Educational and Psychological Measurement
http://epm.sagepub.com/

Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing With Short Test Length

Chun Wang

Educational and Psychological Measurement published online 21 August 2013
DOI: 10.1177/0013164413498256

The online version of this article can be found at:
http://epm.sagepub.com/content/early/2013/08/21/0013164413498256

Published by:
SAGE
http://www.sagepublications.com

Article

Educational and Psychological Measurement
XX(X) 1–19

DOI: 10.1177/0013164413498256
epm.sagepub.com
SAGE

Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing With Short Test Length

Chun Wang[1]

Abstract

Cognitive diagnostic computerized adaptive testing (CD-CAT) purports to combine the strengths of both CAT and cognitive diagnosis. Cognitive diagnosis models aim at classifying examinees into the correct mastery profile group so as to pinpoint the strengths and weakness of each examinee whereas CAT algorithms choose items to determine those strengths and weakness as efficiently as possible. Most of the existing CD-CAT item selection algorithms are evaluated when test length is relatively long whereas several applications of CD-CAT, such as in interim assessment, require an item selection algorithm that is able to accurately recover examinees' mastery profile with short test length. In this article, we introduce the mutual information item selection method in the context of CD-CAT and then provide a computationally easier formula to make the method more amenable in real time. Mutual information is then evaluated against common item selection methods, such as Kullback–Leibler information, posterior weighted Kullback–Leibler information, and Shannon entropy. Based on our simulations, mutual information consistently results in nearly the highest attribute and pattern recovery rate in more than half of the conditions. We conclude by discussing how the number of attributes, *Q*-matrix structure, correlations among the attributes, and item quality affect estimation accuracy.

Keywords

cognitive diagnosis, computerized adaptive testing, Kullback–Leibler information, mutual information

[1]University of Minnesota, Minneapolis, MN, USA

Corresponding Author:
Chun Wang, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN 55455, USA.
Email: wang4066@umn.edu

Computerized adaptive testing (CAT) has been recently adopted by many high-stakes educational testing programs, such as the Graduate Management Admission Test (GMAT), the National Council of State Boards of Nursing (NCLEX), and the Armed Services Vocational Aptitude Battery (ASVAB). A CAT tailors the set of items to the examinee's ability level, so that no examinee receives too many overly easy or difficult items. Consequently, CAT can provide more accurate latent trait estimates using fewer items than required by Paper and Pencil tests (e.g., Weiss, 1982). Traditional CAT selects items to gain general information about a single, continuous latent trait. Alternatively, cognitive diagnosis assessment (CDA) aims to determine whether or not examinees have each of many attributes or skills underlying responses to test items. Unlike traditional CAT, CDA informs the specific cognitive skills required for designing effective remedial interventions in formative instruction (Cui, Gierl, & Chang, 2012, Leighton & Gierl, 2007). Models proposed to facilitate diagnostic assessment (Rupp, Templin, & Henson, 2010) include the rule-space model (Tatsuoka, 1983), Bayesian inference networks (Mislevy, Almond, Yan, & Steinberg, 1999), latent trait models (Yamamoto, 1989), and restricted latent class models such as the Deterministic Input; Noisy And gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001), Noisy Input; Deterministic And Gate (NIDA) model (Maris, 1999), and the Deterministic Input, Noisy Or Gate (DINO) model (Templin & Henson, 2006).

Although statistically sound and substantially useful models provide an infrastructure for cognitive assessment, combining cognitive diagnostic models (CDM) with adaptive procedures (so-called cognitive diagnostic computerized adaptive testing [CD-CAT], Cheng, 2009, Huebner, 2010) facilitates their practical and realistic implementation. A version of CD-CAT has already been applied in determining whether or not students possess specific skills (Jang, 2008). In the past decade, many traditional item selection algorithms have been adapted for CD-CAT, including the Kullback–Leibler (KL) Information Index (Xu, Chang, & Douglas, 2003), the Shannon entropy method (Xu et al., 2003), and the posterior weighted KL Index (PWKL, Cheng, 2009), among others. One method yet to be adapted to cognitive diagnosis settings is based on mutual information. Mutual information was recently proposed by Weissman (2007) for continuous trait-based classification testing and generalized by Mulder and van der Linden (2010) and Wang and Chang (2011) for use in multidimensional CAT. Mutual information has been shown to be more efficient than competing item selection methods, such as those based on KL-information and Shannon entropy, especially for short tests.

The objective of this article is to introduce a mutual information method for use in CD-CAT and evaluate its performance against the KL Index, the PWKL Index, and the Shannon entropy method, the three most widely used methods for CD-CAT (Huebner, 2010; Liu, You, Wang, Ding, & Chang, in press). The mutual information index, as originally proposed, requires several difficult and computationally demanding calculations. Therefore, we also derive a simplified calculation method intended to greatly reduce computational intensity.

The rest of the article is organized as follows. First, we discuss the CDM used in this study and introduce the three existing item selection algorithms for CD-CAT. Next, we introduce the mutual information method as applied to CDMs and derive a computationally simplified formula. We then evaluate the mutual information method (using simplified formula) against the existing item selection algorithms via two simulation studies. Finally, we discuss consequences of the simulation results and provide suggestions for further research.

The DINA Model

In cognitive diagnosis, one attempts to identify the tasks, subtasks, cognitive processes, and/or skills involved in responding to items on an assessment. Each task or skill is generally referred to as an *attribute*. The attributes of a math test might include, for example, converting mixed numbers to improper fractions, finding a common denominator, or multiplying fractions. Typically, content experts determine the attributes required for correct item responses. The general purpose of cognitive diagnosis is to identify which attributes each examinee has mastered based on the examinee's responses.

A critical component underlying almost all cognitive diagnosis models is the Q-matrix (Tatsuoka, 1995), which links individual items with one or more of the attributes. Given a set of J test items and K total attributes, the element in row j and column k of the Q-matrix should be 1 if item j requires attribute k, and 0 otherwise. Researchers typically assume that the Q-matrix is constructed by subject matter experts and test developers, but, in practice, the development of the Q-matrix using these means has proven to be quite time consuming and costly (Roussos, Templin, & Henson, 2007). As a result, de la Torre (2008) and Liu, Xu, and Ying (2012) have recently proposed to estimate the Q-matrix using statistical methods.

One commonly used cognitive diagnosis model is the DINA model (Haertel, 1989; Junker & Sijtsma, 2001). The DINA model assumes that, in principal, an examinee must have mastered *every* attribute associated with a particular item to respond correctly to that item (''And Gate''). However, the model recognizes that examinees might respond contrary to predictions (''Noisy''). Certain examinees will slip on an item, that is answer the item incorrectly even though they have all of the required attributes, whereas other examinees will successfully guess on an item, that is answer the item correctly even though they miss at least one of the required attributes. Given these properties, the DINA model–predicted probability that examinee i will respond correctly to item j is given by

$$P(Y_{ij} = 1 | \boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \quad (1)$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$—$\alpha_{ik}$ is 1 if examinee i has mastered attribute k, and 0 otherwise—denotes the examines' mastery profile, s_j is the probability that an examinee with all of the required attributes will ''slip'' and answer item j incorrectly, and g_j is the probability that an examinee with at least one missing attribute will

successfully guess the correct answer. Note that $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ij}^{q_{jk}}$ is 1 if and only if examinee i has mastered all the attributes required for a correct response to item j. The DINA model differs from other types of CDMs by how attributes interact to arrive at correct responses (i.e., conjunctive as in the DINA model vs. disjunctive as in the DINO model) or how parameters are included to model examinees' stochastic behavior (item level parameters as in the DINA model vs attribute-level parameters as in the NIDA model). For a full review of the different diagnostic classification models, please see Rupp, Templin, and Henson (2010).

Current Item Selection Methods

When applying adaptive testing algorithms to CDMs, one must determine how to choose items to administer to examinees. With respect to traditional CAT, item selection methods generally involve either maximizing information near the estimated location of the examinee in ability space or minimizing the error in that estimation (Reckase, 2009). For example, the most widely used information measure in adaptive testing, Fisher information, measures the amount of information that an observable random variable X carries about the unknown parameter θ. Because Fisher information requires the conditional distribution of X given θ to be continuous with respect to θ, one must develop alternative information measures for CD-CAT. Several other information measures, such as KL information and mutual information, are directly applicable to models measuring discrete latent traits.

Kullback–Leibler Information Index. In CD-CAT, items should be sequentially selected to optimize an objective function of the estimated attribute profile $\hat{\boldsymbol{\alpha}}$. A commonly used objective function, based on information theory, is the KL Information Index (Chang & Ying, 1996, Xu et al., 2003). By definition, KL information measures the divergence between two probability distributions, f and g, such that (Lehmann & Casella, 1998, sec.17; Cover & Thomas, 1991)

$$KL(f||g) = E_f\left[\log\frac{f(\theta)}{g(\theta)}\right]. \tag{2}$$

One generally thinks of Equation (2) as a distance-like measure in the sense that $KL(f||g)$ is greater than or equal to zero, equal to zero if and only if f and g are identical distributions, and increases as the distributions diverge. Chang and Ying (1996), Eggen (1999), and Chen, Ankenmann, and Chang (2000) adopted KL information in the context of unidimensional CAT, Henson and Douglas (2005) used KL information for constructing CDM-based tests, Xu et al. (2003) and Cheng (2009) applied KL information to CD-CAT, Veldkamp and van der Linden (2002) and Wang and Chang (2011) generalized the KL Index to multidimensional CAT.

To apply the KL Information Index to CD-CAT, we must determine distributions f and g in Equation (2). For examinee i, we seek to measure the unknown latent vector, $\boldsymbol{\alpha}_i$ (denoting the true cognitive profile for examinee i). Therefore, we must

differentiate $\boldsymbol{\alpha}_i$ from the remaining possible cognitive profiles. The appropriate KL Information Index would thus calculate the divergence between the conditional distribution of Y_{ij} (denoting the response of examinee i on item j) given the current estimated state, $f(Y_{ij}|\hat{\boldsymbol{\alpha}}_i)$ and the conditional distribution of Y_{ij} given the true latent state, $f(Y_{ij}|\boldsymbol{\alpha}_i)$, which is computed as

$$KL_j(\hat{\boldsymbol{\alpha}}_i||\boldsymbol{\alpha}_i) = \sum_{y=0}^{1} \log\left(\frac{P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i)}{P(Y_{ij}=y|\boldsymbol{\alpha}_i)}\right) P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i). \quad (3)$$

Because the true latent profile, $\boldsymbol{\alpha}_i$, is generally unknown, one cannot directly calculate Equation (3) except in trivial cases. Xu et al. (2003) proposed to sum up the KL divergences between $f(Y_{ij}|\hat{\boldsymbol{\alpha}}_i)$ and the conditional distribution of Y_{ij}, given each of the other latent ability states. Their KL Index can be written as

$$KL_j(\hat{\boldsymbol{\alpha}}_i) = \sum_{c=1}^{2^K} \left[\sum_{y=0}^{1} \log\left(\frac{P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i)}{P(Y_{ij}=y|\boldsymbol{\alpha}_c)}\right) P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i)\right], \quad (4)$$

where c indexes attribute profile and runs from 1 to 2^K. Equation (4) summarizes the discrimination power of an item j in differentiating the provisional estimated profile from all other possible attribute profiles. Xu et al. (2003) has shown that selecting items to maximize Equation (4) accurately recovers examinees' cognitive profiles.

Posterior Weighted Kullback–Leibler Information Index. One problem with the original KL Index, as defined by Equation (4) is that each $\boldsymbol{\alpha}_c$ is assumed equally likely to be the true attribute profile. In practice, some $\boldsymbol{\alpha}_c$s are more likely to be the true profile than others, given the response pattern of the examinee. Therefore, the decision of which item to select should be informed more by those profiles more likely to be the actual profile. For instance, if $\boldsymbol{\alpha}_c$ is more likely to be the examinees' true cognitive profile, then items that can better distinguish $\boldsymbol{\alpha}_c$ from $\hat{\boldsymbol{\alpha}}_i$ should be preferred. To quantify the contribution of each attribute profile to the KL Index, Cheng (2009) proposed a Bayesian version of Equation (4) by multiplying $KL_j(\hat{\boldsymbol{\alpha}}_i||\boldsymbol{\alpha}_c)$ with the corresponding posterior probability $\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1})$. This modified, posterior weighted KL Index (PWKL) can be written as

$$PWKL_j(\hat{\boldsymbol{\alpha}}_i) = \sum_{c=1}^{2^K} \left\{\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1}) \sum_{y=0}^{1} \left[\log\left(\frac{P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i)}{P(Y_{ij}=y|\boldsymbol{\alpha}_c)}\right) P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i)\right]\right\}, \quad (5)$$

where $\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1}) = p(\boldsymbol{\alpha}_c)\prod_{j=1}^{n-1} P(Y_{ij}=1|\boldsymbol{\alpha}_c)^{y_{ij}}[1-P(Y_{ij}=1|\boldsymbol{\alpha}_c)]^{1-y_{ij}}$. $p(\boldsymbol{\alpha}_c)$is the prior probability, and $\boldsymbol{y}_{n-1}$ is the vector of responses on $(n-1)$ items for examinee i (but subscript i for person is omitted hereafter in $\boldsymbol{y}_{n-1}$ to avoid notational clutter). Cheng (2009) has shown that selecting items by maximizing PWKL information yielded

higher measurement precision as compared with alternative item selection methods. The improvement of measurement accuracy was substantial for item banks with large slipping and guessing parameters and thus less informative items.

Shannon entropy method. An alternative information metric to KL-based measures is Shannon entropy. Shannon entropy quantifies the uncertainty inherent in the distribution of a single random variable (Shannon, 1948). Shannon entropy is maximized if distribution is uniform and minimized if the distribution is a single point mass. In CD-CAT, one would ideally like the posterior distribution of $\hat{\boldsymbol{\alpha}}$ to be a point mass. Therefore, Tatsuoka (2002) and Tatsuoka and Ferguson (2003) suggested selecting items that minimize the expected Shannon entropy. Specifically, let $\boldsymbol{y}_{n-1}$ again denote the response vector for examinee i, then the posterior expected Shannon entropy can be calculated as

$$\sum_{y=0}^{1}\left[\sum_{c=1}^{2^K}\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1},Y_n=y)\log\left(\frac{1}{\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1},Y_n=y)}\right)\right]\left[\sum_{c=1}^{2^K}P(Y_n=y|\boldsymbol{\alpha}_c)\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1})\right]. \tag{6}$$

Tatsuoka (2002) has shown that selecting items to minimize (6) should outperform the KL approaches. Wang and Chang (2011) recently found that applying a Shannon entropy item selection algorithm to continuous trait-based multidimensional CAT resulted in accurate ability estimates. As shown by Wang and Chang, the Shannon entropy–based method can be viewed as a special case of mutual information. Therefore, one would expect the more general mutual information–based method to be at least as optimal as Shannon entropy. Because of the desirable properties of mutual information, we next introduce a mutual information–based index in CD-CAT.

Mutual Information in Cognitive Diagnostic Model

The Expected Mutual Information Index

Given two random variables, X and Y, mutual information is defined as the KL divergence between their joint distribution, $f(X, Y)$, and the product of their marginal distributions, $f(X)$ and $f(Y)$. Using KL divergence as defined in (2), mutual information can thus be written as[1]

$$I(X;Y)=\sum_x\sum_y f(x,y)\log\left[\frac{f(x,y)}{f(x)f(y)}\right]. \tag{7}$$

$I(X;Y)$ measures the divergence between the true joint distribution and the joint distribution under independence. In other words, mutual information evaluates the dependence between X and Y—the more useful information X carries about Y (or

vice versa, because mutual information is a symmetric measure), the larger $I(X;Y)$ will be.

In CD-CAT, replace$f(y)$ in (7) by the posterior distribution of $\boldsymbol{\alpha}$ given the first $(n-1)$ items, $\pi(\boldsymbol{\alpha}|\boldsymbol{y}_{n-1})$, and $f(x)$ by the binomial distribution of the next response given all previous responses, $p(y_n|y_{n-1})$. Then the mutual information between $\boldsymbol{\pi}(\boldsymbol{\alpha}|\boldsymbol{y}_{n-1})$ and $p(y_n|\boldsymbol{y}_{n-1})$ indicates the information gained about unknown $\boldsymbol{\alpha}$ when another item is added to the test. As shown by Wang and Chang (2011), maximizing this mutual information is equivalent to maximizing the expected posterior KL divergence between two subsequent posterior distributions of $\boldsymbol{\alpha}$.

We will now describe the explicit equation for mutual information in CDM. After administering the nth item, the posterior density is updated to a new posterior according to Bayes theorem,

$$\pi(\boldsymbol{\alpha}|y_{n-1},y_n)=\frac{p(y_n|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}|\boldsymbol{y}_{n-1})}{p(y_n|\boldsymbol{y}_{n-1})} \tag{8}$$

where $p(y_n|\boldsymbol{y}_{n-1})$ is the posterior predictive probability. Then the KL distance between subsequent posterior distributions is equal to

$$KL(\pi(\boldsymbol{\alpha}|\boldsymbol{y}_{n-1},y_n)||\pi(\boldsymbol{\alpha}|\boldsymbol{y}_{n-1}))=\sum_{c=1}^{2^K}\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1},y_n)\log\frac{\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1},y_n)}{\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1})} \tag{9}$$

Unfortunately, we cannot calculate the previous equation because of the nth response being unknown. Therefore, we must take expectations to eliminiate the actual response. After modifying the previous equation, the expected KL distance between two subsequent posterior distributions is calculated as

$$\sum_{y=0}^{1}p(Y_n=y|\mathbf{y}_{n-1})\left[\sum_{c=1}^{2^K}\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1},Y_n=y)\log\frac{\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1},Y_n=y)}{\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1})}\right]. \tag{10}$$

Any item that maximizes (10) will also maxmize the expected mutual information between $\pi(\boldsymbol{\alpha}|\boldsymbol{y}_{n-1})$ and $p(y_n|\boldsymbol{y}_{n-1})$. As emphasized in Wang and Chang (2011), the expected mutual information Index does not rely on intermediate estimates of $\hat{\boldsymbol{\alpha}}$ and is thus less contaminated by the associated measurement error in $\hat{\boldsymbol{\alpha}}$.

Computational Simplification

Although, in principle, mutual information has benificial properties, the corresponding equation is computationally intensive. Equation (10) requires a triple summation (because both $p(Y_n=y|\boldsymbol{y}_{n-1})$ and $\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1})$ are calculated via summations) over 2^K possible cognitive profiles. In this subsection, we algebraically transform (9) to greatly simplify the calculation. Specifically, each component in (9) can be computed separately as

$$p(Y_n = y|\boldsymbol{y}_{n-1}) = \sum_{c=1}^{2^K} p(Y_n = y|\boldsymbol{\alpha}_c)\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_{n-1}) = \frac{\sum_{c=1}^{2^K} p(\boldsymbol{y}_{n-1}, Y_n = y|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)}{\sum_{c=1}^{2^K} p(\boldsymbol{y}_{n-1}|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)} \tag{11}$$

and

$$\pi(\boldsymbol{\alpha}|\boldsymbol{y}_{n-1}, Y_n = y) = \frac{p(\boldsymbol{y}_{n-1}, Y_n = y|\boldsymbol{\alpha})p(\boldsymbol{\alpha})}{\sum_{c=1}^{2^K} p(\boldsymbol{y}_{n-1}, Y_n = y|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)}. \tag{12}$$

Next, define $h_1 = \sum_{c=1}^{2^K} p(\mathbf{y}_{n-1}|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)$ and $h_2 = \sum_{c=1}^{2^K} p(\boldsymbol{y}_{n-1}, Y_n = y|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)$. Then (10) can be written as

$$\begin{aligned}
&\sum_{y=0}^{1} \frac{h_2}{h_1}\left[\sum_{c=1}^{2^K} \frac{p(\boldsymbol{y}_{n-1}, Y_n = y|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)}{h_2} \log \frac{p(Y_n = y|\boldsymbol{\alpha}_c)h_1}{h_2}\right] \\
&= \frac{1}{h_1}\left[\sum_{\boldsymbol{y}=0}^{1}\sum_{c=1}^{2^K} p(y_{n-1}, Y_n = y|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c) \frac{\log p(Y_n = y|\boldsymbol{\alpha}_c)}{h_2} + \log h_1 \sum_{y=0}^{1}\sum_{c=1}^{2^K} p(\boldsymbol{y}_{n-1}, Y_n = y|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)\right]
\end{aligned} \tag{13}$$

Note that $\sum_{\boldsymbol{y}=0}^{1}\sum_{c=1}^{2^K} p(\mathbf{y}_{n-1}, Y_n = y|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c) = \sum_{c=1}^{2^K} p(\boldsymbol{y}_{n-1}|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c) = h_1$ is a constant that does not depend on candidate items. Removing this constant term, a revised expected mutual information Index can be written as

$$\frac{1}{h_1}\sum_{\boldsymbol{y}=0}^{1}\left[\sum_{c=1}^{2^K} p(\mathbf{y}_{n-1}, Y_n = y|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c) \frac{logp(Y_n = y|\boldsymbol{\alpha}_c)}{h_2}\right]. \tag{14}$$

The rank ordering of prospective items will be the same regardless of whether using Equation (10) or (14). Although the mutual information Index will always be nonnegative by definition, (14) could be negative depending on the value of the constant term h_1. Therefore, multiplying the modified mutual information index, as defined in (14), by a positive exposure control or constraint management index (such as the maximum priority index, Cheng & Chang, 2009) might incorrectly order the items. One could, of course, add a large enough constant to (14) to force nonnegativity and prevent an inappropriate item from being selected.

Simulation Studies

We conducted two simulation studies to compare the mutual information item selection algorithm against three current item selection methods. The first simulation study

assumed that a test measures five attributes (a medium number that is often considered in literature, see Cheng, 2009 or Chen, Xin, Wang, & Chang, 2012). However, five attributes might underestimate the dimensionality underlying a typical application of CDM. Therefore, the second simulation study assumed that the test measures eight attributes (a larger number that is more likely to be observed in real test data, e.g., DeCarlo, 2011; Liu et al., in press). In both studies, we simulated data using the DINA model because of its popularity and simplicity. However, all the discussed item selection algorithms can straightforwardly be applied to any CDM.

Simulation Design

In the first study, item banks differed primarily in Q-matrix structure. Half the banks obeyed simple structure, in which one fifth of the items exclusively measured each of the five attributes. The remaining banks followed complex structure, in which each attribute was roughly measured by half the items in the item bank. To form a complex Q-matrix, every item-by-attribute was accompanied by a random uniform number between 0 and 1. If the random number was smaller than 0.5, then the corresponding Q-matrix entry was 1, indicating that the attribute was measured by the item. Otherwise, the corresponding Q-matrix entry was forced to be 0. Finally, to prevent trivial rows in the Q-matrix, every item was constrained to measure at least one of the five attributes. The total number of items in the bank is 350.

Aside from Q-matrix structure, conditions also differed in item selection algorithm, bank information, and test length. We used five item-selection algorithms: (a) maximum KL information, (b) maximum PWKL information, (c) minimum Shannon entropy, (d) maximum MI, and (e) random selection. With respect to bank information, the DINA model evaluates item quality by the magnitude of item-level slipping and guessing parameters. In the ''high'' information condition, guessing parameters were generated from a uniform distribution with a minimum of .05 and a maximum of .2, and slipping parameters were generated from a uniform distribution with a minimum of .1 and a maximum of .3. In the ''low'' information condition, guessing parameters were generated from a uniform distribution with a minimum of .1 and a maximum of .3, and slipping parameters were generated from a uniform distribution with a minimum of .15 and a maximum of .4. Test length was either set to 5 or 10 items. Therefore, we had 2 (Q-matrix structure) $\times$ 5 (item selection algorithm) $\times$ 2 (bank information) $\times$ 2 (test length) = 40 total conditions for the first study.

Across conditions, we generated a 1000-by-5 (person-by-attribute) $\boldsymbol{\alpha}$-matrix with each row representing a simulee's cognitive profile. Entries in the $\boldsymbol{\alpha}$-matrix were simulated based on the ''higher-order DINA model'' procedures proposed by de la Torre and Douglas (2004). For each simulee, we generated a standard normal value representing the score on a higher order latent trait (θ_i for simulee i). Given each θ_i, $\boldsymbol{\alpha}_i$ was generated by dichotomizing a five-dimensional continuous vector computed via a logistic function. The slope in the logistic function governs the correlation among the attributes, and in this study the slope was chosen such that all the attributes

correlated approximately .35 with all other attributes (Wang, Chang, & Douglas, 2012).

Real tests normally contain a large number of skills, such as the eight-attribute structure identified from the fraction subtraction data (DeCarlo, 2011). Hence, we also wanted to estimate the performance of mutual information given a bank composed of a larger number of attributes. We therefore devised a second simulation study by using an item bank designed to measure eight attributes. The item bank size is the same as in the first study. We also varied the Q-matrix structure (between simple and complex) and attribute correlations (between all correlating .6 and all correlating 0). Unlike the first study, we fixed bank information to be high and adaptive test length to be 10 to keep the simulation study more focused and because item quality and test length are less interesting factors that have been previously and extensively studied (see Cheng, 2009 or Wang et al., 2012). Therefore, we had 2 (Q-matrix structure) $\times$ 5 (item selection algorithm) $\times$ 2 (correlation levels) = 20 total conditions for the second study.

For both simulation studies, two measures were computed to evaluate both the attribute and pattern recovery rates. The attribute-level recovery rate (AR) is defined as the marginal proportion of attributes that are correctly identified, or

$$AR_k = \frac{\sum_{i=1}^{N} A_{ik}}{N} = \frac{\sum_{i=1}^{N} \left(I_{(\hat{\alpha}_{ik}, \alpha_{ik})}\right)}{N}, \quad (k = 1, 2, \ldots, K). \tag{15}$$

The pattern recovery rate (PR) is defined as the proportion of entire attribute patterns that are correctly identified, or

$$PR = \frac{\sum_{i=1}^{N} R_i}{N} = \frac{\sum_{i=1}^{N} (I_{(\hat{\boldsymbol{\alpha}}_i, \boldsymbol{\alpha}_i)})}{N}. \tag{16}$$

Note that for both Equations (15) and (16), I is an indicator function. Therefore, in Equation (14), when $\hat{\alpha}_{ik} = \alpha_{ik}$ indicating that the attribute k is correctly classified for person i, $A_{ik} = 1$, otherwise, A_{ik} is set to be 0. Similarly in Equation 16, if $\hat{\boldsymbol{\alpha}}_i$ is identical to α_1, then $R_i = 1$. If one of those conditions does not hold, then the corresponding value is set to 0. Moreover, because of the more restrictive condition for Equation 16, pattern recovery rate is always lower than or equal to the attribute recovery rate.

In all conditions, the profile estimates, $\hat{\boldsymbol{\alpha}}$, were obtained via maximum a posteriori (MAP) with a uniform prior. One consequence of assuming a uniform prior on α is that the mutual information method should be most advantaged for highly correlated attributes. Unlike the remaining item selection algorithms, mutual information does not depend on accurate estimates of $\hat{\boldsymbol{\alpha}}$ and therefore would not be as adversely affected if the actual distribution of $\boldsymbol{\alpha}$ is far away from the prior uniform distribution.

Table 1. Number of Items Measuring (or Examinees Mastering) Each Attribute in Study 1.

		Attributes				
		1	2	3	4	5
Number of items	Simple Q	76	64	73	64	73
	Complex Q	168	162	170	176	176
Number of examinees		791	869	676	337	521

Results

Tables 1 to 4 present the descriptive statistics of the Q- and $\boldsymbol{\alpha}$-matrices for both studies. As shown in Table 4, when all the eight attributes are independent from each other, most of the examinees master three to five attributes. If all the attributes are independent, then all the cognitive profiles are equally likely, and 182 ($= C_8^3 + C_8^4 + C_8^5$) of the 256 possible profiles have three to five 1s. Conversely, when all attributes are highly correlated, then examinees typically either master very few attributes or many attributes, resulting in more frequencies at the two ends of the distribution (notice that there are 166 examinees mastering 3 attributes in the condition with highly correlated attributes, but this phenomenon might be just because of the sampling variability).

The attribute level and pattern recovery rates corresponding to Simulation 1 are shown in Table 5.

Notice that the pattern recovery rate is always smaller than the attribute recovery rate, as explained earlier. However, lengthening the test or improving the item bank information increases both recovery rates. Moreover, varying the Q-matrix structure does not seem to have a consistent effect on either recovery rate. Although strange at first blush, complex Q-matrices have counterbalancing effects on attribute recovery rates. On the one hand, allowing an item to measure more than one attribute results in an increased number of items measuring each attribute, as shown in Tables 1 and 2. If more items measure an attribute, then that attribute has additional, useful diagnostic information. On the other hand, allowing an item to measure more than one attribute results in a decreased discrimination of that item to differentiate two cognitive profiles. Henson and Douglas (2005) proposed a KL Information Index, as a generalization of Chang and Ying's (1996) original idea, to quantify the discrimination power of item j (denoted as $\bar{D}_j$) in differentiating any two different cognitive profiles. This KL Discrimination Index can be written as

$$\bar{D}_j = \frac{1}{2^K(2^K-1)} \sum_{u \neq v,\, u,\, v=1}^{2^K} D_{juv}, \text{ where } D_{juv} = \sum_{y=0}^{1} \log\left[\frac{P(Y_j = y|\boldsymbol{\alpha}_u)}{P(Y_j = y|\boldsymbol{\alpha}_v)}\right] P(Y_j = y|\boldsymbol{\alpha}_u), \quad (17)$$

with u and v indexing different cognitive profiles. If item j only measures attribute k, then D_{juv} will be nonzero if $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ differ by at least attribute k. Using the DINA model, it can be verified that

Table 2. Number of Items Measuring (or Examinees Mastering) Each Attribute in Study 2.

		Attributes							
		1	2	3	4	5	6	7	8
Number of items	Simple Q	45	38	42	47	47	42	38	51
	Complex Q	103	98	112	97	128	104	105	119
Number of examinees	Correlated	712	579	546	740	507	733	487	578
	Independent	512	526	485	517	495	476	509	478

Table 3. Number of Items Measuring (or Examinees Mastering) Each Possible Number of Attributes in Study 1.

		Number of attributes					
		0	1	2	3	4	5
Number of items	Simple Q	0	350	0	0	0	0
	Complex Q	0	88	88	116	50	8
Number of examinees		56	87	161	221	253	222

Table 4. Number of Items Measuring (or Examinees Mastering) Each Possible Number of Attributes in Study 2.

		Attributes								
		0	1	2	3	4	5	6	7	8
Number of items	Simple Q	0	350	0	0	0	0	0	0	0
	Complex Q	0	94	91	94	52	15	3	1	
Number of examinees	Correlated	125	9	26	166	52	40	98	175	309
	Independent	3	29	96	208	318	222	99	23	2

$$\bar{D}_j = (1 - s_j) \log \frac{1 - s_j}{g_j} + g_j \log \frac{g_j}{1 - s_j}. \tag{18}$$

In (18), $\bar{D}_j$captures the discrimination power of item j in differentiating examinees with and without mastering attribute k. If $\bar{D}_j$is high, then one can accumulate more information in recovering attribute k. If item j measures t different attributes, then

$$\bar{D}_j = \frac{2^{K-t+1}(1 - 2^{-t})}{2^K - 1} \left[(1 - s_j) \log \frac{1 - s_j}{g_j} + g_j \log \frac{g_j}{1 - s_j} \right] \tag{19}$$

Table 5. Attribute and Pattern Recovery Rate for Simulation Study 1.

Q-structure	Test length	Item quality	Item selection	Attribute					Pattern
				1	2	3	4	5	
Simple	5	High	PWKL	0.962	0.95	0.969	0.942	0.950	0.800
			Shannon	0.963	0.944	0.967	0.965	0.957	0.849
			Mutual	0.992	0.948	0.985	0.969	0.956	0.859
			KL	0.993	0.950	0.935	0.721	0.717	0.489
			Random	0.899	0.918	0.836	0.798	0.831	0.483
		Low	PWKL	0.967	0.938	0.959	0.93	0.919	0.757
			Shannon	0.948	0.936	0.953	0.93	0.939	0.757
			Mutual	0.989	0.938	0.972	0.937	0.948	0.797
			KL	0.988	0.945	0.88	0.686	0.705	0.448
			Random	0.867	0.896	0.8	0.75	0.786	0.408
	10	High	PWKL	0.998	0.996	0.999	0.997	0.995	0.985
			Shannon	0.995	0.993	0.992	0.989	0.989	0.976
			Mutual	1	0.992	0.994	0.994	0.992	0.973
			KL	1	0.962	0.972	0.808	0.887	0.67
			Random	0.932	0.937	0.904	0.892	0.887	0.635
		Low	PWKL	0.999	0.989	0.993	0.988	0.986	0.956
			Shannon	0.983	0.982	0.989	0.988	0.964	0.914
			Mutual	0.997	0.985	0.986	0.986	0.985	0.944
			KL	0.999	0.944	0.965	0.822	0.835	0.616
			Random	0.885	0.921	0.86	0.835	0.821	0.505
Complex	5	High	PWKL	0.984	0.974	0.974	0.96	0.97	0.896
			Shannon	0.99	0.982	0.982	0.952	0.968	0.934
			Mutual	0.992	0.984	0.998	0.982	0.984	0.948
			KL	0.938	0.908	0.946	0.878	0.932	0.69
			Random	0.88	0.88	0.826	0.898	0.834	0.584
		Low	PWKL	0.994	0.936	0.962	0.954	0.96	0.856
			Shannon	0.978	0.982	0.988	0.961	0.965	0.902
			Mutual	0.984	0.982	0.998	0.97	0.966	0.912
			KL	0.848	0.868	0.782	0.874	0.754	0.492
			Random	0.996	0.888	0.97	0.884	0.91	0.692
	10	High	PWKL	0.998	1	1	0.998	1	0.996
			Shannon	0.992	0.998	0.996	0.992	0.998	0.978
			Mutual	1	1	0.998	0.994	1	0.992
			KL	0.99	0.98	0.992	0.924	0.978	0.876
			Random	0.888	0.9	0.89	0.95	0.912	0.688
		Low	PWKL	0.996	0.994	0.998	0.998	1	0.986
			Shannon	0.994	0.998	0.998	0.984	0.988	0.976
			Mutual	0.998	0.996	1	0.99	0.996	0.98
			KL	0.988	0.974	0.994	0.934	0.974	0.872
			Random	0.858	0.886	0.84	0.932	0.868	0.596

Note. PWKL = posterior weighted Kullback–Leibler information method; Shannon = Shannon entropy method; KL = Kullback–Leibler Index method; Mutual = mutual information method; Random = random item selection method.

because only 2^{K-t} attributes will lead to $\eta = 1$. Note that the item information in Equation 19 is smaller than that in (18). In addition, only $\frac{1}{t}\bar{D}_j$ amount of information will discriminate examinees who have and have not mastered attribute k. Therefore, although allowing items to measure multiple attributes leads to more items loading on each attribute, the contribution of each item to every attribute is largely reduced. As a result, the relationship between a complex Q-matrix structure and the recovery rate depends on the interaction between item quality and the specific Q-matrix.

As shown in Table 5, the mutual information item selection algorithm generates nearly the most accurate attribute pattern recovery in more than half of the conditions. The advantage of mutual information over the other conditions is increased for short tests. This result is consistent with the performance of a mutual information item selection algorithm in multidimensional CAT (Wang & Chang, 2011). Because MI does not rely on intermediate estimates of $\hat{\boldsymbol{\alpha}}$, it is more robust to short test lengths than the PWKL and KL indices. Shannon entropy also does not depend on proximate estimates of $\hat{\boldsymbol{\alpha}}$. Wang and Chang (2011) showed that minimizing the expected Shannon entropy is equivalent to maximizing the KL divergence between the expected posterior distribution of $\hat{\boldsymbol{\alpha}}$ and a ''uniform prior.'' Because the mutual information method maximizes the KL divergence between the expected posterior distribution of $\hat{\boldsymbol{\alpha}}$ and the current posterior, mutual information should also be more informative, and thus result in slightly more efficient tests, than Shannon entropy.

Results of the second simulation study are shown in Table 6. Based on Table 6, higher correlations lead to more accurate results regardless of the Q-matrix structure. These results are consistent with research in multidimensional item response theory: By using multidimensional models, one can obtain greater precision in estimating dimensional ability by ''borrowing strength'' from other dimensions (i.e., de la Torre & Patz, 2005; Wang, Chen, & Cheng, 2004). And as is clearly shown by de la Torre and Patz (2005), higher correlations among the dimensions results in improved estimation accuracy.

Discussion and Conclusion

Most modern psychometrics focuses on tests designed to measure a unidimensional latent trait. These tests are often used to make summative decisions concerning admission, placement, and scholarships/fellowships. In the past few decades, teachers and administrators increasingly desire tests designed to assess several, finer-grained chunks of knowledge (DiBello & Stout, 2007). These ''formative'' assessments imply that their results are used to directly support teaching and learning. In contrast, ''summative'' testing evaluates students' overall proficiency at the end of the instruction. To fulfill the demand for formative diagnostics and remedial instruction, one must efficiently and accurately pinpoint examinees' strengths and weaknesses across one of a number of content areas. CD-CAT provides an intriguing solution to such a demand.

Table 6. Attribute and Pattern Recovery Rate for Simulation Study 2.

Q-structure	Correlation level	Item selection	Attribute								Pattern
			1	2	3	4	5	6	7	8	
Simple	Independent	PWKL	0.955	0.922	0.985	0.95	0.944	0.958	0.942	0.965	0.683
		Shannon	0.914	0.852	0.991	0.941	0.845	0.854	0.837	0.881	0.57
		Mutual	0.993	0.925	0.987	0.976	0.986	0.969	0.964	0.973	0.792
		KL	0.953	0.911	0.996	0.932	0.759	0.531	0.545	0.808	0.14
		Random	0.641	0.575	0.628	0.59	0.63	0.57	0.594	0.559	0.009
	High	PWKL	0.988	0.854	0.904	0.929	0.898	0.872	0.893	0.904	0.702
		Shannon	0.974	0.944	0.975	0.965	0.961	0.946	0.951	0.942	0.779
		Mutual	0.996	0.83	0.989	0.989	0.983	0.973	0.903	0.971	0.738
		KL	0.974	0.778	0.855	0.916	0.836	0.709	0.837	0.828	0.348
		Random	0.74	0.699	0.683	0.768	0.742	0.761	0.719	0.733	0.277
Complex	Independent	PWKL	0.964	0.788	0.858	0.824	0.79	0.774	0.771	0.808	0.399
		Shannon	0.941	0.882	0.952	0.959	0.906	0.886	0.889	0.877	0.622
		Mutual	0.99	0.859	0.976	0.971	0.932	0.964	0.868	0.896	0.664
		KL	0.962	0.75	0.699	0.82	0.831	0.746	0.749	0.831	0.3
		Random	0.567	0.557	0.595	0.527	0.561	0.567	0.541	0.599	0.008
	High	PWKL	0.946	0.917	0.99	0.923	0.946	0.942	0.958	0.966	0.677
		Shannon	0.972	0.908	0.986	0.967	0.908	0.949	0.879	0.937	0.725
		Mutual	0.996	0.943	0.983	0.982	0.981	0.959	0.969	0.978	0.81
		KL	0.925	0.894	0.997	0.87	0.766	0.924	0.633	0.89	0.405
		Random	0.666	0.583	0.639	0.617	0.647	0.646	0.585	0.607	0.156

Note. PWKL = posterior weighted Kullback–Leibler information method; Shannon = Shannon entropy method; KL = Kullback–Leibler Index method; Mutual = mutual information method; Random = random item selection method.

This article introduces a new item selection algorithm in CD-CAT, namely, the mutual information method. Because mutual information can be computationally intensive, this article also presents a simpler formula designed to make mutual information algorithms amenable in real-time CD-CAT.

We compared mutual information against already existing item selection algorithms for CD-CAT via two simulation studies. In these simulations, we manipulated several factors that might affect estimation accuracy, including number of attributes, test length, *Q*-matrix structure, item quality, and interattribute correlations. Not surprisingly, increasing the number of attributes measured by a CAT resulted in decreased reliability and accuracy. Because test length was limited, more attributes measured by a test implies fewer items per attribute. Moreover, increasing test length and item quality resulted in better recovery of the attribute profiles. Interestingly, one cannot easily predict the change in recovery rates for different *Q*-matrices. As explained in the previous section, a counterbalancing relationship exists between items measuring multiple attributes (more attributes for an item, worse measurement of each attribute) and more items measuring each attribute (more items for an attribute, more information for that attribute). DiBello and Stout (2007) explained that

> the items, possibly including complex items or, even more generally, open-ended tasks, must be designed from a cross-disciplinary evidentiary perspective to effectively measure well all of the specified skills. An open research topic is whether fewer complex items or greater numbers of simple (perhaps multiple choice) items are most informative for a given assessment application, both from the validity and the reliability perspectives. (p. 287)

The ''complex items'' of their statement might indicate open-ended question but could also be interpreted as items measuring numerous skills. Whether including such ''complex items'' is beneficial to attribute recovery depends on the *Q*-matrix of the test as well as the test blueprint.

One might wonder why we chose very short test lengths for all our comparisons. DiBello and Stout (2007) wrote that ''there is a stronger form of formative assessment called embedded assessment, which postulates that formative assessments should be seamlessly and periodically embedded in the curriculum for the purpose of improving teaching and learning'' (p. 289). If giving students an ''embedded assessment'' at the end of each instruction period, one must prefer short tests to prevent unnecessary loss of lecture time. Therefore, the CD-CAT item selection algorithm must be able to recover examinees' mastery profiles with short tests. Based on our simulations, the mutual information method generates more accurate results than competing methods in almost all conditions assuming relatively short tests.

The simulation studies discussed in this article are by no means exhaustive. Future simulations should be conducted to determine the effect of exposure control on profile recovery using the various item selection algorithms discussed in this article (Wang, Chang, & Huebener, 2011). Moreover, nonstatistical constraints, such as content constraints, word counts, or balancing item keys in the test, could also be added to fulfill the requirement in the test blueprint (Mao & Xin, 2013). Although we

mainly used examples from educational measurement in this article, one could apply our results to CDM of psychological assessment and diagnosis. Templin and Henson (2006) demonstrated that various CDMs can aid accurate diagnosis of psychological disorders. Ensuring efficient and accurate psychological tests will allow more time to be spent on treating rather than diagnosing the disorder.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. Whenever possible, without introducing ambiguity, we ignore the distinction between random variables and their realizations in the formulae.

References

Chang, H., & Ying, Z. (1996). A global information approach to computerized adoptive testing. *Applied Psychological Measurement*, *20*, 213-229.

Chen, P., Xin, T., Wang, C., & Chang, H. (2012). On-line calibration methods in cognitive diagnostic computerized adaptive testing. *Psychometrika*, *77*, 201-222.

Chen, S. Y., Ankenmann, R. D., & Chang, H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*, 241-255.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619-632.

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*, 369-383.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York, NY: Wiley.

Cui, Y., Gierl, M., & Chang, H. (2012). Evaluating item selection algorithms in computerized adaptive testing for cognitive diagnosis: a simulation study. *Journal of Educational Measurement*, *49*, 19-28.

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343-362.

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

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295-311.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: the DINA model, classification, latent classes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8-26.

DiBello, L. V., & Stout, W. (2007). Guest editors' instruction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*, 285-291.

Eggen, T. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249-261.

Haertel, E. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333-352.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*, 262-277.

Huebner, A., (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation*, *15*(3). Retrieved from http://pareonline.net/getvn.asp?v=15&n=3.

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

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

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed). New York, NY: Springer.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *27*, 3-16.

Liu, H., You, X., Wang, W., Ding, S., & Chang, H-H. (in press). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*, 609-618.

Mao, X., & Xin, T. (2013). The application of the Monte Carlo approach to cognitive diagnostic computerized adaptive testing with content constraints. *Applied Psychological Methods*. doi:10.1177/0146621613486015

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

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (437-446). San Francisco, CA: Morgan Kaufmann.

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 77-101). New York, NY: Springer.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT -based latent class models. *Journal of Educational Measurement*, *44*, 293-311.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379-423.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *51*, 337-350.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society Series B*, *65*, 143-157.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345-354.

Tatsuoka, K. K. (1995). Architecture of knowledge structure and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-361). Hillsdale, NJ: Lawrence Erlbaum.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287-305.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575-588.

Wang, C., & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive tests-Gaining information from different angles. *Psychometrika*, *76*, 363-384.

Wang, C., Chang, H.-H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, *44*, 95-109.

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic CAT. *Journal of Educational Measurement*, *48*, 255-273.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*, 116-136.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473-492.

Weissman, A. (2007).Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*, 41-58.

Xu, X., Chang, H., & Douglas, J. (2003, April). Computerized adaptive testing strategies for cognitive diagnosis. *Paper presented at the annual meeting of National Council on Measurement in Education*, Chicago, IL.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class model* (ETS Research Rep. RR-89-41). Princeton, NJ: Educational Testing Service.