5

Construct Validity and Cognitive Diagnostic Assessment

Xiangdong Yang and Susan E. Embretson

INTRODUCTION

Cognitive diagnostic assessment (CDA) is increasingly a major focus in psychological and educational measurement. Instead of inferring a general response tendency or behavior consistency of an examinee over a target domain of measurement, diagnostic assessment results provide a detailed account of the underlying cognitive basis of the examinee's performance by mining the richer information that is afforded by specific response patterns. Sophisticated measurement procedures, such as the rule-space methodology (Tatsuoka, 1995), the attribute hierarchy method (Leighton, Gierl, & Hunka, 2004), the tree-based regression approach (Sheehan, 1997a, 1997b), and the knowledge space theory (Doignon & Falmagne, 1999), as well as specially parameterized psychometric models (De La Torre & Douglas, 2004; DiBello, Stout, & Roussos, 1995; Draney, Pirolli, & Wilson, 1995; Hartz, 2002; Junker & Sijtsma, 2001; Maris, 1999), have been developed for inferring diagnostic information.

Although measurement models for diagnostic testing have become increasingly available, cognitive diagnosis must be evaluated by the same measurement criteria (e.g., construct validity) as traditional trait measures. With the goal of inferring more detailed information about an individual's skill profile, we are not just concerned about how many items have been correctly solved by an examinee. We are also concerned about the pattern of responses to items that differ in the knowledge, skills, or cognitive processes required for solution. Similar to traditional tests, empirical evidence and theoretical rationales that elaborate the underlying basis of item responses are required to support the inferences and interpretations made from diagnostic assessments.

Construct validation, as elaborated by Messick (1989), is the continuing scientific process of collecting various kinds of evidence to justify the inferences that are drawn from observed performances of examinees. The covert nature of psychological constructs, however, makes construct validation an inherently difficult process. To make inferences about the unobservable traits or qualities, observable indicators must be either identified or designed to elicit the examinee behaviors. The process of item design and test assembly results in establishing such behavioral indicators for the unobservable latent traits or qualities.

A systematic and defensible approach to item design is especially significant for the construct validity of diagnostic assessment (see Gorin, this volume). However, the traditional item design approach does not achieve such requirements for several reasons. First, traditional item design has been primarily concerned with developing items for stable, self-contained latent traits (Messick, 1989; Mislevy, 1996). Empirical relationships of test items with each other or with other external traits are often deemed to establish item quality in this case. Cognitive diagnosis, in contrast, requires a more direct understanding of the mental processes involved in test items. Second, traditional item design has long been viewed as an artistic endeavor, which mainly depends on the item writer's expertise, language skills, and creativity in the subject domains. The traditional item design approach normally lacks either theories to understand how specific features of items impact the cognitive basis of performance or relevant research methods to test such constructs (Embretson, 1983). As a result, other than some general guidelines or principles about item format, content, and mode, no detailed description or empirical evidence is available to support the relationship between content features of the items and the constructs under investigation.

As noted by Leighton et al. (2004), cognitive item design can provide a basis for diagnostic assessment. In cognitive item design, cognitive theory is incorporated into test design (Embretson, 1994, 1998). If the theory is sufficiently well developed, it elaborates how item stimuli influence the cognitive requirements of solving the item. Therefore, item performance is explicitly linked to its underlying cognitive variables, and the cognitive theory explicates the underlying measurement construct of the test.

This chapter concerns the implications of construct validity for CDA. The material in this chapter is organized as follows. First, current views on construct validity are discussed and summarized. Second, the unique issues of construct validity within the framework of CDA are discussed. The discussion includes a description of the cognitive design system approach to item design and how it relates to construct validity issues of CDA. Third, an example is presented to illustrate and contrast the implications of cognitively designed items for the construct validity of both traditional trait measures and CDA. Fourth, a summary of the approach that is taken in this chapter and discussions of the relevant issues are provided.

CONSTRUCT VALIDITY: GENERAL FRAMEWORK

Construct validity is a multifaceted yet unified concept. *Construct validity* concerns the degree to which empirical evidence and theoretical rationales support the inferences made from test scores. Numerous articles on validity have been published since Cronbach and Meehl's (1955) elaboration of the concept. The most important results from these developments are (a) the presentation in the current *Standards for Educational and Psychological Testing* (American Education Research Association/ American Psychological Association/National Council on Measurement in Education, 1999), in which validity is conceptualized differently than in the previous versions, and (b) an extensive integration of several aspects of validity into a comprehensive framework (Messick, 1989, 1995). According to both developments, no longer may validity be considered to consist of separate types, as emphasized in the prior version of the *Standards for Educational and Psychological Tests*. The separate types of validity were construct validity, criterion-related validity, and content validity, which were differentially appropriate, depending on test use. Instead, the concept of construct validity is now articulated within a unifying framework of construct validity.

Messick differentiated six aspects of construct validity to apply to all tests. Two traditional types of validity – content validity and criterion-related validity – are conceptualized as different sources of evidence for construct validity by Messick (1989). First, the *content* aspect concerns the relevancy and representativeness of test content to the construct. For any test, including ability tests, test content is important to evaluate for appropriateness to the inferences made from the test. Test content may concern either surface or deep structural features of content. Second, the *substantive* aspect concerns the theoretical rationale and evidence about the processes behind test responses. On an ability test, the relevancy of the processes employed by examinees to solve items to the intended construct should be assessed. For example, if solving quantitative reasoning items depends primarily on using information from the distracters, the measurement of reasoning as a general top-down approach to problem solving would not be supported. Third, the *structural* aspect concerns the relationship of the scoring system to the structure of the construct domain. Factor analytic studies are relevant to this aspect of validity. If scores are combined across items and factors, empirical evidence should support this combination. Fourth, the *generalizability* aspect concerns the extent to which score interpretations may be generalized to varying populations, conditions, and settings. Research on adverse impact, use of paper-and-pencil versus computerized testing, are relevant to this aspect of validity. Fifth, the *external* aspect concerns the correlations of test scores with criteria and other tests. Studies of predictability of criteria, as well as multitrait-multimethod studies, are relevant to this aspect of validity. Sixth, the *consequential* aspect concerns the social consequences of test use, such as bias, fairness, and distributive justice.

The *substantive* aspect of construct validity is especially related to item design and interpreting the basis of an examinee's performance. As with other aspects of construct validity, it requires empirical evidence. The type of evidence required for substantive validity goes beyond the individual differences studies that were envisioned as supporting construct validity (Cronbach & Meehl, 1955). Embretson (1983) distinguished an interpretation of construct validity, *construct representation*, that requires evidence previously more typical of experimental psychology studies. That is, *construct representation* refers to the cognitive processing components, strategies, and knowledge stores that persons apply directly in solving items. The evidence that is required to support *construct representation* includes studies on how performance is affected by aspects of the item stimuli that influence various underlying processes. Thus, evidence such as experimental studies to manipulate item stimuli, mathematical modeling of performance, and eye tracker studies are needed to support this aspect of validity. Such studies are also relevant to item design because their results indicate how the cognitive complexity of items can be specified by variations in an item's stimulus features.

For all aspects of construct validity, Messick (1989, 1995) notes two major threats: (a) construct *underrepresentation* and (b) *construct irrelevant variance*. *Construct underrepresentation* occurs when important aspects or facets of what is being measured are omitted. A test of quantitative reasoning, for example, that included only algebra problems would be too narrow and consequently would underrepresent the reasoning construct. *Construct irrelevant variance*, in contrast, occurs when performance depends on qualities that are not considered part of the construct. For example, if a test for quantitative reasoning involves an undue dependence on language, then construct irrelevant variance is introduced. For individuals with less competence in a language, such quantitative reasoning tests become more a measure of language proficiency than of quantitative reasoning.

CONSTRUCT VALIDITY: UNIQUE ISSUES OF COGNITIVE DIAGNOSTIC ASSESSMENT

In addition to the aspects of construct validity for general testing practice, CDA bears its own distinctive issues in terms of construct validity. This section focuses on some aspects of construct validity for CDA. However, it doesn't mean that the aspects of validity that are not given special discussion here are not important. On the contrary, diagnostic assessments that fail to satisfy the fundamental validity requirement of sound measurement instruments most certainly fail to be defensible diagnostic instruments.

The Meaning of Diagnosis

Probably the best place to start the discussion is to ask the following question: "What makes CDA distinctive relative to other type of assessment?" To begin with, we examine the meaning of the word "diagnosis". For example, in the *American Heritage Dictionary of the English Language* (2000), diagnosis is defined as the following:

1. *Medicine.*
   a. The act or process of identifying or determining the nature and cause of a disease or injury through evaluation of patient history, examination, and review of laboratory data.
   b. The opinion derived from such an evaluation.
2. a. A critical analysis of the nature of something.
   b. The conclusion reached by such analysis.
3. *Biology.* A brief description of the distinguishing characteristics of an organism, as for taxonomic classification. (p. 500)

From such a definition, it may be safely inferred that diagnosis has been primarily applied in the field of medicine and biology, while different meanings are attached to the same word in different fields. Ignoring the differences across fields, it seems that at least three aspects of diagnosis can be extracted: (a) a description of the distinguishing characteristics of a thing or phenomenon, (b) identifying or determining the nature of a thing or causes of a phenomenon, and (c) the decision or conclusion that is made or reached by such description or analysis. Such a decision or conclusion could be a decisive classification of the thing into some prespecified categories such as diagnosing a patient as having pneumonia, or it could be an assertive statement of the mechanism that leads to the observed phenomenon, such as failure of the cooling system that leads to the overheating of the car engine.

Goals of Cognitive Diagnostic Assessment

Both the descriptive and causal-seeking approaches of cognitive diagnostic testing have been studied. For example, in his LISP tutor (Anderson, 1990), which is an intelligent tutoring system for teaching programming in LISP, learner's cognitive characteristics are represented by a detailed production system of cognition. Both how students actually execute the production rules and how these rules are acquired are specified in the LISP tutor. Diagnostic evaluation in the system is to assess the set of production rules that have been mastered at a particular stage by a student. Similarly, White and Frederiksen (1987) developed a tutoring system called QUEST that teaches problem solving and troubleshooting of electricity circuit problems. Students' mental characteristics are represented by various qualitatively different mental models, which are mental representations of the students' declarative and procedural knowledge, as well as strategic behaviors afforded by such representations. Other representations that are employed in diagnostic assessments include the conceptual or semantic network of declarative knowledge (Britton & Tidwell, 1995; Johnson, Goldsmith, & Teague, 1995; Naveh-Benjamin, Lin, & McKeachie, 1995). Diagnostic evaluations are conducted by comparing such presentations of domain experts with those of novices to detect possible misconceptions of the latter.

Alternatively, Brown and Burton (1978) represented students' problem-solving processes in basic mathematical skills through a directed procedural network, in which the set of declarative and procedural knowledge is connected with each other in a goal-structured fashion. The various misconceptions student hold in the domain are also represented in the network as incorrect alternatives of the correct procedure. The resulting diagnostic modeling system, the DEBUGGY system, provides a mechanism for explaining why a student is making a mistake in basic algebraic problems, not just identifying the mistake itself. In contrast, Embretson and Waxman (1989) developed several cognitive processing models for spatial folding tasks in which four cognitive components were identified as (a) encoding, (b) attaching, (c) folding, and (d) confirming. By empirically establishing the relationships between task features and the underlying components, specific patterns of examinees' responses to spatial tasks can be explained by identifying the specific sources of cognitive complexity involved in item solution. A similar approach can also be found in the work of Das, Naglieri, and Kirby (1994).

In short, cognitive diagnostic testing in a psychological or educational setting mainly focuses on at least three aspects of cognitive characteristics:

1. Skill profiles or knowledge lists that are essential in a given cognitive domain. Those skill and knowledge sets represent the most important skills and concepts of the domain, and serve as the basic building blocks for developing any other higher-order competency.
2. Structured procedural and/or knowledge network. Knowledge and skills are represented in our minds in a highly structured fashion (Collins & Loftus, 1975; Rumelhart, 1980). Expertise in a domain is represented not only by the number of basic skills or pieces of knowledge possessed in the domain, but also by the structure or organization of such skills and knowledge (Chi, Glaser, & Farr, 1988; Ericsson & Charness, 1994).
3. Cognitive processes, components, or capacities. The information processing paradigm of cognitive research provides methods to tap into the internal processes of cognition so specific cognitive models can be developed for a particular type of cognitive task. Observed performances therefore can be explained by looking into examinees' underlying cognitive processes when they perform such tasks.

These three aspects of cognitive characteristics are not exhaustive. Higher-order thinking skills such as cognitive strategy, strategy shifting, and metacognitive skills, should also be included in diagnostic assessment but may be limited by the development of testing techniques at present (Samejima, 1995; Snow & Lohman, 1993).

Issues of Construct Validity for Cognitive Diagnostic Assessment

The special goals of CDA inevitably bring forth particular issues of construct validity. The following discussion of construct validity focuses on both the representation of measured construct and the design of CDAs.

Construct Representation

As mentioned in the previous section, construct representation refers to the theory or rationale of the specification of the construct itself. For diagnostic purposes, the construct is usually represented in fine-grained forms, such as the procedural networks and processing models, to adequately capture the complexity of examinees' cognitive characteristics.

One important aspect of construct validity is the appropriateness and completeness of construct representation. Appropriateness of construct representation addresses whether the form or the symbol system we adopted to describe the cognitive characteristics is suitable. For example, the traditional latent trait perspective of intelligence, mainly through the dimensional theory of factor analysis, leads to enormous knowledge about the interrelationships among such latent traits, but little knowledge about the representation of the latent traits themselves. Similarly, cognitive psychologists have recognized the necessity of distinguishing between declarative and procedural knowledge because of their distinctive characteristics, which requires different approaches to represent such knowledge appropriately. Therefore, for a specific diagnostic testing purpose, adopting an appropriate representation of the construct is essential.

The completeness of construct representation addresses whether the construct has been identified adequately. As mentioned previously, two threats to construct validity are construct underrepresentation and construct irrelevant variance. For diagnostic assessment, however, complete representation of the construct may not be a practical option given the complexity of the examinee's cognitive characteristics in a given domain. This is exemplified in most of the intelligent tutoring systems that aim to capture the detailed processes of cognitive performance and learning mechanisms. It is a daunting task even for a highly structured cognitive domain. For example, to accurately diagnose students' misconceptions in solving algebraic problems, the diagnostic system has to encode all misconceptions that might appear in students' performances (Brown & Burton, 1978). Alternatively, restrictions on the breath of the construct might have to be imposed to achieve the depth of such representation.

Another important aspect of construct representation for cognitive diagnostic testing is the issue of granularity. Granularity is closely related to the issue of completeness but has its own implications. From a psychometric modeling perspective, granularity might be related to the capacity of measurement models, the affordance of the data, and the availability of computational power that are in contrast with the specificity of the cognitive representation (DiBello et al., 1995). Relative to the detailed representation of examinees' cognitive performance, current development in psychometrics might not provide a practical modeling approach. Thus, some degree of simplification of the cognitive representation might be needed. Besides the technical limitations, granularity has substantial implications that are related to the validity of diagnostic inferences as well. That is, at what level should the construct be represented so that both adequacy of representation and generalizability of the diagnostic inferences are maintained? For example, the production rule system representation of the LISP encodes examinee's problem-solving process in such a detail that 80% of the errors in student performance can be captured (Anderson, 1990). So it is adequate in terms of the construct representation. In contrast, however, the fine-grained diagnoses that are afforded by the representation makes such inferences confined to the highly limited situation, which results in very limited generalizability (Shute & Psotka, 1996). Thus, depending on the purpose of the diagnostic assessment, setting the appropriate level of granularity in terms of the representation is crucial for the validity of the inferences that are made from testing results.

Test Design and Administration for Diagnosis

Once the appropriate representation of the construct for diagnosis has been determined, one step for diagnostic testing is to design the items and to assemble the test to elicit examinees' observable behaviors so desirable diagnostic inferences can be made. This is a central step for defensible CDA, for which several issues will become relevant in terms of construct validity.

As mentioned previously, the covert nature of psychological constructs make the task of measuring such constructs inherently difficult. Two questions are essential to the mapping between observable indicators and the unobservable constructs. The first question is how can we be confident of the fact that the items we designed are actually measuring the construct that we are intending to measure? This indeed is a central question for validity, which has been bothering researchers and test specialists in the field of measurement from the beginning. Traditionally, as mentioned in the previous sections, such confidences come primarily from our faith in the opinions of the subject-matter experts. After the creation of test items, techniques such as factor analysis serve as post-hoc verification of such knowledge with regard to these items. Recent development of item design, however, brings both methodology and findings from cognitive research into the design process (Irvine & Kyllonen, 2002). By modeling the cognitive processes involved in item solution and identifying various sources of item difficulty that are caused by different cognitive components, confidence in the new approach of item design comes from how sound the cognitive theory behind the proposed models is, how well the proposed cognitive models fit the observed data that are collected over the designed items, or both. Such an item design approach has many implications for CDA. More discussion of this point is presented in the next section.

Aside from the effort to design and select items to be valid indicators of the construct, it is also a general concern whether certain aspects of cognitive characteristics are indeed measurable, at least with some item formats or approaches of item administrations. For example, Snow and Lohman (1993) discussed how the conventional approach of item design and administration might not be sufficient to measure cognitive characteristics such as qualitatively distinct strategies, strategy shifting, and the adaptiveness and flexibility that individuals exhibit in their problem-solving processes during testing. However, those aspects are undoubtedly facets of cognitive characteristics that are essential to diagnostic assessment. When an individual faces a cognitively complex task, not only the possession of the basic component skills, but also the ability of dynamically restructuring and innovatively using such component skills are essential to successfully solving the task. Recognizing the importance of incorporating such higher-order thinking skills into diagnostic assessment and the feasibility of assessing such characteristics, Samejima (1995) categorized cognitive processes into three categories: (a) processes that are assessable by conventional paper-and-pencil tests, (b) processes that are assessable by computerized tests with the possibility of using innovative test designs or administration, and (c) processes that are not assessable by either of the two testing methods, which require extensive experimental observations. Clearly, for diagnostic assessment to be valid, the approaches to designing the item, as well as the methods to administering the test, have to be considered.

Another important aspect of the test design and administration for diagnosis is sampling, which is relevant to both construct representation and the content aspect of validity. With the fine-grained representation of cognition, item sampling for diagnostic assessment becomes more complicated than for conventional assessment. In conventional assessment, the goal is to infer the general tendency to solve items in a given domain, where the general tendency is usually defined as a single or a few latent theoretical constructs. A sampling theory to measurement applies in this case, in which item sampling is done through defining a universe of items that are indicators of the latent constructs, and then selecting a random (representative) sample of items from the universe (Kane, 1982). When the goal of assessment is to infer the knowledge structure or procedural network of an examinee in a given domain, as diagnostic assessment does, definition of an item universe is much more sophisticated, if not impossible. For example, to measure knowledge networks or schemas (Marshall, 1990), items have to be sampled to measure both the nodes (i.e., the declarative facts or procedural knowledge) and the lines (that connects the nodes or the set of knowledge points) in the network. There are many unsolved issues in sampling items for this type of construct, such as number of items required for each node or line, how to estimate the prevalence of different nodes (statistical distributions across nodes), and so forth.

If the goal of diagnosis is to identify the mental processes of problem solving like those taken by the cognitive approach to measurement (Embretson, 1994), sampling schema must change accordingly. In such cases, a structural theory of measurement could apply (Guttman, 1971; Lohman & Ippel, 1993). In the cognitive process modeling approach to measurement, systematically varying different features of the tasks and the testing conditions could differentially exert influence on different cognitive processes of solving the task. A structural theory of measurement, therefore, states that measurement should focus on the pattern (structure) of observations under various situations, and not the random sample of observations from a defined universe. For example, in the cognitive processing model developed by Embretson and Waxman (1989) for the spatial folding tasks, two important task features are the number of pieces to be folded and the orientation of the markings on the pieces. These task features will systematically but differentially affect the complexity of different cognitive operations in solving the items.

Therefore, examinee's cognitive processing under various conditions can be elicited through a systematic arrangement of these task features. In this case, variations of responses to different tasks or under different testing situations are not random fluctuations, but rather authentic reflections of the differential requirement of cognitive loadings. Item sampling under this perspective of measurement, therefore, should take into account the structural relations among tasks.

EXAMPLE OF CONSTRUCT REPRESENTATION STUDIES FOR TRAIT MEASUREMENT VERSUS COGNITIVE DIAGNOSIS

In trait measurement, which includes ability testing, the goal is to estimate examinees' standings on one or more latent traits that are postulated to underlie test performance. For example, scores from a particular cognitive test may indicate the examinee's standing on one or more abilities that underlie performance. The link between test performance and the latent trait is established through construct validity studies. Traditionally, trait measurement has been the most popular goal of psychological testing.

Some tests may be appropriate for both trait measurement and cognitive diagnosis. In the following example, items for an intelligence test were designed from a theory about the underlying cognitive sources of complexity. For trait measurement, a construct representation study for construct validity would involve relating the design features of items to item difficulty. For cognitive diagnosis, the same design features can form the basis of classifying examinees in terms of their competences.

In this section, the cognitive design principles behind an intelligence test, the Abstract Reasoning Test (ART; Embretson, 1995) are described. Then, two empirical studies are presented to elaborate the construct representation aspect of construct validity. The first study examines the construct representation of ART for trait measurement, whereas the second study examines the structure of ART for cognitive diagnosis.

Abstract Reasoning Test

The ART was developed by Embretson (1995) using the cognitive design system approach. ART items are matrix completion problems. Matrix completion problems appear on a variety of intelligence and ability tests. They are considered a stable measure of fluid intelligence (Raven, 1965).

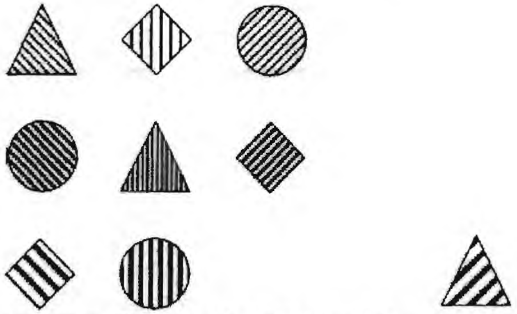

FIGURE 5.1. Example of matrix completion item and the missing entry.

Figure 5.1 gives an example of a matrix completion item. To solve the item, examinees must identify the relationships among the set of entries so the missing entry can be constructed.

Carpenter, Just, and Shell (1990) proposed a theory of processing progressive matrix problems based on results from a variety of experiments. According to their studies, matrix completion problems can be decomposed into different rules or relationships across the rows and columns. Finding and evaluating these relationships are the major processes involved in solving the matrix completion items. Carpenter et al. postulated that five relations are involved in solving matrix completion problems, such as those that appear on the Advanced Progressive Matrix Test (Raven, 1965). These relationships, in order of complexity, are identity, pairwise progression, figure addition or subtraction, distribution of three, and distribution of two. Carpenter et al. (1990) suggested that examinees applied one rule at a time following the hierarchy of these rules. That is, the objects are examined first for identity relationships before considering pairwise progression relationships.

Based on Carpenter et al.'s (1990) processing models for matrix completion problems, Embretson (1995, 1998) developed the ART using the cognitive design system approach. Item structures of ART were specified through the formal notational system (Embretson, 1998), which determines the type and number of relationships in an item. An example of the formal notional system for the item shown in Figure 5.1 is given in Figure 5.2. Two pairwise progressions and one distribution of three are involved in the item. The formal notation system specifies the relations in letters and numbers in which A stands for the triangle, C for diamond, and D for circle. B stands for the grids within the

AB_{41}	CB_{21}	DB_{11}
DB_{42}	AB_{22}	CB_{12}
CB_{43}	DB_{23}	AB_{13}

FIGURE 5.2 Formal notation system for item in Figure 5.1.

objects that show different patterns. The subscripts denote the systematic changes of the grids across rows and columns. Although changes of the first subscript denote the changes of the orientations of the grids, changes of the second subscript denote the changes of their intensities (or weights). Items with the same structure can then be generated by varying different objects or attributes. A set of different item structures can be generated by specifying different combinations of relations and/or abstraction components. The current ART has 30 item structures, each of which has five structurally equivalent items.

Cognitive Psychometric Modeling of Abstract Reasoning Test Item Properties

In this approach, the construct representation aspect is explicated by mathematically modeling ART item properties from the difficulty of the processes involved in item solution. The approach involves postulating a processing model and then specifying the stimulus features in items that determine process difficulty, such as the number of relationships, number of separate objects, and so forth.

The processing theory underlying the cognitive psychometric modeling of ART items was based on Carpenter et al.'s (1990) theory plus an encoding process. In the Carpenter et al. theory, two major processes for matrix completion items such as ART include correspondence finding and goal management. Correspondence finding is primarily influenced by the level of the relationship, as described previously. The highest-level relationship, distribution of two, involves the most abstraction. Goal management, however, depends on the number of relationships in the problem. Carpenter et al. did not include encoding in their theory, primarily because their automatic item solver program required verbal descriptions of the item stimuli.

Embretson (1995, 1998) further developed the processing theory in two ways, that is, combining the relational processing variables into a single variable, namely, memory load, and including an encoding stage that is influenced by several perceptual features of items. The memory load variable includes both the number and the level of relationships. Carpenter et al. (1990) hypothesized that individuals attempted to solve the item with the lowest-order relationships before attempting higher-order relationships. Accordingly, for the highest-level relationship (i.e., distribution of two), all lower-order relationships are assumed to be attempted. The memory load variable is a count of the total number of relationships attempted before reaching the required relationships in the problem. Encoding is influenced by the number of unique attributes, degree of stimulus integration, object distortion, and object fusion. In matrix completion problems, more than one object may appear in a single cell of the design. *Number of unique attributes* refers to the number of separately manipulated objects in the problem stem. *Stimulus integration* refers to the arrangement of the objects. The most integrated display occurs when objects are overlaid, while the least integrated display occurs when two or more objects are displayed around a platform, such as a "+". *Object distortion* refers to corresponding objects for which the shape of one or more is distorted. Finally, *object fusion* occurs when overlaid objects no longer have separate borders.

Estimates of item difficulty from the one-parameter logistic item response theory (IRT) model for 150 ART items were available. The estimates were based on a large sample of young adults. For this set of items, five different items had been generated from each of 30 different item structures. The five variant items differed in the objects or display features. Scores for the items on all cognitive variables were available.

Estimates of item difficulty were regressed on the cognitive variables scored for each item. Table 5.1 presents the overall model summary. Two

TABLE 5.1. *Regression of item difficulty on cognitive model variables*

					Change statistics				
Model	R	R square	Adjusted R square	Std. error estimate	R square change	F change	df1	df2	Sig. F change
1 – Structural only	.758	.575	.569	.92893	.575	99.475	2	147	.000
2 – Structural perceptual	.782	.612	.598	.89737	.036	4.508	3	144	.005

TABLE 5.2. *Coefficients for final cognitive model*

	Unstandardized coefficients		Standardized coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	2.822	.302		–9.332	000
Memory Load	.199	.019	.601	10.664	.000
Number of unique elements	.172	.044	.225	3.923	.000
Object Integration	.387	.168	.129	2.301	.023
Distortion	.507	.260	.105	1.953	.053
Fusions	–.279	.185	–.084	–1.508	.134

variables, memory load and number of unique elements, are specified by the item's abstract structure. These structural variables had a strong and highly significant impact on item difficulty ($R^2 = .575$). The perceptual variables are variations in the display of the objects within a structure. Adding the perceptual variables to the model significantly increased prediction, but the impact was relatively small ($R^2 = .036$).

Table 5.2 presents the standardized and unstandardized regression weights for the final model, which included both structural and perceptual variables. All variables except object fusion had significant weights in prediction. Memory load had the largest beta weight, followed by number of unique elements, again indicating the dominance of the structural variables in predicting item difficulty. Distortion and object integration had smaller and similar beta weights.

In general, the results support the construct representation aspect of construct validity, as specified in the design of the items. The cognitive model variables yielded strong prediction of item difficulty. Of the five variables in the model, ART item difficulty was most strongly affected by the working memory load of the items. These results imply that individual differences in working memory capacity have a strong impact on performance. The pattern of the prediction also supported the feasibility of generating items with predictable properties. That is, the two variables that are specified in the generating structure for ART items, memory load and number of unique elements, had the strongest impact in predicting item difficulty. The perceptual features, which can vary in the same structure, had relatively small impact on item difficulty. Thus, the results support the equivalence of items generated from the same structure, even though specific stimulus features and perceptual displays differ.

Construct Representation Study for Cognitive Diagnosis

Alternatively, the construct of ART may also be represented by the skills of correctly identifying each of the five relationships (rules) involved in solving an ART item. Using a discrete representation of each of the five skills, the ability of a particular examinee to solve an ART item can be regarded as a direct result of whether he or she possesses the required skills to solve the item. For example, the item in Figure 5.1 contains two pairwise progressions and one distribution of three. An examinee is expected to be able to solve the item if he or she possesses the skills of correctly identifying the two relations. To represent the construct measured by ART in this way, however, both the set of skills involved in ART items and the interrelationships among the skills have to be specified. For ART items, the set of skills are the skills associated with identifying the five relations, which constitute the primary source of item difficulty. As for the structural relations among the five skills, Carpenter et al. (1990) speculated that the five relations followed the order: identity (ID) → pairwise progression (PP) → figure addition and substraction (FA) → distribution of three (D3) → distribution of two (D2) (letter in parenthesis stands for an abbreviation of the relation and are used hereafter). The arrow → stands for the surmise relation between the relation to the left and right of the arrow. For example, ID → PP stands for the relation that identification of relation ID is surmised from relation PP. Embretson (1998) found that figure addition/subtraction did not conform to the order given above. Identification of this relation can either be very difficult or very easy. Accordingly, one structural representation among the five relations may be given as follows:

Given the structural relations among the five relations, the following family of admissible subsets of the five skills can be constructed: {Ø}, {ID}, {ID, PP}, {ID, FA}, {ID, FA, PP}, {ID, PP, D3}, {ID, PP, D3, FA}, {ID, PP, D3, D2}, {ID, PP, D3, D2, FA}, where {Ø} refers to the null set

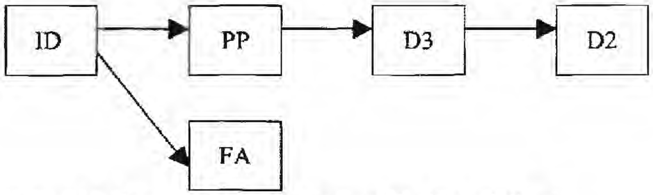

FIGURE 5.3. Structural relations among the five relations.

TABLE 5.3. *Mapping between ability states and ART item structures*

Ability state	Item structure	Ability state	Item structure
{ID}	None	{ID, FA}	10, 12, 19, 21, 23, 24, 27, 28, 32, 34, 35}
{ID, PP}	6, 7, 16, 17, 18, 26, 33	{ID, PP, FA}	{ID, PP} + {FA}
{ID, PP, D3}	8, 9, 13, 20, 25, 29, 39, 40, 41, 46	{ID, PP, D3, FA}	{ID, PP, D3} +{FA}
{ID, PP, D3, D2}	42, 43	{ID, PP, D3, D2, FA}	{ID, PP, D3, D2} + {FA}

that none of the five skills is possessed. We prefer to label each skill set as a latent ability state. In doing so, we imply that a particular latent ability state represents not only a simple list of skills being possessed, but also the higher-order thinking skills that are afforded by the subset of basic skills.

To analyze the ART data under this alternative representation, mapping between different latent ability states and ART item structures needs to be established. Table 5.3 presents the mapping relations for the 30 item structures (see Yang, 2003, for a detailed description of the mapping rules). For illustration, four item structures (7, 21, 29, 42) are selected to measure each of the four skills (PP, FA, D3, D2) (the skill {ID} is excluded because there is no item alone to measure it). For a more detailed analysis of the ART items with regard to inferring diagnostic information from the latent ability-state lattice representation of ART construct, see Yang (2003). From these four item structures, the corresponding ideal response patterns, denoted as the items that can be correctly answered given a particular ability state, are then given as {Ø}, {7}, {21}, {7, 21}, {7, 29}, {7, 29, 21}, {7, 29, 42}, {7, 29, 42, 21}. An intuitive examination of the effectiveness of the hypothesized latent ability structure can be done by simply looking at the proportion of observed response patterns that belong to the ideal response patterns. Table 5.4 gives such an examination based on the data collected from 818 young adults (Embretson, 1998). It can be seen that 88.6% of the examinees fall into one of the eight ideal response patterns, which suggests that the derived ability space fits the data fairly well.

A formal examination of the adequacy of the derived latent ability space to the ART items can be conducted through statistical modeling. Recent developments in both statistical modeling and psychometrics provide the possibility of modeling sets of discrete latent abilities under the framework of latent class/discrete psychometric models (Junker & Sijtsma, 2001; Templin, 2004). For example, a latent class model for the particular example given Table 5.4 can be given as follows:

TABLE 5.4. *Proportion of examinees falling in ideal response patterns*

Pattern	Item {7, 21, 29, 42}	Observed frequency	Ideal response pattern
1	0000	32	0000
2	1000	63	1000
3	0100	16	0100
4	0010	21	–
5	0001	3	–
6	1100	67	1100
7	1010	115	1010
8	1001	6	–
9	0110	29	–
10	0101	4	–
11	0011	6	–
12	1110	248	1110
13	1101	17	–
14	1011	41	1011
15	0111	7	–
16	1111	143	1111
Total		818	725
Percentage			88.6

$$P(\mathrm{x}) = \sum_{c=1}^{C} \pi(\alpha_c) \prod_{j=1}^{J} \left[(1 - s_j)^{\alpha_j q_j} g_j^{(1-\alpha_j)q_j}\right]^{x_j} \left[1 - (1 - s_j)^{\alpha_j q_j} g_j^{(1-\alpha_j)q_j}\right]^{1-x_j},$$

where x, $\mathrm{x} = (x_1, x_2, \ldots, x_J)$ is the observed item response pattern; C is the number of permissible latent ability states; α_c, $\alpha_c = (\alpha_{c1}, \alpha_{c2}, \ldots, \alpha_{cK})$, is the vector that represents the latent ability state c; and α_{ck} (1 or 0) indicates whether ability k, $k = 1, 2, \ldots, K$, is included in the latent state c. The probability of being in the latent ability state c is denoted as π (α_c). q_{jk} (= 1 or 0) indicates whether item j, $j = 1, 2, \ldots, J$, requires ability k and is collected in the matrix $\mathrm{Q}_{J\times K}$, and s_j and g_j are the slip and guessing parameters, respectively (Embretson, 1985; Junker & Sijtsma, 2001). Alternative models can be derived by imposing constraints on s_j and g_j. Table 5.5 presents the results from three alternative models by imposing that (a) $s_j = s$ and $g_j = g$ (constant error rate across items; Dayton & MaCready, 1976), (b) $s_j = g_j$ (item-specific error rate model), and (c) $s_j = g_j = e$ (Proctor, 1970).

TABLE 5.5. *Latent class models fitted to ART data*

		Model I[1]		Model II		Model III	
Item[2] {7,29,21,42}	Frequency	Predicted frequency	Pearson residual	Predicted frequency	Pearson residual	Predicted frequency	Pearson residual
0000*	32	32.86	–0.150	33.03	–0.179	33.76	–0.30
1000*	63	60.62	0.306	61.20	0.230	61.82	0.15
0100	21	15.00	1.549	15.06	1.531	15.90	1.28
1100*	115	120.36	–0.489	120.37	–0.490	119.25	–0.39
0010*	16	16.75	–0.184	16.70	–0.171	16.62	–0.15
1010*	67	66.58	0.052	66.71	0.035	66.67	0.04
0110	29	26.56	0.473	27.31	0.322	27.61	0.26
1110*	248	250.39	–0.151	249.63	–0.103	249.36	–0.09
0001	3	3.65	–0.341	3.19	–0.108	2.53	0.30
1001	6	9.11	–1.030	8.53	–0.867	7.88	–0.67
0101	6	4.46	0.732	4.55	0.678	4.58	0.66
1101*	41	40.95	0.008	41.06	–0.010	41.28	–0.04
0011	4	2.86	0.676	2.72	0.773	2.55	0.91
1011	17	18.05	–0.246	17.86	–0.203	18.42	–0.33
0111	7	13.96	–1.864	14.46	–1.962	14.44	–1.96
1111*	143	135.85	0.614	135.60	0.635	135.33	0.66
Total	818	818		818		818	
G^2		9.68		9.55		8.81	
df		7		6		4	
P value		0.208		0.145		0.066	

Note: [1] Model I, Proctor model; Model II, constant error rate across items; Model III, item-specific error rate model.

[2] Response patterns with asterisk (*) are ideal response pattern.

Table 5.5 can be partitioned into two portions. In the top portion, all observable response patterns from the four items, as well as the associated numbers of examinees in each of the 16 response patterns, were given in the first two columns. Then, the predicted numbers of examinees and the associated Pearson residuals for each of the three fitted models were presented. These results provide detailed information about how adequate a particular model fits the data. Specifically, these results inform us which portion of the data is fitted adequately and which is not. Because Pearson residual approximates to a standard normal distribution for large samples, any number in these columns that is greater than 2 in absolute value indicates a statistically significant deviation (at .05 level) of model-predicted frequency from the observed frequency. It can be seen from Table 5.5 that, for all response patterns, Model I, the Proctor model, fits the data adequately, based on the results of Pearson residual alone.

TABLE 5.6. *Parameter estimates from model 1*

Ability state	Ideal response pattern	Probability	Error
{Φ}	{0000}	0.047	0.093
{PP}	{1000}	0.076	
{FA}	{0010}	0.013	
{PP, D3}	{1100}	0.161	
{PP, FA}	{1010}	0.065	
{PP, D3, FA}	{1110}	0.407	
{PP, D3, D2}	{1101}	0.032	
{PP, D3, FA, D2}	{1111}	0.198	

The bottom portion of Table 5.5 presents values of G^2, degree of freedom (df), and the corresponding P-value for each model. For large samples, G^2 approximates to a chi-square distribution with df $= N-1-m$, where N is the number of all possible response patterns in the data (in this case, $N = 2^4 = 16$) and m is the number of parameters in the model. For example, there are eight parameters in the Proctor model (seven class parameters $\pi(\alpha_c)$ and one slip or guessing parameter e). Therefore, the associate df $= 16 - 1 - 8 = 7$. The G^2 represents an overall goodness-of-fit index for a model. Based on the P-value, it can be seen that all three models in Table 5.5 fit the data well. However, combined with results of the Pearson residual, it seems that Model I, the Proctor model, is the model of choice for this data set.

Table 5.6 presents the estimated probability that a randomly sampled examinee belongs to each ideal response pattern under Model I. It can be seen that examinees are more likely to be in the ability state {PP, D3, FA}. The probability of being in {PP, D3, FA} is .407. They also have substantial probability to be in ability states {PP, D3, FA, D2} and {PP, D3}. The corresponding probabilities are .198 and .161, respectively. This is consistent with a previous study, which showed that ART items are relatively easier for the sample (Embretson, 1998). Table 5.6 also shows that the error probability is small (.093), which can be interpreted either as the probability of getting an item correct without possessing the required skills or as the probability of missing the item with the required skills. This is consistent with the ART item format, which has eight options to choose from and primarily relies on rule identification.

TABLE 5.7. *Inference of latent ability state given response pattern {1011}*

Ability state	$\pi(\alpha_c)$	$P(X \mid \alpha_c)$	$\pi(\alpha_c)^*P(X \mid \alpha_c)$	$P(\alpha_c \mid X)$
{Φ}	0.047	0.0007	0.0000	0.0015
{PP}	0.076	0.0071	0.0005	0.0244
{FA}	0.013	0.0071	0.0001	0.0042
{PP, D3}	0.161	0.0007	0.0001	0.0053
{PP, FA}	0.065	0.0694	0.0045	0.2036
{PP, D3, FA}	0.407	0.0071	0.0029	0.1307
{PP, D3, D2}	0.032	0.0071	0.0002	0.0103
{PP, D3, FA, D2}	0.198	0.0694	0.0137	0.6201
$\Sigma\pi(\alpha_c)^*P(X \mid \alpha_c)$			0.02216	

Given that latent classes in such models are defined theoretically from each latent ability state, a particular examinee's latent ability state can be inferred. In the current example, given the observed response pattern **x** from a particular examinee *i*, the likelihood that such an examinee is in a latent class *c* is given as

$$P\left(\alpha_c \mid \mathbf{x}, \mathbf{s}, \mathbf{g}, \mathbf{Q}\right) = \frac{\pi\left(\alpha_c\right) P\left(\mathbf{x} \mid \alpha_c, \mathbf{s}, \mathbf{g}, \mathbf{Q}\right)}{\sum_{c=1}^{C} \pi\left(\alpha_c\right) P\left(\mathbf{x} \mid \alpha_c, \mathbf{s}, \mathbf{g}, \mathbf{Q}\right)},$$

where $P\left(\mathbf{x} \mid \alpha_c, \mathbf{s}, \mathbf{g}, \mathbf{Q}\right)$ is the probability of observing response pattern **x** conditional on α_c, **s**, **g**, and **Q**. Table 5.7 illustrates the inference for a particular examinee whose observed response pattern is {1011} under Model I. Given an examinee's observed response pattern {1011}, he or she has the highest posterior probability (.6201) of being in the ability state {PP, D3, FA, D2}.

In general, the latent ability state representation of the ART construct provides an alternative approach to cognitive psychometric modeling. This approach to construct representation decomposes the continuous latent construct of ART, such as the memory load, onto a discrete latent ability space. Results from analysis of ART data using latent class/discrete psychometric models show that the hypothetical structure of latent skills involved in solving ART items fits the data quite well. Most of the examinees are likely to possess the majority of the four skills in ART. This result is consistent with previous results from cognitive psychometric modeling, showing that the ART items are relatively easy for this examinee sample. An individual examinee's likelihood of being in a given latent ability state, conditional on his or her observed response pattern, can be inferred from this approach.

SUMMARY AND CONCLUSIONS

CDA has become increasingly important because of its potential to provide a detailed account of the underlying cognitive basis of an examinee's performance. From a broader perspective, the shift from standardized testing to diagnostic testing reflects the response from the measurement field to the challenges from developments in cognitive and educational psychology in recent decades (Mislevy, 1993). With an adequate representation of the measurement constructs, cognitive diagnostic testing could hopefully capture the complexity of examinees' cognitive characteristics, which are reflected in the types of observations and patterns in the data.

For CDA to be valid and effective, its construct validity must be evaluated thoroughly and systematically. In this chapter, contemporary views of construct validity were presented and elaborated. The substantive aspect of construct validity, and particularly construct representation studies, are especially important for cognitive diagnosis. In construct representation studies (see Embretson, 1998), the processes, strategies, and knowledge behind item responses are elaborated. Although construct representation studies are important for elaborating trait measurement constructs, they are even more important for cognitive diagnostic testing. Because cognitive diagnostic testing requires a more fine-gained representation of the measurement construct, several specific issues for construct validity must be addressed: (a) the appropriateness, completeness, and granularity of the construct representation; (b) the design and selection of observable indicators for a fine-grained measurement construct; (c) the measurability of the construct with regard to item formats or test administration procedures; and (d) the appropriateness of the theoretical measurement foundation that is relevant to the specific purpose of diagnostic assessment. This list is far from complete. Because relatively few studies have been done with regard to cognitive diagnostic testing, especially from the perspective of construct validation, many issues of construct validity for diagnostic assessment are left unexplored. Future research in this aspect is essential for gaining a more comprehensive understanding of the complex landscape for cognitive diagnosis.

A systematic and defensible approach to item design is especially significant for the construct validity of diagnostic testing. For instance, in the cognitive design system approach to item design (Embretson, 1998), the relationship of item stimuli to the cognitive requirements of solving the item is elaborated. Item performance is explicitly linked to its underlying cognitive variables, and the cognitive theory explicates the underlying measurement construct of the test. This approach to item design provides an operational mechanism for bridging measurement constructs with item design, which in turn can provide a foundation for the validity of the resultant diagnostic testing.

In this chapter, the cognitive design principles for a reasoning test were examined empirically to illustrate how construct validity can be supported for trait measurement and cognitive diagnosis. For trait measurement, construct representation related the design features of items to item psychometric properties (e.g., item difficulty or discrimination). For cognitive diagnosis, the same design features form the basis for classifying examinees in terms of their competences. Construct validity was supported empirically in the analyses presented for both measurement goals. For cognitive diagnostic testing, which is the main concern in this chapter, the results have several implications. First, the explicit linkage between item stimuli and the underlying cognitive requirements allows items to be generated with targeted cognitive and psychometric properties, which is a feature that is essential to diagnostic testing. Second, the cognitive theory not only identifies cognitive variables that are sources of item difficulties, but also provides a basis for understanding the structural relationship among them. This is important for interpreting test results and sequencing item administration. Third, combined with modern technology, such as an item generation approach, a validated cognitive theory for items has potential for automating individualized diagnostic testing. That is, adaptive testing for individuals can be based not only on the statistical/psychometric properties of items (e.g, IRT-based difficulty estimates), but also on the substantive properties of the items (e.g., knowledge, skills, or cognitive processes that are required for item solution).

Individualized CDA holds many exciting possibilities. Several models and methods have already been developed for diagnostic testing (e.g., Leighton et al., 2004). What clearly remains to be accomplished are construct representation studies on educationally important domains, such as mathematical competency and verbal comprehension, to support the fine-grained aspects of construct validity required for diagnostic testing.

References

American Education Research Association (AERA), American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

The American Heritage Dictionary of the English Language (4th ed.). (2000). Boston: Houghton Mifflin.

Anderson, J. R. (1990). Analysis of student performance with the LISP tutor. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 27–50). Hillsdale, NJ: Erlbaum.

Britton, B. K., & Tidwell, P. (1995). Cognitive structure testing: A computer system for diagnosis of expert-novice differences. In P. Nichols., S. F. Chipman., & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 251–278). Hillsdale, NJ: Erlbaum.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155–192.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404–431.

Chi, M. T. H., Glaser, R., & Farr, M. (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407–428.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). *Assessment of cognitive processes: The PASS theory of intelligence*. Needham Heights, MA: Allyn & Bacon.

Dayton, C. M., & MaCready, G. B. (1976). A probabilistic model for a validation of behavioral hierarchies. *Psychometrika, 41*, 189–204.

De La Torre, J., & Douglas, J.A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

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

Doignon, J. P., & Falmagne, J. C. (1999). *Knowledge spaces*. Berlin: Springer-Verlag.

Draney, K. L., Pirolli, P., & Wilson, M. (1995). A measurement model for complex cognitive skill. In P. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 103–126). Hillsdale, NJ: Erlbaum.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Embretson, S. E. (1985). *Test design: developments in psychology and psychometrics*. Academic Press.

Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Eds.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.

Embretson, S. E. (1995). The role of working memory capacity and general control processes in intelligence. *Intelligence, 20*, 169–190.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 300–326.

Embretson, S. E., & Waxman, M. (1989). *Models for processing and individual differences in spatial folding*. Unpublished manuscript.

Ericsson, K. A., & Charness, N. (1994). Expert performance, its structure and acquisition. *American Psychologist, 49*, 725–747.

Guttman, L. (1971). Measurement as structural theory. *Psychometrika, 36*, 329–347.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality*. Unpublished doctoral thesis, University of Illinois at Urbana-Champaign.

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.

Johnson, P. J., Goldsmith, T. E., & Teague, K. W. (1995). Similarity, structure, and knowledge: A representational approach to assessment. In P. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 221–250). Hillsdale, NJ: Erlbaum.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125–160.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205–236.

Lohman, D. F., & Ippel, M. J. (1993). Cognitive diagnosis from statistically based assessment toward theory based assessment. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 41–71). Hillsdale, NJ: Erlbaum.

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

Marshall, S. P. (1990). Generating good items for diagnostic tests. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 433–452). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Eds.), *Educational measurement* (pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist, 50*, 741–749.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–39). Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416.

Naveh-Benjamin, M., Lin, Y., & McKeachie, W. J. (1995). Inferring student's cognitive structures and their development using the "fill-in-the-structure" (FITS) technique. In P. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 279–304). Hillsdale, NJ: Erlbaum.

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis for Guttman scaling. *Psychometrika, 35*, 73–78.

Raven, J. C. (1965). *Advanced progressive matrices, set I and II*. London: H. K. Lewis. (Distributed in the United States by The Psychological Corporation, San Antonio, TX).

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33–57). Hillsdale NJ: Erlbaum.

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

Sheehan, K. M. (1997a). *A tree-based approach to proficiency scaling* (ETS Research Report No. RR-97–2). Princeton, NJ: Educational Testing Service.

Sheehan, K. M. (1997b). *A tree-based approach to proficiency scaling and diagnostic assessment* (ETS Research Report No. RR-97–9). Princeton, NJ: Educational Testing Service.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present, and future. In D. H. Jonassen (Ed.), *Handbook of educational communications and technology* (pp. 570-600). New York: Macmillan.

Snow, R. E., & Lohman, D. F. (1993). Cognitive psychology, new test design, and new test theory: An introduction. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 1–17). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

Templin, J. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis*. Unpublished doctoral dissertation, University of Illinois at Urbana–Champaign.

White, B., & Frederikson, J. (1987). Qualitative models and intelligent learning environment. In R. Lawler & M. Yazdani (Eds.), *AI and education* (pp. 281–305). Norwood, NJ: Ablex.

Yang, X. (2003). *Inferring diagnostic information from abstract reasoning test items*. Unpublished doctoral thesis, University of Kansas, Lawrence.