

# Language Teaching Research Quarterly

2023, Vol. 37, 248–265



## Another Reason for CDA in Language Assessment: A Critical Synthesis from a Perspective of Validating Test Constructs

Yeon-Sook Yi

SangMyung University, South Korea

*Received 15 June 2023*

*Accepted 20 October 2023*

### Abstract

Cognitive diagnostic assessment is known for its capacity to provide detailed feedback to test-takers, which can inform instruction and guide learning. Besides this core strength of the assessment approach, the measurement procedure also works as a construct validation method in itself. Previous studies do not call attention to this aspect of cognitive diagnostic assessment and how the CDA validation can be different from existing construct validation methods. Thus, the purpose of this article is to synthesize cognitive diagnostic assessment research from a standpoint of seeing it as a construct validation procedure with a focus on its strengths as such. In doing so, the article closely examines studies that applied two earlier cognitive diagnostic models to learner data because they established the research foundations of CDA in language testing. Before looking into the details of the procedure of cognitive diagnostic assessment, the paper critically reviews non-CDA studies on construct validation in language testing. It then scrutinizes the methods and objects of construct validation in CDA research to find any significant differences. The paper concludes with several avenues of future research in CDA, which are related to the matter of verifying test constructs.

**Keywords:** *Cognitive Diagnostic Assessment, CDA, Validation, Test Construct, Q-matrix Construction*

### Introduction

Cognitive diagnostic assessment (CDA) is a relatively new method of measurement that has been gaining attention in recent years. Its theoretical frameworks have been researched in the measurement field and its feasibility and applications in the real-world testing contexts have been examined in specific knowledge domains and particularly in language testing. The strength of CDA is mainly discussed in terms of its capacity to provide detailed feedback to test-takers, which can inform instruction and guide learning. Besides this core strength of the CDA, the measurement procedure also works as a construct validation method in itself.

\* Corresponding author.

E-mail address: uiuc99110@hotmail.com

<https://doi.org/10.32038/ltrq.2023.37.14>

Existing CDA studies do not distinctly call attention to this aspect of CDA, that it can be used as a strong validation tool and, more importantly, how the validation of CDA can be different from other construct validation methods. The purpose of this article is to synthesize CDA research from a standpoint of seeing it as a construct validation method with a focus on its strengths as such, relative to other validation methods. Before looking into the details of the CDA method, the paper first critically reviews (non-CDA) studies on construct validation in language testing in terms of their targets or objects (namely, the internal structure of a test, task characteristics on item difficulty, test-taker characteristics, their strategy use, or test-taker responses) of the analysis and technical methods. The paper then scrutinizes the methods and objects of construct validation in CDA research to find any significant differences in these aspects between the CDA studies and other construct validation research in the field of language assessment.

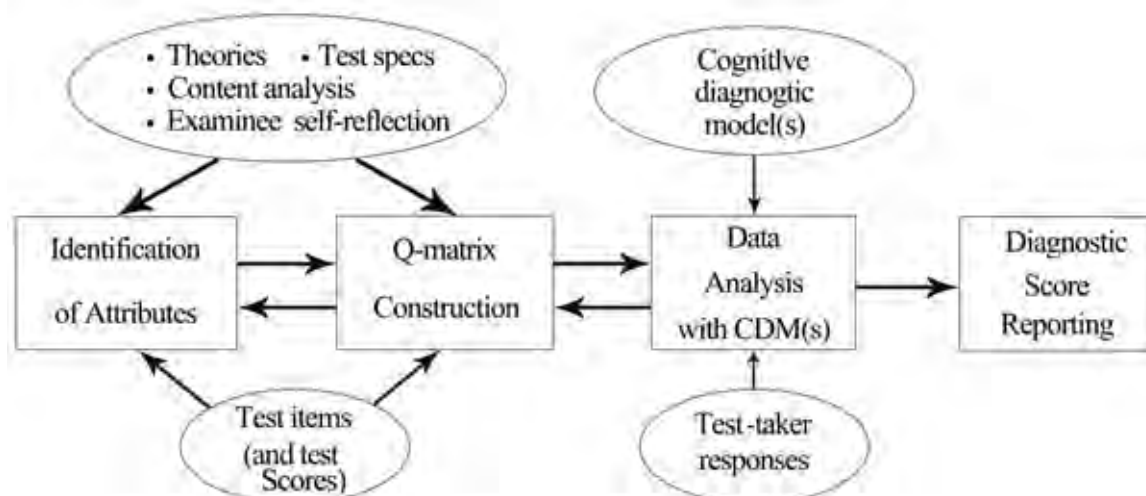
### What is Cognitive Diagnostic Assessment?

In psychometric measurement field, item response theory overcame limitations of classical test theory and was successfully applied for the unidimensional, continuous scaling of test-takers in major subject areas. Though such psychometric modeling is useful and psychometrically dependable for summative assessment, it could not resolve one essential issue in educational assessment. For a few decades, it has been suggested that tests aim at formative assessment where the test results are directly used to inform teaching and learning (Bejar, 1984). The single score-based testing paradigm has been challenged for more in-depth descriptions of student achievement which could provide diagnostic information. To address such challenges, measurement researchers have tried to integrate cognitive psychology and psychometric test theory to devise a test procedure that provides more fine-grained diagnostic information about learners' levels of mastery of cognitive attributes. Such a test procedure will help learners remedy the deficiencies in the skills that they have not acquired sufficiently.

As Figure 1 indicates, among the four steps of CDA, the first three steps are conducted in an iterative and exploratory manner in order to determine a final Q-matrix.

**Figure 1**

*Steps of the CDA Procedure*



\* Note. The first three steps are involved in identifying and validating constructs of a test.

While this measurement method focuses on providing learners' strengths and weaknesses, called mastery profiles, it also provides psychometric properties of test items with corresponding statistical parameter estimates, which is how the third step of the CDA procedure can be utilized as a quantitative tool of construct validation.

One important assumption in discussing cognitive diagnostic modeling, as contrasted with more traditional IRT modeling, is its multidimensionality assumption. Despite some exceptions, the knowledge structure of most cognitive diagnostic models is explicitly multidimensional, as multidimensional cognitive diagnosis models should be necessary in order to effectively assess skill mastery levels, even in situations where students with the same average global ability vary significantly on their mastery levels of the individual skills (DiBello, et al., 2007). This feature of CDM (Cognitive Diagnostic Model) seems particularly relevant to language assessment contexts, if considering the multidimensionality debate about language test constructs which will be examined in the section, 'Construct validation research in language testing.'

### **A Brief Overview of Validation Research in Language Testing**

Before looking at previous studies on construct validation in language testing, a brief look will be taken at validation research in general in language testing, focusing on its analytic methods and techniques. This will enable readers to see construct validation in a larger picture of general validation research and realize the importance of construct validation research and methodological differences of construct validation studies relative to other types of validation research in language assessment.

In examining how the field of language assessment has evolved in its approach to validation since Kunnan (1998) and Hamp-Lyons and Lynch (1998),<sup>1</sup> particularly in reference to methodological approaches, Choi and Schmidgall's study (2011) provides a useful summary. In a review of empirical validation research, they examined 169 studies published between 1999 and 2009 in the journals of *Language Testing*, *Language Assessment Quarterly*, research reports of TOEFL and IELTS and relevant doctoral theses, in terms of types of validity, methods, and objects of validation. Table 1 summarizes their findings.

Looking into some of these research techniques, those listed for quantitative methodologies are correlational analysis, group comparisons, factor analytic techniques, facet analysis, agreement indices, G-theory, chi-square and Rasch modeling. On the other hand, those labeled as qualitative methodologies are questionnaire, interview, content analysis, verbal protocol, discourse analysis, observation, conversation analysis, document analysis, expert review, stimulated recall, etc.

---

<sup>1</sup> These two previous studies provide a review of validation studies in language assessment. Kunnan (1998) reviewed empirical studies in language testing, using Messick's framework of validity to categorize 16 years of research (1980-1996). Hamp-Lyons and Lynch (1998) performed a similar analysis focusing on 16 years of LTRC abstracts.

**Table 1**

*Percentage of Research Techniques Used in Language Testing Validation Research (1999-2009)*

Quantitative methodologies and percentages (55.8%)	Correlational analysis	Group comparisons	Factor analytic techniques	Descriptive Statistics	Facet analysis	Other
	13.4%	12.6%	6.7%	5.9%	3.9%	3.4%
	Agreement indices	G-theory	Chi-square	Rasch models	Discriminant Analysis	
	3.4%	2.8%	2.0%	1.1%	0.6%	
Qualitative methodologies and percentages (44.2%)	Questionnaire	Interview	Content analysis	Verbal Protocol	Discourse analysis	Other
	13.7%	7.3%	6.7%	5.6%	4.2%	0.6%
	Observation	Conversation analysis	Document analysis	Expert review	Stimulated recall	
	2.2%	1.7%	0.8%	0.8%	0.6%	

\* Note. The percentages in the table add up to the percentage noted for each category, i.e., 55.8 % for quantitative and 44.2 % for qualitative methodologies (Choi & Schmidgall, 2011).

Of these methods, those that are frequently used in construct validation studies are correlational analysis (including regression analysis), variations of the factor analytic technique, interview, content analysis, verbal protocol and expert review. According to Choi and Schmidgall (2011), it is the quantitative approach that is dominant (79.67% versus 20.33% for qualitative) in construct validation research, which shows a higher proportion of construct validation research employs quantitative techniques than other types of validation research in language testing. Narrowing our focus down to construct validation studies in the next section, the paper will critically examine the methods of identifying constructs or specific attributes of a language test, in order to see them in comparison with the construct validation methods in the CDA procedure.

### **Construct Validation Research in Language Testing**

Construct validity is defined as the experimental demonstration that a test is measuring the construct it claims to be measuring. Messick (1989) presented a unitary concept of validity which refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. In his unified validity framework, construct validity is defined as the theoretical context of implied relationships to other constructs, which should contribute to the evidential basis for interpreting tests and test scores. In literature, different terms have been used to refer to dimensions of cognitive construct, such as attributes, skills, factors, traits, abilities and subskills. Though slightly differently defined, constructs, attributes or skills are mostly used interchangeably in the literature and in this paper (see Buck & Tatsuoka, 1998; Rupp, et al., 2010; and Li, 2011 for detailed definitions of some of these terms). Prior to discussing methods and objects of construct validation research as well as relevant critical comments on them, a more fundamental issue of determining dimensionality of language ability will be briefly examined in the next subsection.

*Theoretical Basis for Multi-dimensionality: Findings of Applied Linguists*

Research in language testing has tried to explain the relationships between the test tasks and the specific abilities or attributes that they assess. However, unlike other domains of knowledge, a very fundamental line of debate had to be resolved before determining the identity of constructs: that is, whether language ability is, without a doubt, multidimensional or a divisible skill. What it means to be able to use a language is one of the longest-debated and still on-going issues in applied and theoretical linguistics (Spolsky, 1985). Over the years, different models and theories have been proposed to account for the nature of language ability, ranging from multidimensional models at one end to unidimensional model at the other and some moderate models in between.

The early structuralist claim of language ability which focused on structural descriptions of the language and its skills and components model (Carroll, 1961; Lado, 1961) were questioned by the notion of an underlying competence that was thought to govern other subcomponents of the language. Oller (1979), inspired by Spolsky's (1973) concept of overall language proficiency, proposed that a single general language proficiency factor (general factor) accounted for the performance on a variety of language tests. However, the strong version of his unitary competence hypothesis has been criticized for its methodological and theoretical drawbacks (Bachman & Palmer, 1981, 1982; Kunnan, 1995). Subsequently, the unitary competence assumption was replaced by the concept of communicative competence (Hymes, 1972) and multidimensional or multifaceted models of communicative competence (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980; Celce-Murcia, et al., 1995). Research on this issue thereafter in language testing almost consistently demonstrated that second language ability is composed of multiple factors underlying language proficiency (Choi & Bachman, 1992; Vollmer & Sang, 1983).

One interesting point of view with regard to dimensionality issue in language testing is found in Henning (1992). Distinguishing between psychometric and psychological dimensionality, using principal components analyses and internal consistency reliability estimates, he demonstrated that even when the psychological reality underlying a dataset was multidimensional in nature, psychometric dimensionality could still be unidimensional and vice versa, depending on the distributions of item difficulties and person abilities. He further claimed that psychometric unidimensionality was like internal consistency reliability, existent by a matter of degree, not as a categorical phenomenon. As he mentions in the paper, his insight into this distinction may explain why some researchers (e.g., Davidson, 1988) could not find psychometric multidimensionality, when psychological multidimensionality was apparent. Though Henning (1992) claims that judicious use of probabilistic methods of analysis may still prove informative in the process of construct validation, his results seem to be indicative of the inadequacy of certain statistical methods for the identification and validation of psychological constructs underlying test performance, supporting the difficulty of detecting dimensions in language test data due to the complexity of language abilities and processes.

*Methods and Objects of Construct Validation Research and Critical Views*

With the above foundations on multiple factors of language ability as a theoretical backdrop, there has been a body of research attempting to identify the constructs in language tests. Language testing researchers have used various approaches to identifying specific attributes

that L2 language tests were supposed to measure. Sawaki, Kim, and Gentile (2009) sum up these methods under three categories for L2 reading and listening tests and this section expands the scope to include the methodologies for speaking and writing assessments.

First, researchers paid attention to what learners reported in introspective or retrospective studies of test-taking skills and processes. Studies that used verbal protocols or interviews for receptive L2 tests (Anderson, et al., 1991; Barta, 2010; Buck, 1991, 1994; Cohen, 1984; Cohen & Upton, 2007; Cordon & Day, 1996; Farr, et al., 1990; Jang, 2005; Kasai, 1997; Nevo, 1989; Rupp, et al., 2006; Scott, 1998) and those that employed questionnaires and interviews for L2 speaking and writing (Brown, 1993; Swain, 2001) are included in this category. Secondly, an approach focusing on surface task characteristics was used in item difficulty modeling studies. As in the first category, most studies investigated L2 comprehension tests (Bachman, et al., 1995; Carr, 2006; Clapham, 1996; Freedle & Kostin, 1993, 1999; Nissan, et al., 1996; Song, 2008; Wagner, 2004), while a few studies examined both productive and receptive skills (Bachman, et al., 1996; Bachman & Palmer, 1981, 1982). This approach typically employed statistical and/or psychometrical analysis techniques. Thirdly, other approaches were based on subskill theories, in which target constructs of interest in language tests were identified according to theoretical taxonomies of language ability or experts' opinion (Alderson, 1990a, 1990b; Alderson & Lukmani, 1989; He & Dai, 2006; Lumley, 1993).

Among these approaches, the first and third methods can be approximately categorized as qualitative while the second, as quantitative, which means mainly resorting to statistical analyses. As mentioned earlier, Choi and Schmidgall (2011) finds that it is the quantitative approach that is dominant (79.67% versus 20.33% for qualitative) in construct validation research except in validating the constructs of performance assessments (for which 66.67% of research relied on qualitative methods, while 33.33% used quantitative counterparts). Overall, most studies rather counted on one methodology, whether it was quantitative or qualitative.

Though these different methodologies can work as a supplementary tool for each other if employed in the same study, each of these methods was subject to criticism when used separately. First, weaknesses were pointed out in consulting human minds for disclosing their cognitive processes. In using verbal reports and protocols that rely on examinees' self-reflection about their mental response processes, it has been shown that different test-takers may answer the same test items in different ways. Also, as von Schrader (2006) points out, the process of assembling expert groups and/or analyzing a think-aloud study is costly and time-consuming, which is likely to become a larger problem in language testing, as the complexity of language processing would require diverse sources in revealing mental processes.

Critiques on statistical methods seem particularly important, since they were employed in nearly 80% of construct validation research according to Choi and Schmidgall (2011). The two most commonly used techniques of factor analysis and regression analysis have been under some criticism (Adams, et al., 1987) and the tenability of these analyses was seen with skepticism (Buck, et al., 1997; Buck et al., 1998). Listing three reasons for the inadequacy of multiple regression (i.e., too many predictors and too few cases; putting emphasis on item characteristics rather than performance ability; and only providing information about group performance), Buck et al. (1997) make it clear that alternative methods are also to be used for

attribute extraction and their validation, which was one of the reasons they adopted the Rule Space Methodology (RSM) in their research.<sup>2</sup>

Related to the problem of not being able to detect psychometric multidimensionality when psychological multidimensionality is apparent (Henning, 1992), which may be indicative of the deficiency of factor analytic techniques (including principal component analysis Henning used for demonstration) for identifying and validating psychological test constructs, more recent studies particularly focus on the use of factor analysis in comparison to using a cognitive diagnostic model (Aryadoust & Goh, 2013; Kunina-Habenicht, et al., 2009; Wang, 2009). Of these studies, Aryadoust and Goh's study (2013) is relevant to language testing, which applied confirmatory factor analysis (CFA) and a cognitive diagnostic model (CDM) to second language listening test data and found that CFA imposed more restrictions on the data than a CDM. They further suggested that CFA might not be suitable for modelling dichotomously scored data of L2 listening tests, whereas the CDM used in the study (the Fusion Model) appeared to successfully portray the listening sub-skills of the test.

The next section will give readers a comparative view that will help find differences between the previous studies and the CDA approach, and see how CDA combines both quantitative and qualitative techniques in a single validation study undergoing multiple phases of validation.

### **Construct Validation in CDA**

#### *Methods and Objects of Construct Validation in CDA Studies*

The CDA approaches and objects of construct validation will be illustrated in a review of several CDA studies in language testing. In CDA, defining and validating attributes is conducted through proceduralized steps of the method, namely, 1) attribute identification, 2) Q-matrix construction and 3) checking the attributes with test-taker responses. This series of attribute specification and validation constitutes the CDA construct validation methodology. Since earlier models of CDA contributed to establishing this procedure in the area of language testing, this paper focuses on these earlier studies that pioneered cognitive diagnostic modeling for language assessment.

#### *The First Step: Identifying Attributes*

##### *Consulting Relevant Literature and Expert Judgment: Pioneering Work of Buck and Co-authors*

For applying Rule Space Methodology (RSM) to a reading comprehension test, an initial list of attributes was drafted based on the research literature, linguistic theory, teaching experience, test development practice, and self-observations of task-completion strategies in Buck et al. (1997). In Buck and Tatsuoka (1998), the authors also began with an initial list of candidate attributes that came from two main sources of literature: the work of Freedle and Kostin (1996) and of Buck (1990). In verifying the attributes, they resorted to mixed methods: while using both criteria of test-taker classification rates and multiple regression between item difficulty

---

<sup>2</sup> In the RSM applications, however, given the novelty of the method (as the RSM was one of the earliest CDMs), the authors used multiple regression for the purpose of cross-validation, while searching for alternative methodologies for validating RSM analysis results.

parameters and coded attributes, they also exerted human judgment, which kept attributes that had theoretical interest or diagnostic value, and tried to keep a balance of abilities in the whole set. The validating process was also iterative, utilizing both expert decisions and statistical measures repeatedly.

One important issue that came up for repeated consideration in these studies was about at what grain level the skills should be defined. When defining the initial attribute list, the authors in both studies pointed out the importance of considering the interpretability of the attributes as cognitive abilities. They highlighted two general types of attributes that could be defined: abstract (or higher level) skills or more ‘nuts and bolts’ item characteristics. Though they opted for more detailed item characteristics to maximize predictive power for student performance, they accepted that less detailed and more theoretical attributes would be easier to interpret.

#### *Utilizing Learner Perspective and Expert Judgment: Studies of Kasai and Scott*

In order to define attributes for the reading comprehension of TOEFL in the application of RSM, Kasai (1997) and Scott (1998) worked together to hypothesize a global sequence of processes of answering each reading comprehension item, based on Kirsch and Mosenthal’s model (Kirsch & Mosenthal, 1990) and their previous analysis of verbal protocols obtained from five university students while solving practice TOEFL questions. Based on the hypothesized mental processes, they selected attributes, considering the characteristics of correct options and distractors as well as of passages. Kasai himself also answered all the questions in the test and created a tentative set of attributes. Once an initial set of attributes was determined, it was reviewed by a few doctoral students who had strong knowledge of second language reading and experience of teaching reading comprehension. Kasai and Scott assessed the feasibility of each attribute and refined the definition, which led to the final set of attributes that would be used in further analyses. The number of students involved in the protocol study seems controversial, however, considering later studies (Jang, 2005; Li, 2011) hired more than 10 students for the same procedure. Otherwise, it seems they used human resources wisely, consulting experts’ opinions and students’ verbal reports in initially specifying their attributes.

#### *Employing Comprehensive Sources: Studies of Jang and Li*

In applying the NC-RUM (Fusion) model to obtain examinees’ skill level information on the LanguEdge TOEFL reading test, Jang (2005) elaborately identified the test attributes using multiple sources of information. She drew from literature related to both first and second/foreign language acquisition in order to understand the process of reading. Doing an extensive task analysis examining the content and characteristics of each item and textual variables in conjunction with test items, she also used the test developer contents codes (test specifications) provided by ETS and working frameworks in the TOEFL 2000 monograph series for a better understanding of the skills intended to be measured by the test developers. She further explored the skills by asking examinees to think aloud while taking the test. Besides adopting these varied qualitative methods, she performed dimensionality analyses and collected information about item clustering using such procedures as DETECT, DIMTEST and HCA/CCPROX to provide an in-depth description of the multidimensional latent structure of a test. Based on the data collected, Jang defined nine primary reading skills measured by the 37 items on the test.



In Li's research (2011) which used the Fusion model with the MELAB (Michigan English Language Assessment Battery) data, an initial cognitive framework was proposed based on second language reading theories and related literature. In particular, she used Gao's model of cognitive processes based on verbal reports from Chinese ESL students and content experts to inform the construct validation of the MELAB reading (Gao, 2006). Since both MELAB and TOEFL have very similar content areas and cognitive structures, she also referred to the taxonomies for TOEFL reading used in cognitive diagnostic analyses.

#### *Conducting Factor Analysis Based on Existing Linguistic Research*

In addition to the approaches discussed thus far, a statistical analysis was also used to verify the test constructs proposed in the literature. As seen earlier in the paper, factor analysis has been frequently used to detect cognitive structures of language tests. Based on the prior literature suggesting that the three dimensions of morphosyntactic form, cohesive form, and lexical form were measured by the grammar section of the Certificate of Proficiency in English (ECPE) test (Liao, 2007), Henson and Templin (2007) used a three-factor exploratory model to identify basic clusters of the items that might measure similar abilities. Using factor analysis to initially confirm test constructs in the CDA procedure is interesting, because many studies adopted factor analysis for the purpose of construct validation and for most of them it was the main instrument that determined the number and identity of test constructs. However, the skills defined at this phase will continue to be examined in the steps that will ensue, which provides a clear contrast with quantitatively oriented, previous studies discussed earlier.

#### *The Second Step: Constructing Q-matrix and Checking Reliability*

This section will address the actual process of creating a table called a Q-matrix, using the attributes defined in the ways examined thus far. The work of developing a Q-matrix does not only entail numeric weighting procedure but also validating it using both qualitative and quantitative measures in order to arrive at a finalized Q-matrix. Here, it will be examined how multiple raters reached one single set of Q-matrix and how the researcher(s) checked the inter-rater reliability.

The study of Buck et al. (1997) followed the usual paths of reaching a consensus that are shown in RSM applications: After refining the initial attributes using classification rates and multiple regression, two raters created a Q-matrix independently. The inter-rater reliability between the two coders was checked with a Pearson correlation coefficient in the case of the six continuous attributes, and the percentage agreement between coders in the case of the remaining 20 dichotomous attributes. Throughout the remaining process until the end of the RSM application, they resolved their discrepancies by discussion.

In Buck and Tatsuoka's study (1998), the authors went through a number of steps, before applying a CDM, in reducing the number of attributes in their draft Q-matrix from 71 attributes to fewer than 20. Through visual inspection, they removed attributes that did not occur in enough items (as a minimum of three occurrences is likely to produce more stable results), attributes attempting to get at the same idea, or an attribute that appeared to be unrelated to performance. They also deleted attributes with the lowest correlations with item difficulty, as well as attributes that had high correlations with others. They further used multiple regression,

predicting item difficulty as the dependent variable, and eventually agreed on a set of 17 attributes.

Kasai's study (1997) examined more advanced indices of statistics and measurement theories. To determine how consistent each rater's coding was, Kasai computed the Rater Agreement Proportion (RAP) statistic (Bachman et al., 1991) for each item. Generalizability theory was also employed: The generalizability coefficient for the relative decision was computed to determine the extent to which coders agreed with their ratings.

In Jang's research (2005), five raters reviewed the LanguEdge test items and selected skills from the list of skills provided with brief descriptions. Inter-rater reliability was assessed using Fleiss' Kappa statistic for categorical ratings of more than two raters. An individual Kappa value was calculated for each skill, and all the values were averaged into an overall coefficient. Considering that Kappa values are sensitive to the number of rating categories and items, they were used to examine relative agreement rates among the nine skill categories rather than absolute agreement rates.

In building a Q-matrix, Li (2011) invited four experts (doctoral students with experience of teaching ESL reading) to identify the reading skills required by each item. When they completed rating each passage, Spearman's rho was calculated to indicate the agreement between the ratings given by each expert. The values of Spearman's rho were all higher than 0.30, indicating moderate agreement. In creating a Q-matrix, the evidence from the think-aloud verbal reports was also considered.

### *The Third Step: Validating with CDM Analysis of Response Data*

A Q-matrix determined thus far goes in the CDM analysis as input along with test-taker responses. This process is usually iterative: repeating the cycle of fitting a CDM, evaluating item parameters and refining the entries of the Q-matrix. Though this step consults quantitative indices, human expertise and judgment also play a crucial role in determining the final Q-matrix. As the details of the process differ for each different CDM, this section will examine CDA studies according to the specific CDM, focusing on validation processes only unique to each CDM.

### *Rule-Space Model (RSM)*

In applying the RSM, which is one of the earliest CDMs and has been employed in a number of language testing studies, the adequacy of the attributes in the Q-matrix was judged by looking at the percentage of examinees successfully classified into one of the mastery patterns or knowledge states. In the studies of Buck et al. (1997) and Buck and Tatsuoka (1998), they found 96% and 91% of the test-takers, respectively, were classified by the RSM. To cross-validate whether these scores did explain performance on the test, each student's total score was regressed on the attribute mastery pattern to see how well attribute mastery explained performance. The resulting adjusted R-square was 0.79 and 0.77, in the respective study, indicating that for the classified test-takers, the attribute mastery scores (1 for mastery, 0 for non-mastery) explained 79% and 77% of the variance in total scores. Running the RSM four times, Buck et al. (1997) refined the Q-matrix, basically keeping the ones with good statistical properties and deleted those with bad. Following a heuristic strategy and balancing a few conflicting factors, they kept attributes that had theoretical interest or diagnostic value for

teaching purposes. Looking at the attributes as a whole set and trying to keep a balance of abilities, they tried not to delete similar ones altogether, even though the remaining ones did not have very good statistics. Buck and Tatsuoka (1998) examined the responses of the 38 unclassified examinees (9% of the test-takers) to find the attributes that they mastered but that were not included in the attribute list. Consequently, they added another attribute and also deleted three other attributes on account of redundancy and inappropriacy. This led to a reduced list of 15 attributes and the classification rate improved to 96% as a result. The adjusted R-square was also raised to 0.97 and 0.94, for respective studies. This entire procedure of attribute refinement and Q-matrix validation was identical, in principle, in other RSM studies (e.g., Kasai, 1997; Scott, 1998).

#### *Reduced NC-RUM (Fusion Model)*

Fusion model is a relatively new model (Hartz, 2002) but it has been applied more than any other CDM in the arena of language testing in the past several years. With a more refined modeling structure, it is claimed to have important advantages over other CDMs, such as the ability to compensate for the incompleteness of the Q-matrix and the capacity to evaluate the diagnostic capacity of the test items (Li, 2011), which will be very informative in this phase of validating test constructs.

In the studies that employed this CDM (e.g., Jang, 2005), multiple indices were systematically investigated to assess the initial Q-matrix. The first phase in this step started with checking the MCMC (Markov chain Monte Carlo) convergence indicating the statistical estimation was reliable. After achieving the statistical convergence, most Q-matrix entries that were practically insignificant were eliminated, based on the parameter estimates. The item parameters useful for evaluating the item coding are  $\pi_i^*$  and  $r_{ik}^*$ . Particularly,  $\pi_i^*$  values indicate the probability of correctly responding to an item after an examinee has mastered all required attributes for that item. Thus, a low  $\pi_i^*$  value may suggest extreme item difficulty levels or incompleteness of the Q-matrix. The values of another parameter,  $r_{ik}^*$  can be used to interpret the degree to which the mastery of a skill influences the successful performance on the item. If its value is larger than 0.9, it indicates that the mastery of the particular skill has little discriminating power or influence on successful performance on the item. Using these item parameter values, the coding in the Q-matrix was refined by removing some entries with high  $r^*$  values and low  $\pi^*$  values. Because the presence or absence of one skill element affects the ways other skills influence the item response function, they removed only one entry at a time from the Q-matrix following step-wise reduction algorithm.

Researchers also exerted substantive judgment as an expert revisiting various sources before finally revising a Q-matrix. Checking relevant item parameters, Li (2011) found that  $r_{15.5}$ , i.e., the discrimination capacity of item 15 to skill 5 was bigger than 0.9 (0.913). Upon closer examination of the item and also the think-aloud verbal reports, the item was found incorrectly assigned to skill 5 (making inferences), despite the use of the word ‘infer’ in the item stem. Thus, item 15 was reclassified as requiring another skill (synthesizing and connecting). Li subsequently deleted other Q-matrix entries based on both statistical criteria and substantive knowledge.

This step of construct validation using the analysis of test-taker responses primarily resorted to quantitative indices generated from running a specific CDM. Researchers, however,

tried to substantively interpret the numerical evidence. Since determining Q-matrix entries and refining them requires in-depth theoretical and substantive understandings of the processes and skills associated with the items, the substantive soundness of the Q-matrix and justifiable interpretation of the estimation results were given more weight than statistical indices in these CDA studies.

### *Significances of CDA Validation*

As Brown (2000) mentions in his explication of construct validity, the more methods are used to prove the validity of a test, the more assurance test users have in the construct validity of the test. Even though this statement is not explicitly aimed at CDA, the procedure of CDA realizes this important principle of construct validation. This is just what CDA does. It goes through systematically integrated, multiple steps of construct verification, which is one major reason why CDA procedure can be more dependable as a construct validation tool. In comparison to the construct identification and validation studies that used other methodologies than the CDA procedure, the unique features that could constitute the significance of CDA research can be summarized as the following.

First, under the CDA framework, both qualitative and quantitative methods are actively sought and utilized for extracting attributes and creating a Q-matrix. The process is iterative until the researchers feel that the set of attributes sufficiently explains examinee performance. By resorting to both types of analysis techniques, CDA has the potential to compensate for the limitations of each side of the quantitative or qualitative method for establishing multiple dimensions of test constructs. The more systematic, multi-layered approach of CDA to uncovering attributes and building a Q-matrix can be deemed as a synergistic merger between both methods and it is the significance of CDA methodology in language testing.

Secondly, CDA procedure can be more content-based than other item analysis methods that employ test-taker response data. Davidson (2010) puts it as follows:

CDA is a procedure, and it does one very important thing that normative item analysis has long overlooked: It values the content of a test task. Even more, it asks that test developers portray that content in a well-reasoned and conscientious manner. ... CDA gives us a way to articulate rich discussions about test content *in a procedural manner*. ... Test content has never really been part of the response data procedure, until we had CDA, and that is why CDA is so important (p.106, italics added by the author).

The last part of the above quotation refers to one of the unique features of CDA that the analysis results of response data (whether they are unidimensional total scores or multidimensional attribute scores calibrated by a CDM) direct the researchers' attention back to the individual test items and make them scrutinize the items to possibly moderate the list of constructs for more justifiable interpretation. This process is incorporated in the steps of the CDA procedure that the researchers must go through, hence, the phrase 'in a procedural manner.' The process is, as noted repeatedly, iterative between the steps of the procedure.

Lastly, humans play a more active role in CDA. It is human judgment that has the ultimate control over what kinds of attributes should be defined and later deleted or retained in the test constructs. This is how CDA can lessen the chance of *statistical determinism* by using

statistical information like lamp-posts (that is, for support rather than illumination) in the matter of attribute definition and Q-matrix construction. In other words, quantitative methods of examining statistical parameter estimates that a CDM generates basically serve as confirmatory or supplementary tools of various qualitative attribute defining methods, such as literature review, experts' task analysis, examinees' verbal reports and Delphi method for multiple Q-matrix coders to reach a consensus among them. While benefiting from a powerful and advanced psychometric tool, CDA makes it clear that only reliable human knowledge and well-grounded human decision can bring about a harmonious play of multiple sources of attribute identification. Human resources are also more actively utilized in CDA by fully utilizing both types of human resources, not just focusing on either content experts or test-takers, especially during the initial stage of identifying attributes.

### **Conclusion and Suggestions for Future Directions**

This article has reviewed a relatively new test method procedure of CDA from a different angle. CDA has its essential capacity to yield a highly articulated score report and better link tests to performance standards. Another strength of the assessment method is that it also provides a rigorous procedure for validating test constructs, which is the focus of the present paper. Due to these strengths of the method, CDA can 'survive and flourish' (Davidson, 2010) in the measurement field and in language testing, in particular. However, there certainly are issues to be resolved and areas to do more research on implementing CDA. Among these issues, this paper will address a few of them, which are related to the matter of verifying test constructs.

First, logistical challenges can occur in consulting human minds, i.e., experts' knowledge, as was briefly touched upon earlier, though the problem is not unique to CDA. One way to possibly mitigate this potential problem in identifying specific test attributes is having a kind of inventory of skills. The rationale is simple: Testing experts can draw from the existing skill inventory to get a clue about what initial attributes they need to set up for CDA. They will have to, of course, undergo the entire process of skill definition and modification but some sorts of existing skills could definitely make initial work of CDA application easier, as Li (2011) benefited from Jang's work (2005).


Determining the degree of abstractness or concreteness of attributes in adopting CDA can be another important avenue for future research. If attributes are more abstract and theoretical, they are easier to interpret but it might become harder to predict student performance. On the contrary, if attributes are more concrete and detailed, the interpretation of student performance on the general level might become more difficult, but the predictive power of learner performance can be maximized. Deciding on which type of attributes should be identified seems to depend on the utility of the constructs in actual teaching and learning. Thus, it would be useful to experiment with both types of attributes in actual classroom teaching and find out which type turns out to be more conducive to teaching and learning by providing better interpretability or predictability of performance. Extracting constructs at a higher and abstract level or lower and detailed level is also contingent on the facility of obtaining justifiable statistical information, which tells us that considering all relevant factors in resolving this matter is indeed important.

Third, a more interesting and significant line of research will be about using the rigorous content-based approach of CDA in the actual test development process. Most of the prior CDA

applications in language assessment have retrofitted CDMs to existing tests that were not developed to provide diagnostic feedback (Sessoms & Henson, 2018). In this context, the steps of defining and validating test constructs in CDA can be utilized as a content-based *test construction* tool before any response data become available (Ranjbaran & Alavi, 2017). A test developed using this method will be implemented more easily for the purpose of diagnosis, because many complexities and challenges of the iterative processes among the first three steps of CDA could be alleviated in that case.

CDA as a means of construct validation has been discussed thus far focusing on the studies that applied two classification models, RSM and Fusion model. These studies were closely examined because they established the research foundations of CDA in language testing. While these CDMs, particularly RUM model, are still actively utilized for language assessment (Dong et al., 2021; Ranjbaran & Alavi, 2017; Shahsavar, 2019), several recent studies are paying attention to the utility of generalized models such as G-DINA. Li, et al. (2015), Chen and Chen (2016) and Javidanmehr and Arani Sarab (2019) retrofitted reading comprehension assessment under G-DINA model framework and found the fit and capability of the model for reading comprehension attributes. More recently, Tonekaboni, et al. (2021) documented the interaction of attributes to validate the Q-matrix for a reading test within G-DINA model framework. Ma and Meng's study (2014) is noteworthy in that it does not retrofit but constructs an EFL listening diagnostic test. They could further validate G-DINA analysis of the listening comprehension data. While most CDA studies in language testing examined English learners' data, the study of Li, et al. (2021) investigated Chinese learners' listening comprehension ability based on G-DINA model. A more interesting endeavor is found in the study of Effatpanah, et al. (2019), which used G-DINA model for validating their Q-matrices, though they applied ACDM for their listening test data analysis. All these recent studies tell us that cognitive diagnostic models are continuously evolving and so does test validation via CDA models. Thus, in the future research, it would be worth investigating how using these new models plays a role in validating the test constructs, as we have seen it with earlier cognitive diagnostic models in this article.

## **ORCID**

 <https://orcid.org/0000-0001-6771-9771>

## **Acknowledgements**

Not applicable.

## **Funding**

Not applicable.

## **Ethics Declarations**

## **Competing Interests**

No, there are no conflicting interests.

## **Rights and Permissions**

## **Open Access**

This article is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which grants permission to use, share, adapt, distribute and reproduce in any medium or format

provided that proper credit is given to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if any changes were made.

## References

- Adams, R. J., Griffin, P. E., & Martin, L. (1987). A latent trait method for measuring a dimension in second language proficiency. *Language Testing*, 4(1), 9-27. <https://doi.org/10.1177/026553228700400102>
- Alderson, J. C. (1990a). Testing reading comprehension skills (Part One). *Reading in a Foreign Language*, 6(2), 425-438.
- Alderson, J. C. (1990b). Testing reading comprehension skills (Part Two). *Reading in a Foreign Language*, 7(1), 465-503.
- Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, 5(2), 253-270.
- Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, 8(1), 41-66. <https://doi.org/10.1177/026553229100800104>
- Aryadoust, V., & Goh, C. (2013). Exploring the relative merits of cognitive diagnostic models and confirmatory factor analysis for assessing listening comprehension. In E. D. Galaczi & C. J. Weir (Eds.), *Studies in Language Testing volume of Proceedings from the ALTE Krakow Conference, 2011* (pp. 405-426). University of Cambridge ESOL Examinations and Cambridge University Press.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bachman, L. F., Davidson, F., & Foulkes, J. (1990). A comparison of the abilities measured by the Cambridge and Educational Testing Service EFL Test Batteries. *Issues in Applied Linguistics*, 1, 30-55. <https://doi.org/10.5070/L411004989>
- Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing*, 13(2), 125-150. <https://doi.org/10.1177/026553229601300201>
- Bachman, L. F., Davidson, F., Ryan, K., & Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study*. Cambridge University Press.
- Bachman, L. F., & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning*, 31(1), 67-86. <https://doi.org/10.1111/j.1467-1770.1981.tb01373.x>
- Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16(4), 449-465. <https://doi.org/10.2307/3586464>
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Barta, E. (2010). Test-takers' listening comprehension sub-skills and strategies. *Working Papers in Language Pedagogy*, 4, 59-85. Eötvös Loránd University, Hungary.
- Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*, 21(2), 175-189. <https://doi.org/10.1111/j.1745-3984.1984.tb00228.x>
- Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, 10(3), 277-301. <https://doi.org/10.1177/026553229301000305>
- Brown, J. D. (2000). What is construct validity? *Shiken (JALT Testing & Evaluation SIG Newsletter)* 4(2), 8-12.
- Buck, G. (1990). *The testing of second language listening comprehension* [Unpublished doctoral dissertation]. University of Lancaster, England.
- Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, 8(1), 67-91. <https://doi.org/10.1177/026553229100800105>
- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11(2), 145-170. <https://doi.org/10.1177/026553229401100204>
- Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, 15(2), 119-157. <https://doi.org/10.1177/026553229801500201>
- Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, 47, 423-466. <https://doi.org/10.1111/0023-8333.00016>
- Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). *Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I Verbal: Analogy section* (Research Report, RR-98-19). Educational Testing Service.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1-47. <https://doi.org/10.1093/applin/I.1.1>

- Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing*, 23(3), 269-289. <https://doi.org/10.1191/0265532206lt328oa>
- Carroll, J. B. (1961). *Fundamental considerations in testing for English language proficiency of foreign students*. Center for Applied Linguistics.
- Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specification. *Issues in Applied Linguistics*, 6(2), 5–35. <https://doi.org/10.5070/L462005216>
- Chen, H., & Chen, J. (2016). Retrofitting non-cognitive diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly*, 13(3), 218-230. <https://doi.org/10.1080/15434303.2016.1210610>
- Choi, I-C., & Bachman, L. (1992). An investigation into the adequacy of three IRT models for data from two EFL reading tests. *Language Testing*, 9(1), 51-78. <https://doi.org/10.1177/026553229200900105>
- Choi, I. & Schmidgall, J. (2011). *A survey of methodological approaches employed to validate language assessments: 1999 – 2009*. Paper presented at the annual meeting of the Language Testing Research Colloquium, Ann Arbor, MI.
- Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. University of Cambridge Local Examination Syndicate and Cambridge University Press.
- Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing*, 1(1), 70-81. <https://doi.org/10.1177/026553228400100106>
- Cohen, A. D., & Upton, T. A. (2007). ‘I want to go back to the text’: Response strategies on the reading subtest of the new TOEFL. *Language Testing*, 24(2), 209-250. <https://doi.org/10.1177/0265532207076364>
- Cordon, L. A., & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology*, 88(2), 288-295. <https://doi.org/10.1037/0022-0663.88.2.288>
- Davidson, F. G. (1988). *An exploratory modeling of the trait structures of some existing language test datasets* [Unpublished doctoral dissertation]. University of California, Los Angeles, CA.
- Davidson, F. G. (2010). Why is cognitive diagnosis necessary? A reaction. *Language Assessment Quarterly*, 7(1), 104-107. <https://doi.org/10.1080/15434300903426755>
- DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitive diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics*, Vol. 26: Psychometrics (pp. 979 - 1030). Elsevier Science B.V.
- Dong, Y., Ma, X., Wang, C., & Gao, X. (2021). An optimal choice of cognitive diagnostic model for second language listening comprehension test. *Frontiers in Psychology*, 12: 608320. <https://doi.org/10.3389/fpsyg.2021.608320>
- Effatpanah, F., Baghaei, P., & Boori, A.A. (2019). Diagnosing EFL learners’ writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia* 9, 12. <https://doi.org/10.1186/s40468-019-0090-y>
- Farr, R., Robert, P., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*, 27(3), 209-226. <https://www.jstor.org/stable/1434927>
- Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, 10(2), 131–170. <https://doi.org/10.1177/026553229301000203>
- Freedle, R., & Kostin, I. (1996). *The prediction of TOEFL listening comprehension item difficulty for mini-talk passages: Implications for construct validity*. (TOEFL Research Report RR 96-29). Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.1996.tb01707.x>
- Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL’s minitalks. *Language Testing*, 16(1), 2–35. <https://doi.org/10.1177/026553229901600102>
- Gao, L. (2006). Toward a cognitive processing model of MELAB reading test item performance. *Spann Fellow Working Papers in Second or Foreign Language Assessment*, 4, 1–39.
- Hamp-Lyons, L., & Lynch, B. K. (1998). Perspectives on validity: A historical analysis of language testing conference abstracts. In A. J. Kunnan (Ed.), *Validation in language assessment*. Lawrence Erlbaum Associates.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign, Urbana, IL.
- He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing*, 23(3), 370–401. <https://doi.org/10.1191/0265532206lt333oa>
- Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, 9(1), 1-11. <https://doi.org/10.1177/026553229200900102>



- Henson, R., & Templin, J. (2007, April). *Large-scale language assessment using cognitive diagnosis models*. Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, IL.
- Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics*. Penguin.
- Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign, Urbana, IL.
- Javidanmehr, Z., & Arani Sarab, M.R. (2019). Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA model to a high stakes reading comprehension test. *Language Assessment Quarterly*, 16(3), 294-311.
- Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the TOEFL* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign, Urbana, IL.
- Kirsch, I. S., & Mosenthal, P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly*, 25(1), 5-30. <https://doi.org/10.1080/15434303.2019.1654479>
- Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, 35, 64–70. <https://doi.org/10.1016/j.stueduc.2009.10.003>
- Kunnan, A. J. (1995). *Test-taker characteristics and test performance: A structural modeling approach*. Cambridge University Press.
- Kunnan, A. J. (1998). Approaches to validation in language assessment. In A. J. Kunnan (Ed.), *Validation in language assessment*. Lawrence Erlbaum Associates.
- Lado, R. (1961). *Language testing*. McGraw-Hill.
- Li, H. (2011). *Evaluating language group differences in the subskills of reading using a cognitive diagnostic modeling and differential skill functioning approach* [Unpublished doctoral dissertation]. Pennsylvania State University, University Park, PA.
- Li, L., An, Y., Ren, J., & Wei, X. (2021). Research on the cognitive diagnosis of Chinese listening comprehension ability based on the G-DINA model. *Frontiers in Psychology*, 12:714568. <https://doi.org/10.3389/fpsyg.2021.714568>
- Li, H., Hunter, C. V., & Lei, P.-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33(3), 391-409. <https://doi.org/10.1177/0265532215590848>
- Liao, Y. (2007). Investigating the construct validity of the grammar and vocabulary section and the listening section of the ECCE: Lexico-grammatical ability as a predictor of L2 listening ability. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 5, 37–78.
- Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, 10(3), 211-234. <https://doi.org/10.1177/026553229301000302>
- Ma, X., & Meng, Y. (2014). Towards personalized English learning diagnosis: Cognitive diagnostic modelling for EFL listening. *Asian Journal of Education and e-Learning*, 2(5), 336-348. Retrieved from <https://www.ajouronline.com/index.php/AJEEL/article/view/1669>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). Macmillan.
- Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing*, 6(2), 199-215. <https://doi.org/10.1177/026553228900600206>
- Nissan, S., DeVincenzi, F., & Tang, L. K. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension*. (TOEFL Research Report No. 51). Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.1995.tb01671.x>
- Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. Longman.
- Ranjbaran, F., & Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation*, 55, 167–179. <https://doi.org/10.1016/j.stueduc.2017.10.007>
- Rupp, A.A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, 23(4), 441–474. <https://doi.org/10.1191/0265532206lt337oa>
- Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: theory, methods, and applications*. The Guildford Press.
- Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, 6(3), 190–209. <https://doi.org/10.1080/15434300902801917>
- Scott, H. S. (1998). *Cognitive diagnostic perspectives of a second language reading test* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign, Urbana, IL.
- Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, 16(1), 1-17. <https://doi.org/10.1080/15366367.2018.1435104>

- Shahsavari, Z. (2019). Diagnosing English learners' writing skills: A cognitive diagnostic modeling study. *Cogent Education*, 6: 1608007. <https://doi.org/10.1080/2331186X.2019.1608007>
- Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, 25(4), 435-464. <https://doi.org/10.1177/0265532208094272>
- Spolsky, B. (1973). What does it mean to know a language or how to get someone to perform his competence? In J. W. Oller & J. C. Richards (Eds.), *Focus on the learner*. Newbury House Publishers.
- Spolsky, B. (1985). What does it mean to know how to use a language? An essay on the theoretical basis of language testing. *Language Testing*, 2(2), 180-190. <https://doi.org/10.1177/026553228500200206>
- Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, 18(3), 275-302. <https://doi.org/10.1177/026553220101800302>
- Tonekaboni, F.R., Ravand, H., & Rezvani, R. (2021). The construction and validation of a Q-matrix for a high-stakes reading comprehension test: A G-DINA study. *International Journal of Language Testing*, 11(1), 58-87.
- Vollmer, H., & Sang, F. (1983). Competing hypotheses about second language ability: A plea for caution. In J. W. Oller (Ed.), *Issues in language testing*. Newbury House.
- von Schrader, S. (2006). *On the feasibility of applying skills assessment models to achievement test data* [Unpublished doctoral dissertation]. University of Iowa, Iowa City, IA.
- Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 2, 1-25.
- Wang, Y.-C. (2009). *Factor analytic models and cognitive diagnostic models: How comparable are they? A comparison of R-RUM and compensatory MIRT model with respect to cognitive feedback*. [Unpublished doctoral dissertation]. University of North Carolina, Greensboro, NC.