TRANSITION TO MULTIDIMENSIONAL AND COGNITIVE DIAGNOSIS ADAPTIVE TESTING: AN OVERVIEW OF CAT

Lokman Akbay, PhD
Educational Statistics & Measurement
Rutgers, The State University of New Jersey
lokmanakbay@gmail.com

Mehmet Kaplan, PhD
Educational Statistics & Measurement
Rutgers, The State University of New Jersey

ABSTRACT
Although many early adaptive testing methodologies in the literature are based upon unidimensional item response theory (IRT) models, these methodologies have been generalized to or adjusted for multidimensional cases. Along with the developments on the multidimensional adaptive testing, cognitive diagnosis modeling has also shown rapid development over the past decades. Despite of its novelty, researchers have already conducted studies to manage to implement cognitive diagnosis computerized adaptive testing (CD-CAT). Following these developments, this manuscript aims to compile and highlight the developments in multidimensional computerized adaptive testing and review the advances in the CD-CAT development.

INTRODUCTION

Test administration process may be categorized as (a) individually or (b) group administered with respect to type of administration. Both types have some advantages and disadvantages such as the advantage of uniformity of test situation and vastly reduced cost of mass-administered tests. However, a mass-administered test must take into account the assumption that the examinee ability range is broad. Therefore, in order to effectively measure all examinees' ability levels, the test should be consisted of items with varying difficulty levels (i.e., easy-moderate-hard) so that test difficulty matches examinee group.

Computerized adaptive testing (CAT), however, selects and administers items that are most informative with respect to the current ability estimate. Meijer and Nering (1999) stated the objective of CAT as the attempt to construct an optimal test for each examinee based on his/her ability level. Therefore, it reduces the possibility of administering an item which is too far from examinee's ability level. To achieve this goal, an item is selected from an item pool consisting of items with various difficulty levels for administration based on examinee's responses to the previous items. This item administration procedure continues until a stopping rule (e.g., reaching to a predetermined number of items or threshold for standard error of the estimate) is satisfied.

Although many early adaptive testing methodologies in the literature are based on unidimensional item response theory (IRT) models, these methodologies have been recently generalized to or adjusted for multidimensional cases. Another measurement and assessment related subject that has shown rapid development over the past decade is the cognitive diagnosis models (CDMs). These models are used to extract diagnostic information from cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014). CDMs are used to classify examines one of the latent classes, which are characterized by a K number of discretely defined cognitive competencies, skills, and strategies. A latent class in which examinee is assigned to shows the examinee's attribute profiles in terms of mastery or nonmastery status of attributes. Despite its novelty, researchers have been conducting research to manage to implement cognitive diagnosis computerized adaptive testing (CD-CAT).

Following these developments, this paper aims to (1) compile and highlight the developments in multidimensional computerized adaptive testing (MAT), and (2) review the advances in the CD-CAT development. This study provides an overview of recent developments in adaptive testing with an expectation of providing researchers and practitioners with pragmatic information. The rest of the manuscript is organized as following: The next section will briefly explain the CDM framework. Then, adaptive testing system components and the developments within them will be reviewed. Finally, a discussion section will conclude the manuscript.

COGNITIVE DIAGNOSIS MODELS

In psychometric literature, a generic term *attribute* is used to refer to target discrete skills and strategies to be measured (de la Torre, 2009). Based on examinee's observed responses, CDMs assign each examinee a vector that shows mastery and nonmastery of measured attributes. This vector is typically binary where 1 and 0 indicate presence or absence of each of K attributes, respectively. Although the types and psychometric properties of

CDMs are not in the scope of this paper, it should be noted that various reduced and general CDMs have been recently developed. A broad discussion on these models can be found in Rupp and Templin (2008), and de la Torre (2011).

Most, if not all, CDMs utilize an item-by-attribute matrix, which is referred to as Q-matrix (Tatsuoka, 1985). This matrix specifies the association between the items and attributes. Each row of the matrix corresponds to an item that indicates the necessary attributes for successful completion of the item. For attributes k=1…K and items j=1….J in a test (or more generally in an item bank for CAT), the Q-matrix element q_{jk} is defined as

$$q_{jk} = \begin{cases} 1, & \textit{if item } j \textit{ requires attribute } k \\ 0, & \textit{otherwise} \end{cases}$$

For instance, when K=3, if *jth* item requires the first and the third attributes, then the *jth* row of the Q-matrix becomes {1,0,1}.

Furthermore, a test measuring K attributes partitions latent space into a total of 2^K latent classes. For example, when K=3, 2^3=8 latent classes possible (i.e., {0,0,0}, {1,0,0}, {0,1,0}, {0,0,1}, {1,1,0}, {1,0,1}, {0,1,1}, and {1,1,1}). By employing an appropriate CDM, each examinee is assigned to one of these latent groups where the group labels become examinees' attribute profiles. When examinee i is classified to the latent class {1,1,0}; it implies that examinee i has mastered the first and the second attributes but not the third one.

DEVELOPMENTS IN ADAPTIVE TESTING SYSTEMS

Item Pool

A calibrated item pool is a collection of test items with their item parameters stored in a computer-media (Reckase, 2009). In adaptive test environment, individualized tests are introduced to examinees, which require composition of many different forms of the same test. Flaugher (2000) stated that adaptive algorithm can do better job as the quality of the item pool increases. Flaugher (2000) and Reckase (2009) pointed out that the best and the most sophisticated adaptive testing procedure would not perform well if the quality of items in the pool is poor or items in the pool are not appropriate for target population. Therefore, not only size of an item pool but also characteristics of items within the pool are among the important considerations.

In adaptive testing systems two types of parameter estimations are identified. These distinct types are *initial calibration* and *on-line calibration*. In the former, responses are solicited from examinees only for not yet calibrated items. In the latter type of calibration, examinees give responses to both new and previously calibrated items during the adaptive test administration. According to Wainer and Mislevy (2000), *initial calibration* may be required for situations where (1) a novice test is being developed, and (2) an existing conventional test being adapted to a CAT. However, Wainer and Mislevy (2000) pointed out that, when an existing conventional test is adapted to CAT, an equating step might be required to adjust the item parameters for presentation effect (effect of testing format).

One of the issues with CAT is controlling item exposure rates. Overused items need to be replaced with new items; therefore, the parameters of the new items must be obtained in the established scale. Thus, o*n-line calibration* needs to be considered if the parameters of new items need to be estimated within testing process by introducing new items along with the calibrated items (Wainer & Mislevy, 2000). One way of on-line calibration is to carry out a large, independent, calibration study with some linking items. In this method, one needs to find a linear transformation of new calibration, which matches up to the pre-and post-estimation of linking items. Then, this linear transformation can be used to bring the new items onto the existing scale.

For the item calibration process, expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is a common approach for both IRT and CDM. Many existing commercial and freeware software programs employs EM algorithm for item calibration due to its computation efficiency (see Rupp and Templin (2008) for a list of software programs that can be used for CDM estimation). A practical issue for item calibration using EM algorithm is the required sample size. De Ayala (2009) argued that based on the research on IRT parameter estimates, 1000 individuals were enough for accurate and precise item parameter estimates via marginalized maximum likelihood estimation (p. 130).

In diagnosis model estimations, required sample size for accurate calibration depends on the specific CDM and number of attributes measured in a test. For example, regardless of the number of attributes measured, the deterministic input, noisy and "gate" (DINA: Junker & Sijtsma, 2001) model allows only two item parameters,

whereas, the generalized-DINA (de la Torre, 2011) model specifies $2^{K_j^*}$ item parameters to be estimated, where K_j^* shows the number of required attribute by item *j*. For example, if *jth* item requires four attributes, 16 item parameters need to be estimated for that item. Although Rupp and Templin (2008) claimed that a couple of hundreds examinees per item were sufficient for simplest models such as DINA; there is not enough research on the systematic investigation of the relationships between minimum required sample size, types of CDMs, and number of attributes measured by a test.

Item Selection Rule

To build and carry out an efficient CAT, the most informative item should be selected based on the examinee's the most recent ability estimate. This fact gives rise to use of a number of efficient and practical item selection algorithms. The goal of most IRT based CAT is to select a subset of items that provide sufficient information to accurately locate the examinee within the ability space. Similarly, CD-CAT aims to select a subset of items that provide sufficient information to accurately classify examinees into latent classes. Therefore one of the essential components of adaptive testing is a mechanism for selecting items from the item pool. This item selection procedure (to select m^{th} item) is applied in real time depending on the information available about the examinee's ability after applying *m-1* items. Item selection rules for unidimensional-CAT have been generalized to multidimensional cases.

The *Fisher information* is a common method for measuring the amount of information carried by observable variables about an unknown parameter. In multidimensional cases, the Fisher information matrix is considered as a convenient measure. For item *j*, the information matrix is defined as

$$I_j(\boldsymbol{\theta}) \equiv -E\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\ln f(X_j \mid \boldsymbol{\theta}) = \frac{Q_j(\boldsymbol{\theta})(P_j(\boldsymbol{\theta})-\gamma_j)^2}{P_j(\boldsymbol{\theta})(1-\gamma_j)^2}\boldsymbol{\alpha}_j\boldsymbol{\alpha}_j^{'} \quad (1)$$

where, $\boldsymbol{\theta}$ is the multidimensional ability vector, $P_j(\boldsymbol{\theta})$ is the probability correct response to item *j*, $Q_j(\boldsymbol{\theta}) = 1 - P_j(\boldsymbol{\theta})$, γ_j is the guessing parameter, and $\boldsymbol{\alpha}_j^{'}$ is the transpose of the vector of discrimination parameters in a multidimensional three parameter logistic model.

Because of the additivity property of the Fisher information, the information matrix of a set of *S* items is obtained by summing the item information matrices,

$$\boldsymbol{I}_S(\boldsymbol{\theta}) = \sum I_j(\boldsymbol{\theta}). \quad (2)$$

Once the $\boldsymbol{\theta}$ is substituted by its estimate, $\hat{\boldsymbol{\theta}}$, in the *equation 2*, an estimate of item and test information matrices are obtained. Thus, in the process of selecting the m^{th} item, after administering *m-1* items, the amount of information to be maximized is expressed as

$$\boldsymbol{I}_S^{(m-1)}(\hat{\boldsymbol{\theta}}^{(m-1)}) + \boldsymbol{I}(\hat{\boldsymbol{\theta}}^{(m)}). \quad (3)$$

Then an item selection algorithm based on the Fisher information selects an item (i.e., m^{th} item) such that $\boldsymbol{I}_S^{(m)}(\hat{\boldsymbol{\theta}}^{(m)})$ is the largest.

Segall (1996) introduced a new item selection approach for MAT (i.e., maximizing the determinant of the fisher information matrix) based on the relationship between the Fisher information matrix and the confidence region around the estimates. Segall (1996) emphasized that when

$$\det \mid \boldsymbol{I}(\hat{\boldsymbol{\theta}}^{(m-1)}) + \boldsymbol{I}(X^{(m)}) \mid \quad (4)$$

is maximized, the volume of the confidence ellipsoid is minimized. In *equation 4*, the term on the left is the test information matrix for *m-1* administered items; and the term on the right is the item information matrix obtained by administering item *m*. The purpose of this rule is to select the next item that has an item information matrix that results in the maximum value for the determinant of the sum.

Segall (1996) specified another approach referred to as largest decrement in the volume of the Bayesian credibility ellipsoid. This approach is connected to Bayesian modal approach used for ability estimation. The goal of this method is selecting a candidate item that results in the largest decrease in the volume of the Bayesian credibility ellipsoid for ability estimation. As reported in Reckase (2009), Segall (1996) assumed that the prior distribution for the $\boldsymbol{\theta}$-space was multivariate normal with variance-covariance matrix $\boldsymbol{\Phi}$. Based on this assumption; he claimed that the volume of the Bayesian credibility ellipsoid for the estimate of $\boldsymbol{\theta}$ becomes

$$\det | \boldsymbol{I}(\hat{\boldsymbol{\theta}}^{(m-1)}) + \boldsymbol{I}(X^{(m)}) + \Phi^{-1} | . \quad (5)$$

Once this expression is maximized, the volume of the credibility ellipsoid is minimized.

Veldkamp and van der Linden (2002) adapted Kullback-Leibler (KL) information for item selection in multidimensional adaptive testing. KL information is a non-symmetric distance measure to account for divergence between two probability distributions (Cover & Thomas, 1991). It should be noted here that Chang and Ying (1996) have used the KL information for the unidimensional-CAT. For a binary item response X_j and true examinee ability vector $\boldsymbol{\theta}_0$, KL information for item *j* is defined as

$$K_j(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = E\left[\ln \frac{L(\boldsymbol{\theta}_0 \mid X_j)}{L(\boldsymbol{\theta} \mid X_j)} \right] \quad (6)$$

where $\boldsymbol{\theta}$ is another possible ability vector and L is the likelihood function $P_j(\theta)^{X_j} Q_j(\theta)^{(1-X_j)}$. Veldkamp and van der Linden (2002) showed that the KL information is additive such that information provided by individual items are summed to obtain KL for a total of *m* administered items. Therefore, Veldkamp and van der Linden (2002) suggested selecting an item that maximizes the posterior expected KL information. It has been shown that item selection based on KL information outperforms to traditional Fisher information (Reckase, 2009; Veldkamp & van der Linden, 2002).

Due to the fact that latent variables being measured in CDMs are discrete, Fisher information related item selection rules cannot applied to CD-CAT as they require continuous latent variables (Xu, Chang, & Douglas, 2003). Alternatively, Xu et al. (2003) have suggested using KL information and Shannon entropy (SHE) for CD-CAT item selection. KL information for CD-CAT was defined as

$$KL_{j(\hat{\alpha}_i)} = \sum_{l=1}^{2^K} \left[\sum_{x=0}^{1} \log \left(\frac{P(X_j = x \mid \hat{\boldsymbol{\alpha}}_i)}{P(X_j = x \mid \boldsymbol{\alpha}_l)} \right) P(X_j = x \mid \hat{\boldsymbol{\alpha}}_i) \right] \quad (7)$$

where $l=1,\ldots,2^K$ possible latent classes defined by *K* attributes, $P(X_j = x \mid \hat{\boldsymbol{\alpha}}_i)$ is the probability of correct response of examinee *i* to item *j* given the examinee's current attribute profile estimate $\hat{\boldsymbol{\alpha}}_i$, and $\boldsymbol{\alpha}_l$ is attribute profile other than $\hat{\boldsymbol{\alpha}}_i$.

SHE is a measure of uncertainty in probability distributions (Cheng, 2009), which can be considered as the measure of flatness of posterior distribution of latent classes in CDM. The item selection rule specified in Xu et al. (2003) is based on the minimization of expected SHE of the posterior distribution of $\hat{\boldsymbol{\alpha}}$. Xu et al. reported that, in comparison with KL, SHE required more computation time than KL; yet it was more efficient in terms of classification accuracy. Although both SHE and KL information based item selection rules were reported to be promising, Cheng (2009) achieved higher classification accuracy by modifying KL item selection rule in CD-CAT. She proposed a rule based on posterior weighted KL (PWKL)

$$PWKL_{j(\hat{\alpha}_i^{(m)})} = \sum_{l=1}^{2^K} \left[\sum_{x=0}^{1} \log \left(\frac{P(X_j = x \mid \hat{\boldsymbol{\alpha}}_i^{(m)})}{P(X_j = x \mid \boldsymbol{\alpha}_l)} \right) P(X_j = x \mid \hat{\boldsymbol{\alpha}}_i^{(m)}) \pi_i^{(m)}(\boldsymbol{\alpha}_l) \right] \quad (8)$$

where $P(X_j = x \mid \boldsymbol{\alpha}_l)$ is the probability correct response of the item *j* given the attribute profile $\boldsymbol{\alpha}_l$, $\pi_i^{(m)}(\boldsymbol{\alpha}_l)$ is the posterior probability of examinee *i* at iteration *m* (i.e., after administering *m* items). Posterior probability of examinee *i* after administering *m* items is

$$\pi_i^{(m)}(\boldsymbol{\alpha}_l) \propto \pi_i^{(0)}(\boldsymbol{\alpha}_l) L(X_i^{(m)} \mid \boldsymbol{\alpha}_l) \quad (9)$$

where $\pi_i^{(0)}(\boldsymbol{\alpha}_l)$ is the current prior (e.g., the prior at the beginning of the test administration or posterior after administering *m*-1 items) and $L(X_i^{(m)} \mid \boldsymbol{\alpha}_l)$ is the likelihood of examinee *i*'s response vector $X_i^{(m)}$ given the attribute profile $\boldsymbol{\alpha}_l$. Her results showed that the PWKL yielded higher correct classification rates in comparison to the KL and SHE methods.

Lately, Kaplan, de la Torre, and Barrada (2015) argued that, by using the current estimate $\hat{\boldsymbol{\alpha}}_i^{(m)}$, PWKL assumes that the point estimate is a good summary of posterior distribution $\pi_i^{(m)}(\boldsymbol{\alpha}_l)$ and this assumption might not hold, especially in the early stages of testing. Therefore they proposed a modified version of the index, which is referred to as modified posterior weighted Kullback-Leibler (MPWKL). In the formulation of this item selection index, they consider posterior probabilities for all 2^K attribute profiles. MPWKL is formulated as

$$MPWKL_{ij}^{(m)} = \sum_{d=1}^{2^K}\left[\sum_{l=1}^{2^K}\left[\sum_{x=0}^{1}\log\left(\frac{P(X_j = x \mid \boldsymbol{\alpha}_d)}{P(X_j = x \mid \boldsymbol{\alpha}_l)}\right)P(X_j = x \mid \boldsymbol{\alpha}_d)\pi_i^{(m)}(\boldsymbol{\alpha}_l)\right]\pi_i^{(m)}(\boldsymbol{\alpha}_d)\right], \quad (10)$$

which does not require current estimate of attribute profile (i.e., $\boldsymbol{\alpha}_i^{(m)}$), rather, it considers the entire posterior distribution and weights them accordingly. It is clear from the formulation that this procedure requires an extra summation in comparison with the PWKL that makes estimation computationally cumbersome when *K* is large. However, Kaplan et al. (2015) reported that MPWKL provides higher correct classification rates in comparison to the PWKL.

Kaplan et al. (2015) has proposed another item selection rule using the G-DINA model discrimination index (GDI), which was proposed by de la Torre and Chiu (2010, 2015) as an index for empirical Q-matrix validation. The GDI, denoted as ς_j^2, is a measure of weighted variance of the success probabilities of an item given an attribute profile distribution (Kaplan et al., 2015). To define the index, let κ_j^* be the number of set of attributes required by item *j*; and $\boldsymbol{\alpha}_{lj}^*$ be the reduced set of attribute profiles formed by κ_j^* attributes such that $l = 1, 2, ..., \kappa_j^*$. Then, ς_j^2 is defined as

$$\varsigma_j^2 = \sum_{l=1}^{2^{K_j^*}}[P(X_{ij} = 1 \mid \boldsymbol{\alpha}_{lj}^* - \bar{P}_J]^2 \pi_i^{(m)}(\boldsymbol{\alpha}_l) \quad (11)$$

where $P(X_{ij} = 1 \mid \boldsymbol{\alpha}_{lj}^*)$ is the probability correct on item *j* given the reduced attribute profile, $\bar{P}_j$ is the mean success probability calculated as, $\sum_{l=1}^{2^{K_j^*}} \pi(\boldsymbol{\alpha}_{lj}^*)P(X_{ij} = 1 \mid \boldsymbol{\alpha}_{lj}^*)$ and $\pi_i^{(m)}(\boldsymbol{\alpha}_l)$ is the posterior probability of examinee *i* after administering *m* items.

Kaplan et al. (2015) emphasized the fact that GDI considers only the reduced attribute profiles, which makes it computationally more efficient (i.e., when $K = 6$ and $K_j^* = 2$, GDI computation is based on $2^2 = 4$ latent classes rather than $2^6 = 64$) in comparison to PWKL and MPWKL. Their simulation studies showed that both MPWKL and GDI item selection algorithms resulted in very similar classification rates, which were, in general, higher than the ones obtained through the PWKL as an item selection rule.

Item Exposure Rate and Overlap Rate

One of the vital practical considerations in adaptive testing is the test security. Either organized or individual item theft may seriously damage a high-stake and large-scale adaptive testing program. The features such as flexibility of examination times and testing on demand allow an examinee to communicate with other examinees about the topics and the items administered to them. Lee, Ip, and Fuh (2008) argued that because item selection algorithms tend to select optimal items, they often choose the most discriminating items, which are in return used more often than others. They further stated that this overexposure of some specific items might lead to information sharing among the examinees. Then, these overexposed items will eventually be public knowledge and will be answered by all examinees regardless of their ability levels.

The ratio between the number of times an item is administered and total number of examinees is called *item exposure rate*. According to Revuelta and Ponsoda (1998), item exposure rate depends on three elements of the measurement process which are; (a) psychometric properties of the items, (b) items available in the item pool, and c) the ability distribution of the examinees. They further stated that the strategies for controlling item exposure rate have two substantial goals; (1) preventing overexposure, and (2) increasing the use of infrequently or never-selected items. Although substantial number of research on item exposure rate control conducted thus far (i.e., McBride & Martin, 1983; Sympson&Hetter, 1985; Hetter&Sympson, 1997; Stocking & Lewis, 1998; van der

Linden & Reese, 1998; Chang & Ying, 1999; van der Linden, 2003; Boyd, Dodd &Fitspatric, 2003); Yi, Zang, and Chang (2008) claimed that the Sympson-Hetter (SH) and Stocking-Lewis (SL) exposure control procedures are commonly used for traditional CAT.

Let the size of the item pool be *J* and let the test length be *n*. Ideally, all the items in the item pool are expected to have the same exposure rate, which is calculated as

$$\overline{er} = \frac{n}{J}. \tag{12}$$

However, in applications, because of the psychometric characteristics of the items, some items are more likely to be administered. This disparity in item selection ratio produces an asymmetric item exposure rate distribution. In order to measure this asymmetry, Chang and Ying (1999) proposed a χ^2 distribution:

$$\chi^2 = \sum_{j=1}^{J} \frac{(er_j - er)^2}{\overline{er}}, \tag{13}$$

A low χ^2 value represents a low discrepancy between observed and ideal exposure rates (Lee et al. 2008). It should be noticed here that constraint on item exposure comes at a price. As Finkelman, Nering, and Roussos (2009) argued, all item exposure methods result in a reduction in psychometric precision. So, there is a trade-off between item exposure control and measurement precision. Although a good number of researches had performed to introduce item exposure rate control methods, there exist relatively few studies considering MAT and CD-CAT.

A series of studies proposed methods for item exposure control in unidimensional CAT. One of the popular exposure control methods was proposed by Sympson and Hetter (1985) which is an iterative procedure for controlling item exposure. To define the method, let *P(A)* be the probability of administering an item and let *P(S)* be the probability of selection an item. In this case, selecting an item does not mean to administer the item for sure. In the Sympson and Hetter (SH) method, an item exposure parameter, the probability of administering an item that had already been selected *P(A|S)*, is assigned to each item. If the parameter of a particular item is higher than the prespecified exposure rate, the item cannot be administered when it is selected. However, the main drawbacks of this method involved time-consuming iterations in calculating item exposure parameters and not being able to maintain the exposure rates of all items at or below the prespecified desired exposure rate (Barrada, Abad, & Veldkamp, 2009). Later Finkelman et al. (2009) proposed *the Generalized Sympson-Hetter Method*, which combines the SH exposure control method and *Kullback-Leibler* $(KL_j^{(m-1)})$ index, which is

$$KL_j^{(m-1)} \equiv \int_0 K_j(\hat{\boldsymbol{\theta}}^{(m-1)}, \boldsymbol{\theta}) f^{(m-1)}(\boldsymbol{\theta}) d(\boldsymbol{\theta}), \tag{14}$$

where $K_j(\hat{\boldsymbol{\theta}}^{(m-1)}, \boldsymbol{\theta})$ is defined in *equation 6*, $f^{(m-1)}(\boldsymbol{\theta})$ is the posterior density of $\boldsymbol{\theta}$ after *m-1* items. The overall goal of this method is to keep administration rate below a prespecified *desired exposure rate (r)*. Let *A* to be the administration, and $\pi_{j,f_0(\boldsymbol{\theta})}(A)$ the probability of administering item *j* with respect to the prior distribution $f_0(\boldsymbol{\theta})$. Then, the method sets a relation

$$\pi_{j,f_0(\boldsymbol{\theta})}(A) \le r \tag{15}$$

for all *j*. To succeed in keeping the *inequality 15*, items are initially ranked based on some psychometric criterion (e.g., based on $KL_j^{(m-1)}$) and the item maximizing $KL_j^{(m-1)}$ becomes a *candidate* item, which is administered with a conditional probability of $P_j(A/S)$, where *A* and *S* indicate the number of administration and number of selection, respectively. The algorithm searches for a new candidate item, as the current candidate item is not allowed for administration. This process is carried on until a candidate is approved for administration.

In general, SH method requires computation of $P_j(S)$ and $P_j(A)$, which can be carried out regardless of the dimensionality of the psychometric criterion for item selection. Thus, the SH method is applicable to both unidimensional and multidimensional cases. However, according to Finkelman et al. (2009), there are two substantial differences; (*1*) a multivariate prior distribution must be used to generate the abilities for simulees in MAT to set threshold *r* and (*2*) the psychometric index differs for these two cases which means that although the Fisher information can be used in unidimensional-CAT item selection algorithm; multidimensional item selection indices have to be employed in MAT.

Security can be an issue with one of the two ways: (*1*) item theft by an organized group and (*2*) peer-to-peer communication. Stocking-Lewis (1998) pointed out that if the test security threat of peer-to-peer communication is

of concern, examinees with similar abilities are most likely to share information about the contents and items of a test. Consequently, it requires exposure control conditional on ability. They presented a method to control exposure rate conditional upon examinee ability level (i.e., Stocking-Lewis [SL]). This method is applicable for unidimensionalCAT procedures. Later, for multidimensional cases, Finkelman et al. (2009) introduced a generalized version of SL method known as generalized Stocking-Lewis (GSL) method.
The SL method sets the exposure control boundary along a set of U discrete θ-levels, $\theta_{1,}...,\theta_{U,}$ which approximately satisfies the boundary for all ability levels. Notice that the constraint is not implemented over all ability levels, it rather is implemented over U meaningfully selected values of θ. Therefore, this method establishes the relation below for all j and all $u = 1,...,U,$

$$\pi_{j,\boldsymbol{\theta}_U}(A) \leq r \tag{16}$$

Inequality16 is used as surrogate for the desired relation of

$$\pi_{j,\boldsymbol{\theta}}(A) \leq r. \tag{17}$$

The computed proportion of selection and administration for each θ_u are then denoted as $P_{j,\theta u}(S)$ and $P_{j,\theta u}(A)$, respectively. When item j is selected as a candidate item, the item exposure control parameter $P_{j,\theta u}(A|S)$ that corresponds to θ_u is used as it is closest to the current theta estimate.

In multidimensional cases, direct extension of SL method would require *inequality 16* for all j and all u over an D-dimensional grid where ability is represented by a vector $\boldsymbol{\theta}$ $(\theta_1,..., \theta_D)$. Finkelman et al. (2009) discussed that because the number of $\boldsymbol{\theta}$ values requiring *inequality 16* exponentially increases as the number of dimensions increases, complete crossing of discrete values in the grid is intractable even when D is moderate. The GSL method proposed by Finkelman et al. (2009), ideally, maintains a good number of quadrature points regardless of D. Instead of using *inequality16*, this method performs an exposure control conditional on θ^*, where θ^* is considered to be a function of $\boldsymbol{\theta}$. However θ^* is a scalar such that $\theta^* = \lambda' \boldsymbol{\theta}$ where λ is a set of weights. Therefore, item exposure control is conditional on a linear combination of D dimensions of $\boldsymbol{\theta}$. Then the established relationship between probability of administering item j at θ^* and desired exposure rate is

$$\pi_{j,\boldsymbol{\theta}^*}(A) \leq r, \tag{18}$$

for all j and all values θ^*. All operational steps are same for SL and GSL methods, the only difference is that the conditioning variable in GSL becomes θ^* rather than θ.

For item exposure control in the context of cognitive diagnosis, Wang, Chang, and Huebner (2011) proposed a restrictive stochastic item selection method, which is a modified version of the *progressive method* (PM; Revuelta, 1995) for traditional CAT procedures. The PM method weights items based on an information component with a random part, where as the test progresses the relative impact of information increases. Wang et al. (2011) modified the PM by adding a stochastic component such that it would not always select the item producing the highest information at the current stage. The restrictive progressive (RP) method that employs PWKL information is defined as

$$RP-PWKL_j = \left(1-\frac{exp_j}{r}\right)\left[\left(1-\frac{m-1}{n}\right)R_j + \frac{PWKL_j\beta_{(m-1)}}{n}\right], \tag{19}$$

where exp_j is the preliminary exposure rate of item j, r is the pre-defined exposure control rate, $m-1$ is the number of administered items, n is the test length, R_j is a random value that is drawn from a uniform distribution between 0 and $PWKL_j$ for the items in the pool, and β is an arbitrary number to control the balance between the test security and estimation accuracy. Smaller β tends to produce more secure test, however, tests with small β yield less accurate estimation.

Kaplan (2016) has recently incorporated the RP method for exposure control with the GDI and MPWKL item selection rules for CD-CAT application. The notation below is the RP method representation where Δ_j indicates the information on item j (e.g., $MPWKL_j$ and ζ_j)

$$RP-\Delta_j = \left(1-\frac{exp_j}{r}\right)\left[\left(1-\frac{m-1}{n}\right)R_j + \frac{\Delta_j\beta_{(m-1)}}{n}\right], \tag{20}$$

Notice that the current form of RP method is applicable to fixed length (i.e., n) tests. In his study, Kaplan (2016) modified it such that minimum of the maximum of the posterior distribution is used as the test termination rule. Then, the modified item selection index incorporating the RP method is

$$RP-\Delta_j=\left(1-\frac{exp_j}{r}\right)\left[f(x)R_j+\Delta_j\beta\frac{\pi(\hat{\boldsymbol{\alpha}}_l\mid \boldsymbol{X}_j)}{P}\right], \quad (21)$$

Where $f(x)=\min\left(0,1-\frac{\pi(\hat{\boldsymbol{\alpha}}_l\mid \boldsymbol{X}_j)}{P}\right)$ and P is the predetermined minimax value.

Stopping Rule

A rule determining when to stop administering items adaptively is referred to as *stopping rule*. Stopping rules can be set for fixed- and variable-length tests (Reckase, 2009; Frey & Seitz, 2009). A stopping rule for a fixed-length test sets a predetermined number for item administration. Once the test reaches this pre-specified length, it is terminated and final ability estimate is computed. A variable-length test is terminated based on a pre-specified standard error of the ability estimate. In other words, stopping rule in a variable-length test becomes a statistical criterion on the measurement precision. In variable-length tests, the number of items to be administered depends on the location of the examinee in the latent ability-space, consistency of examinee responses, and the information provided by the item pool relative to the ability level (Reckase, 2009) or his/her attribute profile in CDM cases.

Reckase (2009) argued that test length could be determined based on some practical considerations such as testing time, or by taking the average of the variable length tests. Wang, Chang, and Boughton (2012) argued that the literature on MAT focuses on the stopping rules for fixed-length tests, which provide less accurate ability estimates for examinees whose ability locations are substantially different than the average difficulty level of the item bank.

Despite the fact that fixed- and variable-length stopping rules were well explored in the unidimensional CAT, Wang et al. (2012) noted that because the precision of multiple ability dimensions should be considered simultaneously, these well-defined stopping rules cannot straightforwardly be generalized to multidimensional adaptive testing situations. They further discussed that in order to set a stopping criterion for MAT, firstly, an index such as generalized variance, total variance, or entropy should be set to quantify the estimation accuracy of θ-vector. Moreover, Kaplan and de la Torre (in press) is among the limited variable-length CD-CAT studies where they use the minimum of the maximum of the posterior distribution as the test termination rule.

DISCUSSION

This paper intended to compile and highlight the recent developments in CAT procedures by reviewing the generalization and/or modification of traditional CAT components for MAT and CD-CAT applications. This paper also intended to provide researchers and practitioners with pragmatic information such that they can use this information toward their own research and application purposes.

When traditional CAT is intended, unidimensional items need to be written and calibrated in accordance with unidimensional IRT models. Similarly, item development and item calibration need to be in accordance with the MIRT models when MAT is intended. Item pool development for CD-CAT can be much more challenging, as CDM applications require a Q-matrix specifying the item-by-attribute associations. Construction of a Q-matrix requires collaboration among measurement experts and content experts. The misspecification of the Q-matrix can reduce the credibility of CD-CAT applications.

Generalization of traditional CAT procedures for MAT is challenging because the new procedures needs to be tractable and computationally manageable as the number of dimensions increases. Further, due to the discrete nature of CDMs, not all conventional item selection rules and exposure control rates can be modified for CD-CAT implementations. For example, item selection algorithms based on Fisher information cannot be applied in the context of cognitive diagnosis (Xu, Chang, & Douglas, 2003). Alternatively, the MPWKL and GDI can be used as item selection algorithms in CD-CAT.

Another vital practical consideration in adaptive testing is item exposure rates. It should be noticed that constraints on item exposure comes at a price because item selection algorithms can no longer use the most informative items in every step. As discussed by Finkelman, Nering, and Roussos (2009), employment of item exposure control methods on the item selection algorithms results in reduction in estimation accuracy. In other words, there is a trade-off between item exposure control and measurement precision. There is not much research conducted for item exposure control in CD-CAT applications. Wang, Chang, and Huebner (2011) proposed a restrictive progressive item selection method, which is a modified version of PM (Revuelta, 1995; Revuelta & Ponsoda, 1998) for traditional CAT applications. Although the RP was proposed for and used with fixed-length tests, it was recently modified for variable-length tests.

Another practical consideration in adaptive testing is the starting point (i.e., with which item to start), for which there is not enough research for CD-CAT. Similarly, impact of type of attribute profile estimation (i.e., MLE, MAP, and EAP) may impact item selection and consequently exposure rates differently. Impact of ability estimation methods and the prior distribution on CD-CAT can be further research topics.

REFERENCES

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

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

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

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34,* 115-130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76,* 179-199.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educative, 20,* 89-97.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39(1),* 1-38.

Finkelman, M., Nering, M. L., & Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement*, *46*, 84-103.

Flaugher, R. (2000). Item pools. In H. Wainer (Ed.),*Computerized adaptive testing: A primer* (2nd ed., pp. 37-59). Mahwah, NJ: Lawrance Erlbaum Associates, Inc, Publishers.

Frey, A., & Seitz, N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35*, 89-94.

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.

Kaplan, M. (2016). *New item selection and test administration procedures for cognitive diagnosis computerized adaptive testing*. Unpublished doctoral dissertation, Rutgers, The State University of New Jersey, NJ.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied psychological measurement, 39,* 167-188.

Lee, Y., Ip, E. H., &Fuh, C. (2008). A strategy for item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement*, *68*, 215-232.

Meijer, R. R., &Nering, M. L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement*, *23*, 187-194.

Reckase, M. D. (2009). *Multidimensional item response theory, statistics for social and behavioral sciences*. New York, NY: Springer Science Business Media, LLC.

Revuelta, J. (1995). *El control de la exposicion de los items en tests adaptativosinformatiozados [Item exposure control in computerized adaptive tests]*. Unpublished master's dissertation, Universidad Autonoma de Madrid, Spain.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 311-327.

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-art. *Measurement, 6,* 219-262.

Segall, D. O. (1996). Multidimensional adaptive testing.*Psychometrika*, *61*, 331-354.

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

Wainer, H., &Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 61-100). Mahwah, NJ: Lawrance Erlbaum Associates, Inc, Publishers.

Wang, C., Chang, H.-H., &Boughton, K. A. (2011). Kullback-Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, *76*, 13-39.

Wang, C., Chang, H.-H., &Boughton, K. A. (2012). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*, 99-122.

Xu, X., Chang, H.-H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Yi, Q., Zhang, J., & Chang, H.-H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. *Applied Psychological Measurement*, *32*, 543-558.

