on their non-$Q$ skills, then a model with no representation of the non-$Q$ skills may perform very well. If the testing situation is truly concerned only with determination of mastery vs. non-mastery (even in the presence of low positivity of the skills) or if the skills have a high
degree of positivity, then the discrete (for example, the mastery/non-mastery dichotomy) representation of mastery on the skills would seem to be preferred since the estimation will be more efficient than for models with continuous mastery variables. Other trade-offs in complexity between knowledge structure and item structure can also be considered as described in Section 4.2.

### 4.4 Item structure modeling assumptions.

In this section the item structure modeling assumptions will be discussed in general terms to demonstrate the broad variety of such assumptions present in cognitive diagnosis models. The models have many differences in the details of the implementations of these structures, and the reader is referred to Section 4.2 for a more detailed description of the assumptions for each model. This section is divided into three subsections: Item structure for $Q$ skills, Item structure for non-$Q$ skills, and Item structure for multiple strategies. All of the models are included in the first subsection, but only three of the models reviewed in this chapter have item structure for non-$Q$ skills and only two of these models have item structure for multiple cognitive strategies. All of the models assume that conditioned on ability (that is, for examinees of the same level of mastery in the modeled skill mastery space), the item responses of the examinees are statistically independent. This is commonly referred to as the assumption of Local Independence or Conditional Independence.

#### 4.4.1 Item structure for $Q$ skills.

The two main distinctions to be made among the item structure representations for the $Q$ skills are: (1) Number of skills (single vs. multiple) and (2) Skill interaction (compensatory/disjunctive vs. non-compensatory/conjunctive). By *compensatory/disjunctive* is meant that no matter how low a level of mastery an examinee has on one skill, it is possible in the mathematical model of the IRF that a high level of mastery on another required skill can compensate (in terms of probability of a correct response) for the low level of mastery on the first skill to any desired degree. Suppose an item depends on only two skills. Then even if an examinee is extremely weak in mastery of one skill, by having high enough mastery of the second skill the examinee can still have a high probability of a correct response. By *non-compensatory/conjunctive* is meant that an examinee must be of sufficiently high ability on all the required skills of an item in order to have a high probability of getting the item right. Weakness on one skill may be partly compensated by strength on another but it cannot be compensated for to an arbitrarily high degree. Partial compensation is allowed in all the non-compensatory models as one of the relaxations (positivity) of the deterministic cognitive diagnosis model described in Section 4.1. Indeed, any item response model that demonstrates monotonicity in each of the ability variables must demonstrate some amount of partial compensation (van der Linden, personal communication).
Table 4 below summarizes the item structure of the models with respect to these two dimensions. Note that when an item structure models only a single skill, then the skill interaction dimension is not applicable.

<table>
<thead>
<tr>
<th>Number of skills</th>
<th>Skill interaction</th>
</tr>
</thead>
</table>
| Cognitive diagnosis models are meant to be applied in situations where individual items are assumed to be measuring a number of skills, the standard unidimensional IRT models (1PL, 2PL, and 3PL) are seen, as discussed above, to not provide direct links between multiple skill demands of items and multiple skill levels of examinees. All the remaining models allow for the influence of multiple skills in their IRFs. In HYBRID, as mentioned in the previous section, most of the examinees are modeled with a vector of dichotomous variables which indicate mastery or non-mastery of the multiple skills modeled by the test items; but a few examinees are modeled with a unidimensional ability. Thus, HYBRID is listed as a model that mixes a single-skill representation with a multiple-skill representation.

**Number of skills.** Given that cognitive diagnosis models are meant to be applied in situations where individual items are assumed to be measuring a number of skills, the standard unidimensional IRT models (1PL, 2PL, and 3PL) are seen, as discussed above, to not provide direct links between multiple skill demands of items and multiple skill levels of examinees. All the remaining models allow for the influence of multiple skills in their IRFs. In HYBRID, as mentioned in the previous section, most of the examinees are modeled with a vector of dichotomous variables which indicate mastery or non-mastery of the multiple skills modeled by the test items; but a few examinees are modeled with a unidimensional ability. Thus, HYBRID is listed as a model that mixes a single-skill representation with a multiple-skill representation.

**Skill interaction.** Type of skill interaction assumed by a model is most important when items require multiple skills, but, as we have seen above, does not strictly require a multidimensional representation of skill mastery. Specifically, LLTM does include the effects of multiple skills in the item structure portion of its IRF, but skill mastery is essentially represented as some sort of unspecified average of mastery over all the skills, a unidimensional representation. Although the representation of multiple skills is not explicit in LLTM, as we discussed earlier, it can be thought of as equivalent to a compensatory skill mastery interaction model. Consequently, we have so labeled it in our corresponding tables. Clearly, LLTM was not intended for diagnosis of examinee mastery on individual skills, and we include it here as an historically important bridge in the evolution from non-diagnostic unidimensional models to multidimensional diagnostic models.

MIRT-C and MCLCM-C each use a weighted sum of mastery level variables on the skills in their IRFs, thus making them also compensatory models. MCLCM-D is referred to by Maris (1999) as a “disjunctive” model because a high probability of a correct response will occur so long as at least one of the required skills has been mastered, which could be viewed as the maximum possible compensation.

All of the remaining models that have item structures that allow for multiple skills assume that the mastery levels on the skills required by the item interact in a non-compensatory or “conjunctive” manner in the IRF.

From a psychological perspective, the interaction of skills in examinee item response behavior may sometimes be better modeled as compensatory and other times better modeled as conjunctive. From the model fit perspective, however, it may be that either type may fit approximately equally well in practice. For example, even though the compensatory models in principle allow compensation to an arbitrarily high degree among the skills, in practice the skills required for a given item may be highly enough correlated so
as to preclude large amounts of compensation for a given examinee. In situations where this is true, a compensatory model might sometimes be favored over an equivalent non-compensatory model if the parameters of the compensatory model are fewer in number or easier to estimate for a particular setting. However, as will be described below, the available non-compensatory models may also contain other psychologically valuable modeling components – such as item structure for non-Q skills (Section 4.4.2) and multiple cognitive strategies (Section 4.4.3) – that are not found in the compensatory models thus making it difficult to find an equivalent compensatory model. Perhaps the highlighting of such modeling gaps will result in new models to fill these gaps as needed.

4.4.2 Item structure for non-Q skills.

Only three models have item structure for non-Q skills: MLTM, GLTM, and RUM. As mentioned in Section 4.3, among all the models only RUM explicitly models differences in examinee mastery on non-Q skills. MLTM and GLTM essentially assume that all examinees have the same level of mastery on the non-Q skills. MLTM and GLTM have exactly the same item structure for modeling performance on the non-Q skills. They both use a constant probability that is independent of both examinees and items (indeed, this constant could be interpreted as the probability of “not slipping,” although that was not the intended interpretation in Embretson, 1985, 1997). RUM takes a different approach, using a 1PL model with a completeness item parameter that ranges from 0 to 3 (the negative of the 1PL difficulty parameter). Thus, RUM allows different degrees of Q completeness from item to item, and uses a unidimensional parameter to approximately model the non-Q skills.

As discussed in the Knowledge structure section (Section 4.3 above), the modeling of the non-Q skills, as in RUM, is not always necessary. Even though completeness and positivity are theoretically separate issues, the two can become confounded. For example, as discussed earlier, if examinees of the same level of mastery on the skills are also of the same level of mastery on the non-Q skills, performance on the non-Q skills can not be distinguished from performance on the Q skills, and Q incompleteness is totally accounted for within the positivity terms. Of course there is also the trivial case where modeling performance on non-Q skills is not necessary because Q is complete, or very nearly so. Though a high degree of completeness is unlikely on theoretical grounds, we believe that when the number of skills is at least moderately large, the skills in Q may “explain” (much like in a principal components analysis) most of the observed systematic variation. In such cases, we can eliminate the completeness term from the model with little penalty to effective skills mastery assessment.
4.4.3 Multiple cognitive strategies.

The term *strategy* as we use it here refers to a hypothesis about how examinees respond to one or more test items. A strategy is made concrete by the specification of the skills that are required for each item, i.e., the $Q$ matrix. Only MLTM allows for alternate strategies in the item response behavior of examinees. Even though HYBRID can also be viewed as modeling two specific strategies, only one of the strategies is cognitively based. In MLTM the alternative strategies are explicitly modeled in the skills. In HYBRID, even though only a single set of skills is modeled, the alternative unidimensional model within HYBRID can be viewed as modeling a single skill for all the items. As mentioned above, this single skill could be viewed as a unidimensional composite of the multiple skills with the composite being the same for all the items.

Although the modeling of multiple strategies in parametric IRT skills diagnosis has been implemented to only a limited degree, the integration of multiple strategies into unidimensional and latent class IRT models has been studied more extensively, including, for example, Mislevy and Verhelst (1990), Kelderman and Macready (1990), and Rost (1990). Thus, the modeling of multiple strategies in parametric IRT skills diagnosis is an area where advancements in the near future can be expected.

4.5 Summary table of the models.

The above review of cognitive diagnosis models can be summarized with a table highlighting important differentiating characteristics that the above review revealed. Four features seem especially relevant in assessing the degree of complexity of a model:

1. **Multidimensionality.** The skills can be modeled as separate dimensions and more than one can be allowed to occur in a given item. We note here the term multidimensional is used to refer to the indicated dimensions. Even if a four-skill model were shown to have lower dimensionality for a particular dataset, the model itself mediates the true statistical dimensionality for this dataset and the indicated dimensionality that may be based on cognitive or pedagogical theory.

2. **Conjunctive vs. compensatory item response modeling.** The psychological nature of how the cognitive components interact in an item should help determine which model is more appropriate. Conjunctive models are better aligned with cognitive theory in some cases in which it is strongly believed that high levels of ability on all required skills for an item are required for high likelihood of correct response to that item.

3. **Incompleteness of $Q$.** The degree to which $Q$ may be considered complete or nearly complete should be considered in choosing a model. When the number of specified skills is large, $Q$ will usually be
approximately complete, by analogy with principal components analysis. But when the specified number of skills is small, this part of the modeling process could be important.

(4.) **Multiple cognitive strategies.** There will often exist more than one way to solve an item correctly and this can result in more than one set of skills available for success in solving the item.

A fifth important feature is that a model allow for the influence of multiple skills on examinee item performance. As noted above, all the models, except 1PL, 2PL, and 3PL, allow for such an influence, usually through multidimensional modeling of examinee proficiency. Even though this is an important characteristic of a cognitive diagnosis model, since it is shared by all the models meant for cognitive diagnosis, it is not included as a separate entry in this list. Table 5 below summarizes the cognitive diagnosis models reviewed in this chapter by noting whether or not (Y for Yes, N for No) each model contains the above characteristics. The first four models in the table are the unidimensional models; the next four models are the ones with multidimensional continuous skill mastery representations; and the last five models are the models with multidimensional discrete skill mastery representations. Within each of these sets the models are placed in order according to increasing cognitive complexity as determined by the headings of the table, though the ordering is only partial in some cases since there are ties.

Finally, it is important to point out that models having a time component were purposely left out of this review to maintain a reasonable scope. These models include the multidimensional Rasch model for learning and change by Embretson (1991), the LLTM with relaxed assumptions by Fischer (1989), the tutoring systems of Anderson, Corbett, Koedinger, and Pelletier (1995) and of VanLehn, Niu, Siler, and Gertner (1998), and the stochastic learning theory model for a knowledge space by Falmagne (1989). A separate summary of how learning over time is modeled should be addressed in the future.

5 **Summary and Conclusions: Issues and Future Directions**

The purpose of this chapter is to provide for psychometricians and statisticians a broad survey of the basic ideas, models, and statistical tools that constitute the emerging field of skills diagnostic assessment, to provide a focused survey of the field’s rapidly growing literature, and to lay out for researchers and practitioners the steps needed to effectively conduct such assessments. A secondary goal is to illuminate some of the unresolved issues and controversies of the skills diagnostic assessment field, as a guide for future research. This statistical subfield promises exciting research opportunities and promises to play a vital
societal role as assessments developed specifically to support teaching and learning become increasingly important in schools and other instructional settings.

Skills diagnostic assessment research to date has made considerable and important progress. In particular, a rich, varied, and useful collection of parametrically complex and potentially useful stochastic models have been developed, which were carefully and systematically surveyed in the second portion of the chapter. To situate the model survey within a proper context of assessment design, the first portion of the chapter presented a practitioner-oriented framework for the broad basic steps of the process required to conduct a skills diagnostic assessment, including modeling and statistical analysis.

Clear scientific disagreements and controversies exist in both the methods and goals of the approach we advocate, and in the interpretations of progress that has occurred to date. We provide our point of view here, and the reader is encouraged also to read the literature for alternative points of view. Well established and largely agreed-upon recommendations concerning which combinations of model, inference method, computational approach, and evaluation of assessment effectiveness that should work best for practitioners in various settings either do not yet exist or are still subject to controversy. In spite of the lack of consensus concerning various aspects, the foundations and basic building blocks of the field are well developed, numerous models and several statistical approaches are available for application and study, and various appropriate computational software packages exist. Thus, skills diagnostic assessment studies can currently be carried out to pursue the many questions that remain.

Alternative statistical computation methods such as MCMC and EM algorithms can be compared along several important dimensions. For example, the EM algorithm demonstrates advantages over MCMC such as greater computational efficiency and direct focus on estimation of particular parameters. EM disadvantages include greater difficulty in developing the analytical likelihood functions necessary to implement new models, especially complex models, and less straightforward determination of standard errors (Junker, 1999; Junker & Patz 1999a, 1999b).

MCMC advantages over EM include relatively greater ease in setting up new models, and the generation of the estimated posterior distribution of all the parameters. The estimated posterior distribution directly aids determination of multiple modes and better control of standard errors. These advantages come at certain costs, including longer computation time, and greater difficulty of determining convergence since the convergence is not that of individual parameters, but convergence of a joint posterior with many variables (all the model parameters). These issues, as they relate specifically to cognitive diagnostic models, should be a target of future setting-specific research.

As we argue in the chapter, development of effective diagnostic assessment, whether for experiment or for trial use within a program, requires sufficiently rich test data. Better diagnostic performance can be
expected if the data come from a test carefully designed for the purposes of skills diagnosis. The selection of an appropriate cognitive diagnostic model based on an analysis of the cognitive interaction between the skills and the items on the test, almost always requires consultation of the literature and close collaboration among psychometric and substantive experts, in addition to empirical checking and confirmation. As emphasized in the chapter, one should plan a broad-based evaluation of the effectiveness of one’s model-driven skills assessment from a variety of perspectives, including statistical convergence, interpretation of model parameter estimates for reasonableness and internal consistency, evaluation of model diagnostics, reliability estimation, and gathering multiple pieces of evidence in support of internal and external validity. Comparison of fit statistics based on a cognitive diagnostic model with fit statistics based on a standard unidimensional IRT model may have scientific or academic interest, but, as we note in the text of this chapter, whether or not a unidimensional model has better statistical fit does not determine, in and of itself, the appropriateness or usefulness of the diagnostic model. For the fit study to be most useful, the fit measures used to vet the model must be sensitive to the diagnostic purpose of the test.

Given the available foundations of cognitive diagnostic modeling, the field is at a stage where conducting pilot skills diagnostic studies involving real test data and for various test design settings (including tests already designed and originally intended for continuous scaling of examinees rather than diagnosing them) is a natural and necessary next step in the evolution of the field. The resulting feedback from numerous such real-data-based pilot studies should sharpen the focus of future model building research and associated statistical methodology development, should contribute to the foundations of diagnostic assessment design and task development, and should resolve many of the outstanding open issues that remain. Such diagnostic research will make clearer which combinations of models, procedures, and computational approaches are to be preferred, as a function of varying instructional settings and assessment purposes. In addition, future research should address some of the more fundamental questions that have been raised about the saliency of the cognitive diagnostic enterprise.

One fundamental requirement for the approach we advocate here is that the skills and the Q of the CDM represent what substantive experts believe to be the essential skills and their item/skill interactions. That is, to optimize effectiveness, a skills diagnostic analysis of data must accept ab ovo the approximate validity of the proposed cognitive diagnostic model. Then, as long as the diagnostic model with its expert-chosen Q matrix fits the data reasonably well, based on fit measures that are relevant to the skills diagnostic analysis, it will be appropriate to use the model to draw skills diagnostic conclusions, even if the diagnostic model fits slightly less well than a unidimensional model, say. As we note in the body of the chapter, parsimony is vitally important, but only what we have termed parsimony well conceived: The simplest solution is preferred among all possible solutions that satisfy the diagnostic purpose.
We also note that high correlations among the skills and/or low individual skill reliabilities have the potential to cause measurement problems. One implication of this recognition is that it is helpful if the diagnostic model and estimation method used are equipped to model and estimate inter-skill correlations. We note that a most extreme case may occur in which five skills, say, are defined with perfectly unidimensional data. In such a case the ordered individual skill classifications would be completely determined by four thresholds on the unidimensional scale. As long as the individual skills are clearly defined substantively in terms of a $Q$ matrix that connects cognitive skills and items, it may be an extremely useful service to teachers and learners to know that each particular scale score denotes a particular profile of skill proficiencies. The point of the example is that even in the most extreme case that data are perfectly unidimensional, there may be statistically supportable, useful diagnostic information that can be given to teachers and learners, stated in terms of the cognitive diagnostic skills rather than an abstract scale score.

The typical formative assessment student profiles that educators and students require for their teaching and learning purposes are almost always more substantively complex and detailed than what a single score can provide. For example, each examinee might be assigned a nine component discrete mastery/nonmastery algebra skills diagnostic profile as a result of an algebra skills diagnostic assessment. This sort of additional parametric complexity includes a real danger that selected models may be non-identifiable or, more generally, may demonstrate unstable estimation or computation. The careful user needs to watch for this possibility.

We emphasize that model fit is not the goal, rather valid skills diagnosis is. If, for example, the multivariate latent distribution of the skills is such that they are highly correlated or that they satisfy a low dimensional hierarchical structure, the skills diagnosis is no less valid and no less important in the classroom. If such correlational information can be inferred from data, it may on occasion improve the accuracy of the skills diagnosis.

On one hand the use of cognitive diagnostic models presents some very real dangers due to model complexity, and on the other hand, the use of overly simple models fails to provide the diagnostic benefit that is required. Thus, one of the key emerging challenges facing both researchers and practitioners is the finding of a comfortable middle ground of moderate model complexity such that rigorous statistical thinking supports use of the chosen models and the consequent statistical methodology chosen, but also where there is just enough well-estimated parametric complexity to provide student profiles with enough cognitive skill information to be useful for student and instructor.

It is important to note that most test data that have been currently available for skills diagnostic analysis have been from tests intended for continuous unidimensional scaling. By contrast, future skills diagnostic assessments ideally will be generated from tests specifically designed to assess discrete levels of mastery of a targeted set of skills. Such designed tests may display much stronger latent multidimensionality as will
The DSM pathological gambling diagnostic example (Templin & Henson, in press) of Section 3 provides an instance where the parametric complexity resulting from the multiple latent components of the complex multidimensional diagnostic model improves on the practitioner's capacity to draw conclusions about gambling profiles. The multiple dimensions that were found were so strongly dimensionally distinct from each other that, \textit{a fortiori}, information was provided that could not have been made available from a unidimensional model based analysis. The further consequential validity question of whether this additional information obtained can in practice actually improve the diagnosis or treatment of gambling addiction is a subject of interesting future research. Also, many more such real data examples, some more explicitly defended by statistical analyses and especially some involving consequential validity, is another important topic for future research.

Some would argue that relatively simple and perhaps even approximately unidimensional IRT models, accompanied by targeted sophisticated statistical analyses, can produce as much diagnostic information as can be reasonably expected for test lengths appropriate for classroom use. We disagree with that assertion and welcome the opportunity to address the issue in future studies. Of course, the resolution of this controversy may have differing general conclusions for the different types of student test data. In particular, for example, there may be one answer for assessments that are strongly unidimensional and for which the diagnostic question is translated into providing better skills-based information about the score scale. A different answer may be expected for tests that are designed for the explicit purpose of providing more highly multidimensional, discrete skill profile information.

Reliability has always been a conceptual cornerstone of psychometrically based educational testing, and remains so for diagnostic testing. Perhaps the most salient question for skills diagnostic testing is understanding and establishing standards for this new notion of skills classification reliability. The chapter summarized one practical method of defining and estimating an index of reliability. Given estimates of diagnostic classification reliability, the literal transfer of standardized test reliability standards to diagnostic testing may not be appropriate. If, for example, a Cronbach alpha or KR20 of at least 0.9, say, represents high reliability for a high stakes college admission test, it does not necessarily follow that the same level of 0.9 estimated correct classification reliability or skills classification test-retest consistency is required for a successful diagnostic test intended for classroom use. The needed standards of reliability for these diagnostic classification reliability indices must be determined in reference to the changed interpretation of the new scale and the changed diagnostic testing setting and purposes for the test, and must evolve independently through research and experience over time.

A multitude of interesting and important research questions remain concerning skills diagnostic assess-
ment, especially concerning the statistical aspects of such diagnoses. These questions include foundational questions as well as applications questions, and comprise psychometric and statistical aspects as well as applications work that combines the statistics with substantive areas of cognitive science, pedagogy, and learning theory. The benefits of widely providing effective skills diagnostic assessment in the near future, designed to improve teaching and learning in instructional settings, are potentially very large. Moreover, as this chapter argues, based on the current state of the art, practitioners and researchers alike can currently carry out skills diagnostic assessments in multiple ways with tools to evaluate performance.

We hope this concluding discussion spurs increased activities along two different and vital streams of research. The first is to assess how wide and deep the value of skills diagnosis assessment is when applied in real instructional settings. This is partly a question of how much value is added by the diagnostic information from carefully designed and analyzed diagnostic tests. The consequential validity question is whether the added information directly and demonstrably improves teaching and learning.

The second stream of research is statistical and psychometric, namely to continue to refine various diagnostic approaches, to introduce new ones, and to compare the performances of the major approaches with the goal of setting-specific optimization of the retrieval of skills diagnostic information. Some models, statistical methods, computational approaches, and evaluation approaches will fade away but the needs are so rich and varied that the authors predict that many will survive and prosper.

References


Table 1: Effect of diagnosis-based remediation on seriousness of student errors based on pretest and posttest data. Results reproduced from Table 5 in Tatsuoka & Tatsuoka (1997).

<table>
<thead>
<tr>
<th>Pretest</th>
<th>Posttest</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serious</td>
<td>21</td>
<td>93</td>
<td>114 (51%)</td>
<td></td>
</tr>
<tr>
<td>Nonserious</td>
<td>1</td>
<td>107</td>
<td>108 (49%)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>200</td>
<td>222</td>
<td>(10%) (90%)</td>
</tr>
</tbody>
</table>

Table 2: List of reviewed cognitive diagnosis models.

<table>
<thead>
<tr>
<th>Model Abbreviation</th>
<th>Model name</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1PL</td>
<td>One-parameter logistic</td>
<td>Rasch (1961)</td>
</tr>
<tr>
<td>2PL</td>
<td>Two-parameter logistic</td>
<td>Birnbaum (1968)</td>
</tr>
<tr>
<td>3PL</td>
<td>Three-parameter logistic</td>
<td>Birnbaum (1968)</td>
</tr>
<tr>
<td>DINO</td>
<td>Deterministic-Input Noisy-Or</td>
<td>Templin &amp; Henson (in press)</td>
</tr>
<tr>
<td>GLTM</td>
<td>General component latent trait</td>
<td>Embretson (1985, 1997)</td>
</tr>
<tr>
<td>HYBRID</td>
<td>HYBRID</td>
<td>Gitomer &amp; Yamamoto (1991)</td>
</tr>
<tr>
<td>LLTM</td>
<td>Linear logistic test</td>
<td>Fischer (1983)</td>
</tr>
<tr>
<td>MCLCM-C</td>
<td>Compensatory MCLCM (Multiple classification latent class)</td>
<td>Maris (1999)</td>
</tr>
<tr>
<td>MCLCM-D</td>
<td>Disjunctive MCLCM</td>
<td>Maris (1999)</td>
</tr>
<tr>
<td>MIRT-C</td>
<td>Compensatory multidimensional IRT</td>
<td>Reckase &amp; McKinley (1991)</td>
</tr>
<tr>
<td>MIRT-NC</td>
<td>Non-compensatory MIRT</td>
<td>Sympson (1978)</td>
</tr>
<tr>
<td>MLTM</td>
<td>Multicomponent latent trait</td>
<td>Whitely (1980)</td>
</tr>
<tr>
<td>RLCM</td>
<td>Restricted latent class</td>
<td>Hartz (1984, 1990)</td>
</tr>
<tr>
<td>RUM</td>
<td>Reparameterized unified cognitive/psychometric</td>
<td>DiBello, Stout, &amp; Roussos (1995)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hartz (2002), Hartz &amp; Roussos (2005)</td>
</tr>
</tbody>
</table>
Table 3: Knowledge structure of reviewed cognitive diagnosis models.
U = Unidimensional  M = Multidimensional.
C = Continuous    D = Discrete.

<table>
<thead>
<tr>
<th>Skill mastery scale</th>
<th>Dimensionality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U</td>
</tr>
<tr>
<td>C</td>
<td>1-2-3PL, LLTM</td>
</tr>
<tr>
<td>D</td>
<td>DINO, MCLCM-C, MCLCM-D, RLCM, RUM (Q complete)</td>
</tr>
<tr>
<td>C/D</td>
<td>RUM</td>
</tr>
</tbody>
</table>

Table 4: Item structure for Q skills.
S = Single  M = Multiple.
C = Compensatory-Disjunctive  NC = Non-compensatory- Conjunctive
N.A. = Not applicable.

<table>
<thead>
<tr>
<th>Skill interaction</th>
<th>Number skill components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
</tr>
<tr>
<td>C</td>
<td>LLTM, MIRT-C, MCLCM-D, MCLCM-C</td>
</tr>
<tr>
<td>NC</td>
<td>MIRT-NC, GLTM, MLTM, RLCM, RUM, DINO</td>
</tr>
<tr>
<td>N.A.</td>
<td>1-2-3PL</td>
</tr>
<tr>
<td>NC/N.A.</td>
<td>HYBRID</td>
</tr>
</tbody>
</table>
Table 5: Classification of cognitive diagnosis models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Multi-dimensional</th>
<th>Non-compensatory</th>
<th>Q incomplete</th>
<th>Multiple strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1PL</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>2PL</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>3PL</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>LLTM</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>MIRT-C</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>MIRT-NC</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>GLTM</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>MLTM</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>MCLCM-C</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>MCLCM-D</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>DINO</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>RLCM</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>HYBRID</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>RUM</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>