SHIKEN: JALT Testing & Evaluation SIG Newsletter. March 2011. Vol. *15* #1. pp. 2-6.

Cognitive diagnostic assessment as an alternative measurement model

by Vahid Aryadoust (National Institute of Education, Singapore)

Abstract

Cognitive diagnostic assessment (CDA) is a development in psycho-educational measurement which helps assessment researchers examine test takers' mastery of specific sub-skills with greater specificity than other models such as Rasch or item response theory models. This paper discusses the principles of CDA in general and the fusion model (FM) in particular, underscoring their advantages over other models. It concludes by discussing some resources to learn more about CDA.

Keywords: cognitive diagnostic assessment, fusion model, psycho-educational measurement

Cognitive diagnostic assessment (CDA) is a relatively new method in education as well as language assessment to help furnish fine-grained diagnostic information about test takers' degree of mastery of various defined sub-skills (Lee & Sawaki, 2009). Although the development of CDA has largely been motivated by the need for new formative assessment methods, the technique has been recently retrofitted to norm-referenced tests (Jang, 2005, 2008).

Using CDA methods in assessment confers some benefits that other models do not. First and foremost, the majority of item response theory (IRT) and Rasch models assume or require the statistical unidimensionality of data sets as a precondition for item calibration and parameter estimation. Although multidimensional Rasch models lack this requirement, in most other models unidimensionality is considered requisite to locate test takers along a hypothetical continuum. A great asset of CDA is that it does not require unidimensionality. The unidimensionality precondition seems particularly problematic in the language assessment field because most measurement models force the test data to be unidimensionally asymptotic, whereas research shows that language and educational assessment tools typically tap into an array of attributes or sub-skills, each of which could create a statistically separable dimension (Aryadoust, Akbarzadeh, & Akbarzadeh, 2011; Buck, 1994).

"A great asset of CDA is that it does not require unidimensionality. The unidimensionality precondition seems particularly problematic in the language assessment field because most measurement models force the test data to be unidimensionally asymptotic . . ."

Fitting unidimensional models into the data with complicated constituent structures is likely to identify misfitting items or persons, which, irrespective of their content quality, the researcher must then delete, also eliminating useful information about test takers or items.

One application of CDA modeling is to evaluate test takers' mastery of measured sub-skills. For example, in a test of foreign language reading comprehension, test takers might need to understand particular verbs or nouns (sub-skill 1) or make inferences (sub-skill 2). Examinee's mastery in each of these sub-skills is estimable through CDA models. Because they evaluate the influence of multiple test takers' sub-skills on their test performance, CDA models can provide rich feedback on degrees of examinee mastery of each sub-skill and allow for focused learning.

Psychometricians such as Tatsuoka (1983) and Hartz (2002) have developed a number of CDA models, few of which have been applied to language assessment. The objective of this article is to examine the potential that CDA modeling has in language assessment. The article describes the principles of CDA models in general, and the fusion model (FM) in particular. It finishes off by discussing what CDA might imply for foreign language assessment.

An Overview of CDA

As a multidimensional IRT class of models, CDA is a relatively new development in latent trait measurement that uses both mathematical methods of parameter estimation and principles of cognitive psychology to distinguish *masters* from *non-masters* (DiBello, Roussos, & Stout, 2007; Geirl, Cui, & Zhou, 2009). On a language test in which K sub-skills (or attributes) are tested, there are $K \times \alpha_j$ mastery score values (α_1,…., α_k). For example, for $K = 3$, a test taker who receives $\alpha = (1, 0, 1)$ has mastered two sub-skills. There are 2^K possible mastery/non-mastery patterns, called *latent classes* (Gierl, Cui, & Zhou, 2009; Rupp, Templin, & Henson, 2010). Figure 1 displays a few latent classes for a test with three sub-skills.

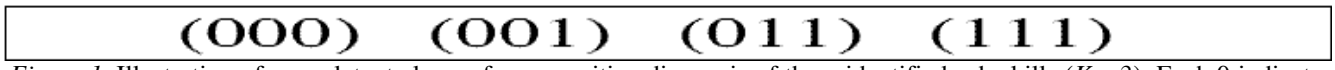

Figure 1. Illustration of some latent classes for a cognitive diagnosis of three identified sub-skills ($K = 3$). Each 0 indicates that the sub-skill is not mastered and therefore not applied to the test item and each 1 indicates that the sub-skill is mastered by the test taker.

Fu and Li (2007) summarized 62 psychometric models used for confirmatory diagnostic purposes. They defined these models as "confirmatory" because the researcher specifies the relationships between manifest and latent variables in a matrix, called Q-matrix. The Q-matrix is constructed by associating test items to hypothesized sub-skills on the basis of an a priori theory (see *The Fusion Model* section below). However, Rupp's (2007) taxonomy of CDA models defines these models more narrowly and includes fewer models. Those include the Deterministic Input, Noisy "And" gate (DINA) model (Junker & Sijtsma, 2001) as well as the FM (Hartz, 2002; Hartz, Roussos, & Stout, 2002).

Another useful classifying criterion proposed by Hartz et al. (2002) is between two general classes: those based on Tatsuoka's (1983) *rule space model* (RSM), which is an ability-based model; and those centered around Fischer's (1977) *linear logistic trait model* (LLTM), which is an item difficulty-based model. All subsequent CDA models have been attempts to expand on these two model classes (Lee & Sawaki, 2009). An RSM model associates test items to the specified cognitive sub-skills "which represent the underlying knowledge and cognitive processing skills that the items assess, and then, based on an test taker's pattern of correct and incorrect responses, infers the probability of each test taker having mastered each sub-skill" (Buck & Tatsuoka, 1998, pp. 126-127). By contrast, LLTM divides unidimensional IRT-based item difficulty parameters into multiple categorical cognitive sub-skills (Leighton & Gierl, 2007).

CDA models have also been classified on the basis of their parameter estimation methods. A number of models (such as DINA and FM) employ Markov Chain Monte Carlo (MCMC) methods based on Bayesian principles. Others use marginal maximum likelihood (MML) estimations. The use of MCMC in parameter estimation approximates distribution features, leaving some residuals, which does not typically occur in MML.

The Fusion Model

The fusion model (FM) is an IRT model developed by Hartz (2002) for CDA purposes. IRT models delineate the probability of test taker j answering item i to be a function of both test taker's ability and item parameters (i.e., difficulty, discrimination, and guessing). As FM specifies, performance on test items is based on test taker's mastery of a set of cognitive sub-skills. Relations between identified sub-skills and test items are specified in a matrix called the Q-matrix (Tatsuoka, 1983). For example, given i test items ($i = 1, 2, 3, 4\ldots, i$) that evaluate k sub-skills ($K = 1, 2,\ldots, k$), the Q-matrix would appear as:

Q = {q_ik}. When sub-skill k is required by item 1, then $q_ik = 1$, and when the sub-skill is not required, then $q_ik = 0$. Table 1 represents a hypothetical Q-matrix:

Table 1. *Illustration of a Q-matrix of four items by three sub-skills*.

Items	Identified sub-skills		
	a	*b*	*c*
1	0	0	1
2	1	1	0
3	1	0	0
4	0	1	1

The items in Table 1 measure three sub-skills. For example, item 1 assesses sub-skill *c* only. For an identified sub-skill to be measured accurately, it should be measured by at least two or three test items (Hartz, Roussos, & Stout, 2002), but each item should seek to test a relatively small number of sub-skills rather than a large array of sub-skills.

FM includes the test taker parameter θ_j, which denotes the overall test taker ability. This parameter is not specified by the Q-matrix, so it is only recognized in the FM, and is expressed as Equation 1:

$$P(X_{ij} = 1|\alpha_j, \theta_j) = \pi^*_i \prod_{k=1}^{K} r^{*(1-\alpha_{ik})\times q_{ik}}_{ik} P_{ci}^{(\theta_j)} \quad (1)$$

where

X_{ij} = response of test taker j to test item i (0 = incorrect; 1 = correct);

α_j = vector of sub-skill mastery (if test taker j has mastered sub-skill k, then $\alpha_{jk} = 1$, and if not, then $\alpha_{jk} = 0$);

θ_j = overall ability of test taker j, which is not specified by the Q-matrix; unlike the ability parameters of IRT, which have continuous θ indices (i.e., data that can hold any value, ranging from minus infinity to infinity), the θ_j index in the FM is a categorical parameter (i.e., data that can only take certain values) (Lee & Sawaki, 2009); $-\infty < \theta_j < +\infty$;

π^*_i = probability of correctly applying the required sub-skills in answering the i^{th} item when the test taker has mastered all relevant sub-skills, or the difficulty of item i according to the Q-matrix; this index ranges from 0 to 1;

$r^*_{ik} = \frac{P(Y_{ij}=1|\alpha_{jk}=0)}{P(Y_{ij}=1|\alpha_{jk}=1)}$, the discrimination parameter of item i and skill k; it ranges from 0 to 1. For each item i, there are k sub-skill values of r^*_{ik} and k_i is the number of sub-skills specified in the Q-matrix as being required to answer item i correctly;

q_{ik} = specification of mastery of sub-skill k which is required to answer item i;

c_i = the degree of reliance of item performance on θ_j in addition to the sub-skills identified in the Q-matrix. This index ranges from 0 to 3.

$P_{ci}^{(\theta_j)}$ = probability of correctly applying the sub-skills which are not specified in the Q-matrix. This index is estimated though the Rasch model.

The sub-skills should be identified on the basis of "test specifications, content domain theories, analysis of item content, think-aloud protocol analysis of test takers' test taking process, and other relevant research results" (Lee & Sawaki, 2009, p. 176). An equally useful method of defining the Q-matrix is iterative runs of the FM to specify the matrix (Sawaki et al., 2009). That is, a panel of experts develops a few rival Q-matrices, whose specified sub-skills will have commonalities and points of

departure depending on their decisions. The researcher should consider multiple factors to partial out rival Q-matrices and retain the best fitting matrix.

Among the aforementioned FM parameters π^*_i , r^*_{ik} and c_i have important roles because they not only provide diagnostic information about each test taker and item, but also highlight the properties of the Q-matrix and the misspecifications observed. An ideal matrix produces estimates r^*_{ik} below .90, indicating that the item discriminates masters from non-masters sufficiently; values below 0.50 are regarded as sub-skills highly necessary to answer the question correctly (Roussos, Xueli, & Stout, 2003). In addition, high π^*_i indices are desirable, as they indicate that masters have a higher probability of successfully applying the sub-skills required by that item. The parameter c_i is a "completeness index" ranging from 0 to 3 (Montero, Monfils, Wang, Yen, & Julian, 2003). It will approach 3 if the sub-skills required to successfully answer the item are fully specified in the Q-matrix, and 0 if they are not specified in the matrix.

Recently, a few informative textbooks on CDA models have been published by educational researchers, including Rupp, Templin, and Henson's (2010) and Tatsuoka's (2009) works.

References

Aryadoust, S. V., Akbarzadeh, S. [Sanaz], & Akbarzadeh, S. [Sara] (2011). Psychometric characteristics of the Persian version of the Multidimensional School Anger Inventory–Revised. *Asia Pacific Journal of Education*, *31*(1), 51-64. doi: 10.1080/02188791.2011.544070.

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11*(2), 145-170. doi: 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. doi: 10.1177/026553229801500201

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* (pp. 45-79). Vol. 26, Psychometrics. Amsterdam: Elsevier Science B.V.

Fischer, G. H. (1977). Linear logistic trait models: Theory and application. In H. Spada & W. F. Kempf (Eds.), *Structural models of thinking and learning* (pp. 203–225). Huber: Bern, Switzerland.

Fu, J., & Li, Y. (2007, April). *Cognitively diagnostic psychometric models: An integrative review*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL. Retrieved from http://www.google.com.sg/url?sa=t&source=web&cd=1&ved=0CBsQFjAA&url=http%3A%2F%2Fwww.stat.cmu.edu%2F~brian%2Fimps2007%2Fjunker-psysoc-2007.pps&ei=iBR6Ta6wMIbIrQewlvHKBQ&usg=AFQjCNE3hLCMty1WXr4SB-U5-dkOeFfqaQ&sig2=V2ZX-Lp1RAOwu0uiBfdwoQ

Gierl, M., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(3), 293-313. doi: 10.1111/j.1745-3984.2009.00082.x

Hartz, S. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practice*. Unpublished doctoral thesis, University of Illinois at Urbana-Champain.

Hartz, S., Roussos, L., & Stout, W. (2002). *Skill diagnosis: Theory and practice* [Computer software user manual for Arpeggio software]. Princeton, NJ: Educational Testing Service.

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.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272. doi:10.1177/01466210122032064

Lee, Y.W., & Sawaki, Y. (2009). Cognitive diagnostic approaches to language assessment: An overview. *Language Assessment Quarterly, 6(3)*, 172-189. doi: 10.1080/15434300902985108

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*(3), 205-237. doi: 10.1111/j.1745-3984.2004.tb01163.x

Montero, D. H., Monfils, L., Wang, J., Yen, W. M., & Julian, M. W. (2003, April). *Investigation of the application of cognitive diagnostic testing to an end-of-course high school examination.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Roussos, L., Xueli, X., & Stout, W. (2003). *Equating with the Fusion Model using Item Parameter Invariance*. Unpublished Manuscript, University of Illinois, Urbana-Champaign.

Rupp, A. A. (2007, April). *Unique characteristics of cognitive diagnosis models*. Paper presented at the annual meeting of National Council on Measurement in Education, Chicago, IL.

Rupp, A. A., Templin, J., & Henson, R. J. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Tatsuoka, K. K. (2009). Cognitive assessment: An introduction to the Rule Space Method. New York, NY: Routledge.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354. doi: 10.1111/j.1745-3984.1983.tb00212.x

HTML: http://jalt.org/test/ary_1.htm / **PDF**: http://jalt.org/test/PDF/Aryadoust1.pdf