Article

A Blocked-CAT Procedure for CD-CAT

Applied Psychological Measurement
2020, Vol. 44(1) 49–64

DOI: 10.1177/0146621619835500
journals.sagepub.com/home/apm
SAGE

Mehmet Kaplan[1] and Jimmy de la Torre[2]

Abstract

This article introduces a blocked-design procedure for cognitive diagnosis computerized adaptive testing (CD-CAT), which allows examinees to review items and change their answers during test administration. Four blocking versions of the new procedure were proposed. In addition, the impact of several factors, namely, item quality, generating model, block size, and test length, on the classification rates was investigated. Three popular item selection indices in CD-CAT were used and their efficiency compared using the new procedure. An additional study was carried out to examine the potential benefit of item review. The results showed that the new procedure is promising in that allowing item review resulted only in a small loss in attribute classification accuracy under some conditions. Moreover, using a blocked-design CD-CAT is beneficial to the extent that it alleviates the negative impact of test anxiety on examinees' true performance.

Keywords

item review, CD-CAT, block design

Introduction

The debate over whether to provide examinees the option to review items and change answers during test administration in the context of computerized adaptive testing (CAT) has continued for several years. Test takers and test developers have different attitudes toward such an option. Allowing test takers to review their answers can reduce test anxiety, and thus increase test scores legitimately. However, test developers are reluctant to provide this option to test takers for several reasons (Vispoel, Clough, & Bleiler, 2005). The two most common concerns are decreased testing efficiency (e.g., longer testing times) and illegitimate score gains resulting from test-taking strategies.

Wise (1996) classified score gains as legitimate and illegitimate. The former refers to a score gain in which examinees, who possess the required knowledge to answer an item correctly, increase their scores after they review and change their answers. The latter refers to a score gain in which examinees, who do not possess the required knowledge, are somehow able to answer the item correctly because, for example, they get a clue from other items. On one hand, when examinees obtain legitimate score gains by using item review and answer change, the validity

[1]Artvin Coruh University, Turkey
[2]The University of Hong Kong, Hong Kong

Corresponding Author:
Mehmet Kaplan, Faculty of Education, Artvin Coruh University, 08000 Artvin, Turkey.
Email: mehmet.kaplan2@gmail.com

of the test increases, and therefore, inferences from the test results become more meaningful and appropriate. On the other hand, providing this option can lengthen testing times and decrease testing precision due to illegitimate score gains (Wise, 1996).

To date, no research has been done to investigate the impact of item review and answer change in the context of cognitive diagnosis CAT (CD-CAT). The goal of this study is to propose a new CD-CAT administration procedure. In this procedure, a block of items is used as a unit of administration. Because there are no difficulty parameters to partition the test into blocks in cognitive diagnosis, blocking is performed based on the information using item selection indices. Using this design, examinees have an opportunity to review and change their answers within the block.

Item Review

There is a common belief among examinees and college instructors that changing initial responses to items about which examinees are uncertain might lower the examinees' test scores (Benjamin, Cavell, & Schallenberger, 1984). Contrary to this belief, researchers have shown that most examinees changed their answers when they were allowed to, and those changes were generally from incorrect to correct. Therefore, those who made changes improved their test scores (Benjamin et al., 1984). Moreover, the results of those studies showed that examinees changed their answers for only a very small percentage of answers, but a large number of examinees made changes for at least a few items.

Having the option to review items and change answers during test administration has several benefits for examinees. This option is beneficial for correcting careless errors, misreading of items, temporary lapses in memory, reconceptualization of answers to previously administered items, and test validity (Vispoel, 1998). In addition, this option can relax the testing environment for examinees who have high test anxiety (Vispoel, 2000). However, reviewable CAT requires more complicated item selection algorithms because most of the item selection algorithms in CAT rely on a provisional ability estimate to select the next item. Changing the answer of an item during the test administration can make the succeeding items no longer appropriate for estimating the ability level (Yen, Ho, Liao, & Chen, 2012). In addition, more flexible structures must be developed for examinees' diverse review styles. For example, some examinees like to review item by item sequentially; however, others mark some of the items and review them later (Wise, 1996).

As noted before, examinees can have illegitimate score increases even if they do not possess the required knowledge to answer the item correctly. This happens when they can get a clue from the characteristics of other items on the test, or they can use specific testing strategies (e.g., the Wainer strategy, the Kingsbury strategy, and the generalized Kingsbury strategy). The use of cheating strategies in CD-CAT may be possible, but its impact may be limited. The Wainer and Kingsbury strategies are based on the difficulty parameter; however, several factors can affect the item selection in CD-CAT, not just item difficulty, because of the multidimensional nature of CDMs. For example, consider an examinee whose true α is (1,1,0,0). This examinee would be administered an item whose q-vector is (1,1,0,0), (0,0,0,1), or (0,0,1,0) if the model is DINA (deterministic input noisy ''and'' gate model), or an item whose q-vector is (0,0,1,1), (1,0,0,0), or (0,1,0,0) if the model is DINO (deterministic input noisy ''or'' gate model). In this case, examinee's true $\boldsymbol{\alpha}$ and the model have an impact on the item selection. Other factors (e.g., item quality and availability of all possible q-vectors) can also affect the item selection.

Stocking (1997) proposed a blocked-design CAT in which items were grouped into blocks *on the fly* or *during* the actual test administration, and examinees were allowed to review items

within a block. The impact of the Wainer strategy with and without item review on the test was investigated. She noted that the bias in the estimates and the standard errors were at acceptable levels using this method. Moreover, researchers have also shown that testing time increased by only 5% to 11% on average with the majority of examinees indicating that they had adequate opportunity for item review and answer change in the blocked-design CAT (Vispoel et al., 2005). However, Han (2013) noted that the blocked-design CAT still did not allow test takers to skip items. As an alternative, he proposed an item pocket method in which examinees had the option to skip items for a certain number in addition to reviewing items and changing answers.

Another procedure, which allows item review and answer change is multistage testing (MST). In MST, the test adaption occurs at the item set level or the testlet level instead of the item level, and items are preassembled into modules *prior* to the test administration. However, Zheng and Chang (2015) recently proposed on-the-fly MST (OMST) procedure in which blocks of items are assembled *on the fly*, and the adaptation occurs between the blocks. The results showed that the estimation accuracy using OMST was comparable with CAT and MST, and because OMST and CAT constructed highly diverse tests, those two provided better test security compared with MST. Similar to the Zheng and Chang (2015)'s work, items in the present study were selected one by one to maximize the information at the latest updated posterior distribution of attribute vectors; however, different from OMST, CAT was performed in the context of cognitive diagnosis.

MST applications beyond unidimensional models are also limited. However, most models used for diagnostic testing require a multidimensional latent trait. More specifically, cognitive diagnosis modeling requires the estimation of a set of discrete attributes, an attribute vector, which consists of several dimensions. Constructing the blocks in MST for cognitive diagnosis can be challenging because there are no difficulty parameters for every relevant dimension in CDMs. MST using CDMs (CD-MST) was first noted by von Davier and Cheng (2014). They discussed several heuristics that can be applied in the CD-MST selection stage, and suggested Shannon entropy for selecting the next block of items. However, the authors did not investigate further how a block of items in the context of cognitive diagnosis for MST can be created.

CDMs

CDMs aim to determine whether or not examinees have a mastery of a set of typically binary attributes, which represents the presence or absence of the skill or attribute. Let $\boldsymbol{\alpha}_i = \{\alpha_{ik}\}$ be the attribute vector of examinee i, where $i = 1, 2, \ldots, N$ examinees, and $k = 1, 2, \ldots, K$ attributes. The k th element of the vector is 1 when the examinee has mastered the k th attribute, and it is 0 when the examinee has not. In cognitive diagnosis, examinees are classified into latent classes based on the attribute vectors. Each attribute vector corresponds to a unique latent class. Therefore, K attributes create 2^K latent classes or attribute vectors. Similarly, the responses of the examinees to J items are represented by a binary vector. Let $\boldsymbol{X}_i = \{x_{ij}\}$ be the i th examinee's binary response vector for a set of $j = 1, 2, \ldots, J$ items. The required attributes for an item are represented in a Q-matrix (K. Tatsuoka, 1983), which is a $J \times K$ matrix. The element of the j th row and the k th column, q_{jk}, is 1 if the k th attribute is required to answer the j th item correctly, and it is 0 otherwise.

To date, a variety of general CDMs has been proposed to increase their applicability. For example, the log-linear CDM (Henson, Templin, & Willse, 2009), the general diagnostic model (von Davier, 2008), and the generalized deterministic inputs, noisy ''and'' gate model (G-DINA; de la Torre, 2011) are examples of general CDMs. The G-DINA model relaxes some of the strict assumptions of the DINA model (de la Torre, 2009), and it partitions examinees into $2^{K_j^*}$ groups, where $K_j^* = \sum_{k=1}^{K} q_{jk}$ is the number of required attributes for item j. A few

constrained CDMs can be derived from the G-DINA model using different constraints (de la Torre, 2011). These include the DINA model, the DINO model (Templin & Henson, 2006), and the additive CDM (*A*-CDM; de la Torre, 2011).

CAT

CAT has become a popular tool in testing because it allows examinees to receive different tests, with possibly different lengths. Compared to paper-and-pencil tests, the mode of test administration changes from paper to computer, and the test delivery algorithms change from linear to adaptive (van der Linden & Pashley, 2010). Therefore, it provides a tailored test for each examinee, and better ability estimation with shorter test lengths (Meijer & Nering, 1999). One of the crucial components of CAT is the item selection methods. In traditional CAT, item selection methods based on the Fisher information are widely used (Lord, 1980); however, those methods are not applicable in CD-CAT because the equivalent latent variables in cognitive diagnosis are discrete. This issue was first noted by C. Tatsuoka (2002) and C. Tatsuoka and Ferguson (2003), and they showed that minimizing the expected Shannon entropy to classify finite and partially ordered sets outperformed the Kullback–Leibler (K-L) information approaches. Later, X. Xu, Chang, and Douglas (2003) proposed two item selection indices for CD-CAT based on the K-L information and Shannon entropy procedure. The efficiency of these indices was compared with random selection using a simulation study. The results of their study showed that both indices outperformed random selection in terms of attribute classification accuracy. Cheng (2009) proposed two item selection indices in CD-CAT, and both were based on the K-L information, namely, the posterior-weighted K-L index (PWKL) and hybrid K-L index (HKL). The results showed that the new indices performed similarly, and both had higher classification rates than the K-L and Shannon entropy procedure. Wang (2013) proposed a mutual information (MI) method as an item selection index in CD-CAT. The results showed that the MI outperformed the PWKL only with short test lengths. In comparison, the G-DINA model discrimination index (GDI) outperformed the PWKL across the different test lengths, and the increase in classification rates was higher compared with that of the MI (Kaplan, de la Torre, & Barrada, 2015). Based on the results of these two studies, it can be expected that the GDI will also outperform the MI.

Recently, Kaplan et al. (2015) proposed two new item selection indices, namely, modified PWKL (MPWKL) and the GDI for CD-CAT. The results showed that the two new indices performed very similarly and had higher attribute classification rates compared with the PWKL. In addition, the GDI had the shortest administration time. The MPWKL can be computed as

$$\text{MPWKL}_{ij}^{(t)} = \sum_{d=1}^{2^K}\left[\sum_{c=1}^{2^K}\left[\sum_{x=0}^{1} log\left(\frac{P\left(X_j=x|\boldsymbol{\alpha}_d\right)}{P\left(X_j=x|\boldsymbol{\alpha}_c\right)}\right)P\left(X_j=x|\boldsymbol{\alpha}_d\right)\boldsymbol{\pi}_i^{(t)}(\boldsymbol{\alpha}_{\mathbf{c}})\right]\boldsymbol{\pi}_i^{(t)}(\boldsymbol{\alpha}_{\mathbf{d}})\right]. \quad (1)$$

In this modified version does not require estimating the attribute vector $\boldsymbol{\alpha}_i^{(t)}$ as it is in the PWKL.

The GDI measures the posterior-weighted variance of the probabilities of success of an item given a particular attribute distribution. To give a summarized definition of the index, define $\boldsymbol{\alpha}_{cj}^*$ as the reduced attribute vector consisting of the first K_j^* attributes, for $c = 1, \ldots, 2^{K_j^*}$. Also, define $P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^*)$ as the success probability on item j given $\boldsymbol{\alpha}_{cj}^*$. The GDI for item j is defined as

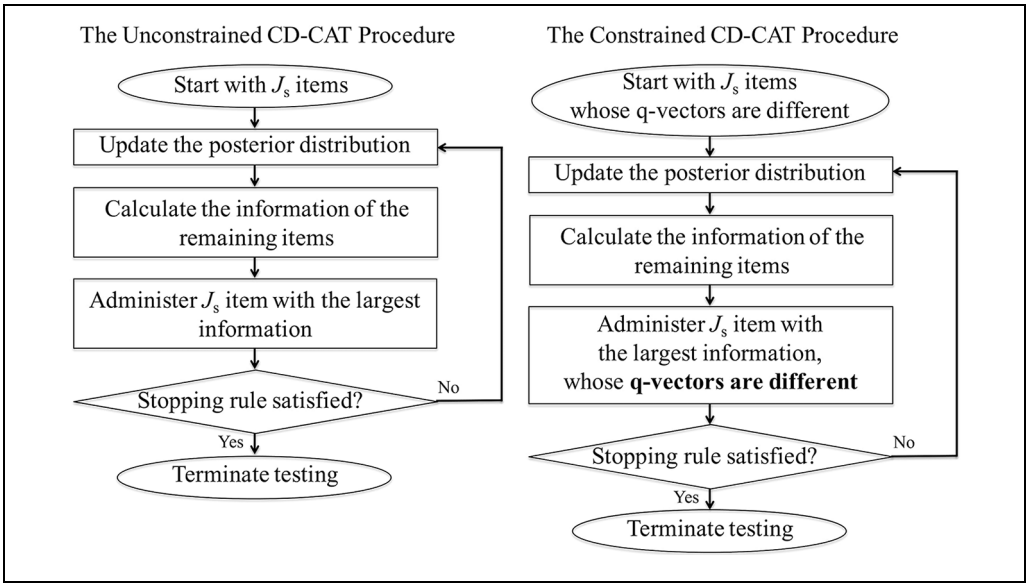

Figure 1. The new CD-CAT procedures.
Note. CD-CAT = cognitive diagnosis computerized adaptive testing.

$$\varsigma_j^2 = \sum_{c=1}^{2^{K_j^*}} \left[P\left(X_{ij} = 1 | \boldsymbol{\alpha}_{cj}^*\right) - \bar{P}_j \right]^2 \boldsymbol{\pi}_i^{(t)}\left(\boldsymbol{\alpha}_{cj}^*\right), \tag{2}$$

where $\bar{P}_j = \sum_{c=1}^{2^{K_j^*}} \boldsymbol{\pi}(\boldsymbol{\alpha}_{cj}^*) P(X_{ij} = 1 | \boldsymbol{\alpha}_{cj}^*)$ is the mean success probability and $\boldsymbol{\pi}_i^{(t)}(\boldsymbol{\alpha}_{cj}^*)$ is the posterior probability of the reduced attribute vector. In this article, the PWKL and GDI were used as item selection indices with the new CD-CAT administration procedure.

Simulation Study

The goal of this study is to investigate how different blocking versions affect the efficiency of the item selection indices in terms of classification accuracy. Different versions of the blocked-design CAT, namely, unconstrained, constrained, and hybrid can be formulated. In the unconstrained version, a block of J_s items is randomly administered first to calculate the examinee's posterior distribution, which is needed to compute the item selection indices. The most informative J_s items remaining in the pool are administered together, and the posterior distribution is updated. This cycle continues until the test termination rule has been satisfied. The unconstrained version of the new procedure is shown in the left panel of Figure 1.

In the constrained version, items are selected based on some constraints, for example, a constraint on the q-vectors. One such constraint is to require that none of the items within the same block can have the same q-vector. A previous study showed that item selection indices did not provide additional information when the same type of items (e.g., the same q-vector) were administered repeatedly (Kaplan et al., 2015). As with the unconstrained version, the first J_s items are randomly selected from the pool; however, the q-vectors of the items are constrained to be different from each other. Again, the posterior distribution is calculated, and the next J_s items are selected from the pool based on the item selection index, with the same constraint. This

procedure continues until the termination criterion has been satisfied. The right panel of Figure 1 shows the constrained version of the proposed procedure.

In the hybrid versions, some blocks are constrained, whereas others are not. In the first version (i.e., Hybrid-1), a block of J_s items with the same constraint as in the constrained version is administered during the first half of the test, and the second half of the test is performed without constraint. In the second version (i.e., Hybrid-2), no constraint is applied during the first half of the test, but the constraint is applied in the second half. A simulation study was designed to examine the viability of the new procedure. The impact of different factors on the attribute classification accuracy of the different versions of the new procedure in conjunction with two item selection indices was investigated.

Design

Data generation. Two levels of item quality, namely, low-quality (LQ) and high-quality (HQ), were considered in the data generation. However, it should be noted that these two terms were used exclusively for this study and in other studies (e.g., in de la Torre, Hong, & Deng, 2010), $P(\mathbf{0})$was drawn from $U(0.20, 0.30)$ and $U(0.05, 0.15)$ for LQ items and HQ items, respectively), they have been defined differently. For the purposes of this study, LQ and HQ can also be viewed as less discriminating and more discriminating, respectively. For LQ items, the lowest and highest success probabilities (i.e., $P(\mathbf{0})$ and $P(\mathbf{1})$) were generated from uniform distributions, $U(0.15, 0.25)$ and $U(0.75, 0.85)$, respectively; and for HQ items, $P(\mathbf{0})$ and $P(\mathbf{1})$ were generated from uniform distributions, $U(0.00, 0.20)$ and $U(0.80, 1.00)$, respectively. Item responses were generated using three reduced models: DINA model, DINO model, and A-CDM. The probability of success was set as discussed above for the DINA and DINO. In addition to these probabilities, the intermediate success probabilities were obtained by allowing each required attribute to contribute equally for the A-CDM. The number of attributes was fixed at $K = 5$.

A more efficient simulation design from Kaplan et al. (2015)'s paper was also used in this study. One representative of each attribute vector (i.e., no mastery, mastery of a single attribute only, mastery of two attributes only, and so forth) was used, and the appropriate weights are applied. Two thousand examinees were generated for each attribute vector, resulting in a total of 12,000 examinees.

Item pool and item selection methods. The Q-matrix was created from $2^K - 1 = 31$ possible q-vectors, each with 40 items. The pool then had a total of 1,240 items. Fixed test lengths were used as a test termination rule. The test lengths were set to 8, 16, and 32 items, and the size of the blocks was set to J_s = 1, 2, and 4. It should be noted that $J_s = 1$ corresponds to traditional CD-CAT administration. Three item selection indices were considered: PWKL, MPWKL, and GDI. For greater comparability, a uniform distribution of the attribute vectors was used as the prior distributions for the indices across all conditions. In the case of the PWKL, when the estimate of the attribute vector was not unique, the provisional attribute vector estimate was chosen from the modal attribute vectors.

To compare the efficiency of the indices, the means of the correct attribute classification (CAC) rate and the correct attribute vector classification (CVC) rate were computed for each condition. For each of the six attribute vectors considered in the design, let α_{ikl} and $\hat{\alpha}_{ikl}$ be the k th true and estimated attribute in attribute vector l, $l = 0, 1 \ldots 5$, for examinee i, respectively. The CAC and CVC rates were computed as

Table 1. The CVC Rates Using the DINA Model.

			PWKL				MPWKL				GDI			
IQ	J	J_s	UC	H1	H2	C	UC	H1	H2	C	UC	H1	H2	C
LQ	8	1		0.40				0.58				0.59		
		2	0.25	0.36	0.31	0.41	0.55	0.55	0.57	0.57	0.54	0.54	0.55	0.55
		4	0.20	0.31	0.28	0.39	0.51	0.50	0.48	0.53	0.51	0.48	0.48	0.54
	16	1		0.72				0.84				0.84		
		2	0.55	0.70	0.63	0.74	0.83	0.82	0.82	0.82	0.83	0.82	0.83	0.83
		4	0.45	0.65	0.59	0.72	0.76	0.78	0.78	0.79	0.76	0.78	0.78	0.80
	32	1		0.96				0.99				0.99		
		2	0.90	0.96	0.93	0.96	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
		4	0.80	0.94	0.90	0.95	0.97	0.98	0.97	0.97	0.97	0.97	0.97	0.97
HQ	8	1		0.83				0.98				0.98		
		2	0.55	0.70	0.64	0.78	0.98	0.98	0.98	0.97	0.98	0.98	0.97	0.97
		4	0.35	0.62	0.54	0.73	0.94	0.95	0.94	0.96	0.97	0.97	0.96	0.97
	16	1		1.00				1.00				1.00		
		2	0.95	0.99	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		4	0.83	0.96	0.96	0.98	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	32	1		1.00				1.00				1.00		
		2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. CVC = correct attribute vector classification; DINA = deterministic input noisy "and" gate; IQ = item quality; J = test length; J_s = block size; PWKL = posterior-weighted K-L index; MPWKL = modified posterior-weighted K-L index; GDI = G-DINA model discrimination index; UC = unconstrained; H1 = Hybrid-1; H2 = Hybrid-2; C = constrained version; LQ = low quality; HQ = high quality.

$$CAC_l = \frac{1}{2{,}000} \sum_{i=1}^{2{,}000} \sum_{k=1}^{5} I[\alpha_{ikl} = \hat{\alpha}_{ikl}] \quad \text{and} \quad CVC_l = \frac{1}{2{,}000} \sum_{i=1}^{2{,}000} \prod_{k=1}^{5} I[\alpha_{ikl} = \hat{\alpha}_{ikl}], \tag{3}$$

where I is the indicator function. Using appropriate weights (described below), the CAC and the CVC were computed assuming the attributes were uniformly distributed for the fixed test-length conditions. This study focused on uniformly distributed attribute vectors; however, the sampling design of the study can allow for results to be generalized to different distributions of the attribute vectors (e.g., de la Torre & Douglas, 2004). Thus, the results based on the six attribute vectors had to be weighted appropriately. For $K = 5$, the vector of the weights were 1/32, 5/32, 10/32, 10/32, 5/32, and 1/32, which represented the proportions of zero-, one-, two-, three-, four-, and five-attribute mastery vectors among the 32 attribute vectors, respectively (for more details, see Kaplan et al., 2015).

Results

Classification accuracy. For all conditions, the CAC rates were, as expected, higher than the CVC rates, but the measures showed similar patterns. Thus, only the CVC rates are discussed. The CVC rates under the different factors are presented in Table 1 for DINA model. Additional results for DINO and *A*-CDM are shown in Tables 1 and 2 in online Supplemental Appendix A, respectively. In Kaplan et al. (2015), differences in the classification rates were evaluated using different cut points to better summarize the findings. Similarly, in this study, differences in the CVC rates were evaluated using a cut point of 0.05. Differences of 0.05 and below were considered negligible, but differences above 0.05 were considered substantial. In addition,

eight-item tests were considered as short, 16-item tests were considered as medium length, and 32-item tests were considered as long tests.

Several results can be noted. First, using different blocking versions with the MPWKL and the GDI resulted in different classification accuracies based on the factors; however, those differences were mostly negligible. However, there are some exceptions, where the differences among the blocking versions were substantial. Using a large block size (i.e., $J_s = 4$) and the DINA model, the constrained version yielded substantially higher CVC rates compared with the Hybrid-2 version when short-length tests and LQ items were used; using a large block size (i.e., $J_s = 4$) and the DINO model, the unconstrained version yielded substantially lower CVC rates compared with the other blocking version when short-length tests were used regardless of the item quality; and using the same factors, the constrained version yielded substantially higher CVC rates compared with the unconstrained and Hybrid-1 versions when medium-test lengths were used with LQ items.

Second, for the DINA and DINO models, the MPWKL and the GDI resulted in higher classification rates compared with the PWKL especially using LQ items with short- and medium-test lengths, and HQ items with short-test lengths. It was expected because these two indices did not use estimated attribute vectors in the calculation of the information, and therefore, they produced better CVC rates. Moreover, the maximum difference between the PWKL and the other two indices occurred when HQ items, short-test lengths, and large block size (i.e., $J_s = 4$) were used. For example, the differences between the PWKL and the MPWKL and the PWKL and the GDI were 0.59 and 0.62 for the unconstrained version with the DINA model, respectively. For the DINO model, the MPWKL and GDI performed very similarly except that the difference in CVC rates were substantial (i.e., 0.8) when the unconstrained version was used with a large block size, short-test lengths, and HQ items.

Third, for the A-CDM, the indices generally performed similarly and the differences in the classification rates were negligible. However, there were some conditions where the differences were substantial. The MPWKL and the GDI yielded substantially higher CVC rates than the PWKL when short-test lengths, HQ items, and large block size (i.e., $J_s = 4$) were used regardless of the blocking version. For example, the MPWKL and the GDI yielded the CVC rates 16% and 18% higher than the PWKL when the unconstrained version was used, respectively. However, the PWKL yielded substantially higher CVC rates than the MPWKL and the GDI in conditions—when short-test lengths, LQ items, and medium block size (i.e., $J_s = 2$) were used with the Hybrid-2 and constrained versions.

Last, using the PWKL with the DINA and the DINO as the generating models, the constrained version with the PWKL had the best classification accuracy among the other blocking versions, whereas the unconstrained version had the worst classification accuracy regardless of the block size, test length, and item quality except on the 32-item test with the HQ item conditions where the classification accuracy was perfect. In the following section, the impact of the factors (i.e., block size, test length, and item quality) on the CVC rates is discussed.

The impact of block size. Regardless of the generating model and using the MPWKL and the GDI as an item selection index, increasing the block size resulted in negligible differences in the classification rates for LQ items with long-test lengths and HQ items with any test length. However, there are some conditions (i.e., LQ items with short- and medium-test lengths) where increasing the block size yielded substantially lower classification rates. For example, for the DINA model, increasing the block size from two to four resulted in substantially lower CVC rates when LQ items and short-test length were used with the Hybrid-1 and Hybrid-2 versions; and increasing the block size from two to four resulted in substantially lower CVC rates when LQ items and medium-test length were used with the unconstrained version. For the DINO model, increasing the block size from one to two resulted in substantially lower CVC rates

when LQ items and short-test length were used with the Hybrid-2 and constrained versions; increasing the block size from two to four resulted in substantially lower CVC rates when short-test length was used with the unconstrained version regardless of the item quality; and increasing the block size from two to four resulted in substantially lower CVC rates when LQ items and medium-test length were used regardless of the blocking version.

For the DINA and DINO models and using the PWKL as an item selection index, increasing the block size generally resulted in lower classification rates, and the differences in the CVC rates were generally substantial, particularly for LQ items with short- and medium-test lengths, and HQ items with short-test lengths. However, using the same index with long-test lengths, increasing the block size resulted in negligible differences in the CVC rates regardless of the item quality except using LQ items and long-test lengths with the unconstrained version.

For the A-CDM and using the PWKL, increasing the block size from two to four resulted in substantially lower CVC rates for LQ items and medium-test lengths, and for HQ items short-test lengths regardless of the blocking version. For the same generating model and item selection index, the differences were negligible in other conditions.

The impact of test length. Using LQ items, increasing the test length resulted in substantial increases in the classification rates regardless of the block size, blocking version, and generating model. Moreover, the increases for the PWKL were greater than those for the MPWKL and the GDI especially using LQ items. For example, for the DINA model with the block size of 1, increasing the test length from 8 to 16 resulted in 0.32, 0.27, and 0.25 increases in the CVC rates for the PWKL, the MPWKL, and the GDI, respectively. Although the PWKL had higher increases in the CVC rates, the MPWKL and the GDI still had higher classification accuracies when LQ items were used. This is due to the dramatically lower CVC rates for the PWKL when the test was short.

However, using HQ items, the impact of the test length was negligible when the MPWKL and the GDI were used as an item selection index regardless of the block size, blocking version, and generating model. Again, using HQ items, increasing the test length from 8 to 16 resulted in substantial increases when the PWKL was used regardless of the block size, blocking version, and generating model; however, increasing the test length from 16 to 32 using the same index yielded negligible differences in the classification rates.

The impact of item quality. As expected, using HQ items instead of LQ items resulted in higher classification rates regardless of the block size, blocking version, generating model, and item selection index, and those differences are generally substantial. However, the impact of the item quality had negligible differences when the MPWKL and the GDI were used with longer-test lengths regardless of the block size and blocking version. Also, using HQ items instead of LQ items yielded negligible differences when the PWKL was used with longer-test lengths and small block size (i.e., $J_s = 1$) regardless of the blocking version; however, the item quality had a substantial impact on the classification rates when the PWKL was used with longer-test lengths and large block size (i.e., $J_s = 4$).

Interestingly, for the PWKL and short-length tests, increasing the block size resulted in smaller CVC rate differences between LQ and HQ items regardless of the blocking version and generating model. For example, using the unconstrained version and DINA model with the PWKL, the classification differences between LQ and HQ items were 0.42 and 0.16 for the block size 1 and 4, respectively. Moreover, for the PWKL and medium-test lengths, and the MPWKL and the GDI and short- and medium-test lengths, increasing the block size resulted in larger CVC rate differences between LQ and HQ items regardless of the blocking version and generating model.

Item usage. To get a deeper understanding of the differences in item usage across the different blocking versions, items were grouped based on their required attributes. An additional simulation study was carried out using the same factors except for the item quality because this design aimed to eliminate the effect of the item quality on item usage. Therefore, the lowest and highest success probabilities were fixed across all of the items for this study, specifically, $P(\mathbf{0}) = 0.1$ and $P(\mathbf{1}) = 0.9$. The test administration was divided into periods, each of which consisted of four items. Item usage was then recorded in each period. Only the results for the GDI, DINA, eight-item tests, and $\boldsymbol{\alpha}_3$ using the unconstrained, Hybrid-1, Hybrid-2, and constrained versions are shown in online Supplemental Appendix B. However, the results using the other conditions can be requested from the first author.

In the first period, which includes the first four items, single attribute items were mostly used regardless of the block size, blocking version, and generating model. This finding is also consistent with another study in the literature (i.e., G. Xu, Wang, & Shang, 2016), which found that administering single-attribute items at the beginning of the test was more useful compared with the items measuring more attributes. Also, because the uniform distribution was used at the beginning of the test for each blocking version and item selection index, the four single attribute items were the same regardless of the blocking version and generating model when the block size was 1. For example, items with the q-vectors of (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), and (0,0,0,0,1), each with 12.5% usage, were used in the first period for each blocking version and generating model when $J_s = 1$. However, interestingly, for larger blocks (i.e., $J_s = 2$ and $J_s = 4$), the different blocking versions resulted in different item usage types, and also, increasing the block size reduced item usage variability for the unconstrained and Hybrid-2 versions. For example, the unconstrained and Hybrid-2 versions used two types of single attribute items (e.g., items whose q-vectors were (0,0,1,0,0) and (0,0,0,1,0), each with 25% usage) when the block size was 2, and only one type of single attribute item (e.g., items whose q-vector was (0,0,1,0,0) with 50% usage) when the block size was four regardless of the generating model. Moreover, using the constrained version increased the variability in item usage type. For example, the Hybrid-1 and constrained versions used four single attribute items, whose q-vectors were different, in the first period regardless of the block size and generating model.

In the second period, item usage differed based on the block size, blocking version, and generating model. When $J_s = 1$, the item usage types were similar to those observed in the Kaplan et al. (2015) study. To summarize, the item usage patterns based on the generating model were described as following. The DINA model used items that required single attributes that were not mastered by the examinee (e.g., items whose q-vectors were (0,0,0,1,0) with 10% and (0,0,0,0,1) with 8% usages) and items that required the same attributes as the examinee's true attribute mastery vector (e.g., items whose q-vectors were (1,1,1,0,0) with 8% usage). The DINO model used items that required single attributes mastered by the examinee (e.g., items whose q-vectors were (1,0,0,0,0) with 13%, (0,1,0,0,0) with 8%, and (0,0,1,0,0) with 10% usages) and items that required the same attributes as the examinee's true attribute nonmastery vector (e.g., items whose q-vectors were (0,0,0,1,1) with 8% usage). The *A*-CDM used items that required single attributes regardless of the true attribute vector. Those item usage rates were the highest among the other item types. In addition to these item usage types for each model, because items with the q-vector of (1,0,0,0,0) were not administered at all in the first block, this item type was used 13% of the time, which was the highest item usage rate, regardless of the blocking version and generating model in the second period when $J_s = 1$.

When $J_s = 2$ and 4, the blocking versions resulted in different item usage types, and again, increasing the block size reduced the variability in item usage types in the second period. Specifically, the unconstrained version used only single attribute items regardless of the generating model, and also, it used only four and two different item types when $J_s = 2$ and 4,

respectively. For example, when $J_s = 2$, the DINA model mostly used items whose q-vectors were (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), and (0,0,0,0,1) with the item usage rates of 14%, 4%, 4%, and 28%, respectively, and when $J_s = 4$, items whose q-vectors were (0,0,1,0,0) and (0,0,0,1,0) with the item usage rates of 48% and 2%, respectively. Considering the total item usage rate for the second period was 50%, the other item types were not administered at all in the second period. The Hybrid-2 version mostly used single attribute items in addition to the two-attribute items when the block size was larger for the DINA and DINO models. For example, the DINA and DINO models used all single attribute items and items with the q-vector of (1,0,1,0,0) with 8% usage when the block size was 2. The Hybrid-1 and constrained versions yielded the same item usage types for the generating model when the block size was 2. However, it used only one type of single attribute items when the block size was four. Again, the A-CDM used only single attribute items regardless of the blocking version and block size.

Longer test lengths (i.e., 16- and 32-item tests) yielded similar item usage types in the first period as on the eight-item test. Moreover, in the last periods, the blocking versions yielded similar item usage types for the generating models, except for the block size of 4 in which different types of items were used because of the constraint.

Item review. As noted earlier, item review can relax the testing environment and reduce test anxiety for examinees (Vispoel, 2000). In addition, most of the research has found a negative correlation between the test performance and several measures of test anxiety, such as questionnaires and heart pulse rates (Cassady & Johnson, 2002). However, there is no available research on the exact relationship between test anxiety level and test performance. Therefore, in this study, the impact of the test anxiety and item review on the test performance was investigated hypothetically. An additional simulation study was carried out to further investigate the impact of item review using some of the same factors in the simulation study above. Specifically, the unconstrained and constrained versions of block designs were used with the GDI, DINA, and DINO models, and 16-item LQ test. In addition, the impact of the test anxiety on test performance was investigated hypothetically by modifying the examinees' probabilities of success on the items. On one extreme, the impact of anxiety level can be so severe that the probability of success reduces to $P(\mathbf{0})$. On the other extreme, the impact of anxiety can be negligible that it does not affect the examinee's success probability. In other words, examinee's probability of success may or may not be affected by test anxiety. The modified probability of success is given by

$$P^*(X = 1|\boldsymbol{\alpha}) = P(X = 0|\boldsymbol{\alpha}) + [P(X = 1|\boldsymbol{\alpha}) - P(X = 0|\boldsymbol{\alpha})] * (1 - \epsilon), \tag{4}$$

where ϵ is the modification parameter for the probabilities of success. Two different designs were considered in the item review study. The first design assumed a constant impact of ϵ on the probability of success, whereas the second design assumed a nonconstant (i.e., decreasing) impact of ϵ across different blocks within the CD-CAT administration. Specifically, in the first design, six different levels of ϵ, namely, $\epsilon = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,$ and 1.0, were considered. Furthermore, anxiety has no impact on the examinee's success probability when $\epsilon = 0$; in contrast, when $\epsilon = 1.0$, the impact of anxiety dominates such that all success probabilities reduce to $P(\mathbf{0})$. The weighted CVC rates using unconstrained and constrained versions of the GDI with LQ items in a 16-item test are shown in Table 2. It can be seen that the CVC rates using $\epsilon = 0$ are the same as those in Table 1 under the same conditions.

Several findings can be culled from the results. First, increasing the ϵ resulted in lower CVC rates regardless of the block design, generating model, and blocking version. For example, using the constrained version, the DINA model, and $J_s = 1$, the CVC rate was reduced from 0.84 to 0.59 when ϵ was increased from 0 to 0.2, or the rate was reduced from 0.59 to 0.31

Table 2. Weighted CVC Rates for the GDI, $J = 16$, and LQ Items.

BV	Model	J_s	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.6$	$\epsilon = 0.8$	$\epsilon = 1.0$
UC	DINA	1	0.84	0.59	0.31	0.13	0.06	0.03
		2	0.83	0.59	0.32	0.15	0.06	0.03
		4	0.76	0.50	0.27	0.12	0.06	0.03
	DINO	1	0.82	0.59	0.31	0.13	0.05	0.03
		2	0.79	0.55	0.29	0.12	0.05	0.03
		4	0.70	0.49	0.28	0.13	0.06	0.03
C	DINA	1	0.84	0.59	0.31	0.13	0.06	0.03
		2	0.82	0.58	0.32	0.14	0.06	0.03
		4	0.79	0.55	0.31	0.15	0.07	0.03
	DINO	1	0.82	0.59	0.31	0.13	0.05	0.03
		2	0.80	0.56	0.29	0.12	0.05	0.03
		4	0.75	0.51	0.27	0.11	0.05	0.03

Note. CVC = correct attribute vector classification; GDI = G-DINA model discrimination index; LQ = low quality; BV = blocking version; J_s = block size; ϵ = anxiety level; UC = unconstrained version; DINA = deterministic input noisy "and" gate; DINO = deterministic input noisy "or" gate; C = constrained version.

when ϵ was increased from 0.2 to 0.4. Second, the impact of the test anxiety on test performance can be reduced to the extent that item review afforded by the block design can lower anxiety. The results in Table 2 can be interpreted in the following way to delineate the benefit of item review in alleviating the negative impact of test anxiety on an examinee's probability of success. The CVC rates obtained from traditional CAT (i.e., $J_s = 1$) using a specific ϵ need to be compared with the rates using larger block size (i.e., $J_s = 2$ or 4) and lower ϵ. For example, using the constrained version and the DINA model, the CVC rate was 0.13 with $\epsilon = 0.6$ (i.e., higher test anxiety) and $J_s = 1$. However, when item review was allowed (i.e., a larger block size was used), the CVC rate was 0.32 with $\epsilon = 0.4$ and $J_s = 2$ or 0.55 with $\epsilon = 0.2$ and $J_s = 4$ for the same blocking version and model. Moreover, for the DINA model, the benefit of item review can be observed for all anxiety levels. A reduction of 0.2 in the anxiety level (e.g., from 0.2 to 0 or 0.8 to 0.6) resulted in the block designs having higher CVC rates compared with the traditional CAT (i.e., $J_s = 1$). For example, the CVC rates at $\epsilon = 0$ for $J_s = 2$ and 4 of 0.83 and 0.76, respectively, were higher than the CVC rate of 0.59 at $\epsilon = 0.2$ for $J_s = 1$. This pattern was true for all ϵ, and a higher reduction in ϵ resulted in better improvement in the CVC rates. Similar results can be seen when the constrained version and the DINO model were used. Last, it is interesting to note that, for a fixed ϵ, increasing the block size did not always result in lower CVC rates. For example, using the constrained blocking version and DINA model, the CVC rates for $J_s = 1,\ 2,$ and 4 were 0.13, 0.14, and 0.15, respectively, when $\epsilon = 0.6$.

To further investigate the unexpected pattern in the previous results, the CVC rates were calculated for different attribute vectors and test lengths, and the results for the unconstrained blocking version are shown in Figure 2 for $J = 8$. Additional results using the same conditions with $J = 16$ are presented in online Supplemental Appendix C. Again, increasing the block size did not always result in lower CVC rates for some attribute vectors even for $\epsilon = 0$. As it can be seen in Figure 2, when $J = 8$, $\epsilon = 0$, and the attribute vector was (1,1,1,1,1), the CVC rates for $J_s = 1,\ 2,$ and 4 were 0.59, 0.45, and 0.50, respectively. However, using the same ϵ and attribute vector, but longer test length (i.e., $J = 16$), the CVC rates decreased when the block size was increased (i.e., the CVC rates were 0.92, 0.90, and 0.80 for $J_s = 1,\ 2,$ and 4, respectively). These results indicate that the CVC rates may not go down as J_s increases when the test is not

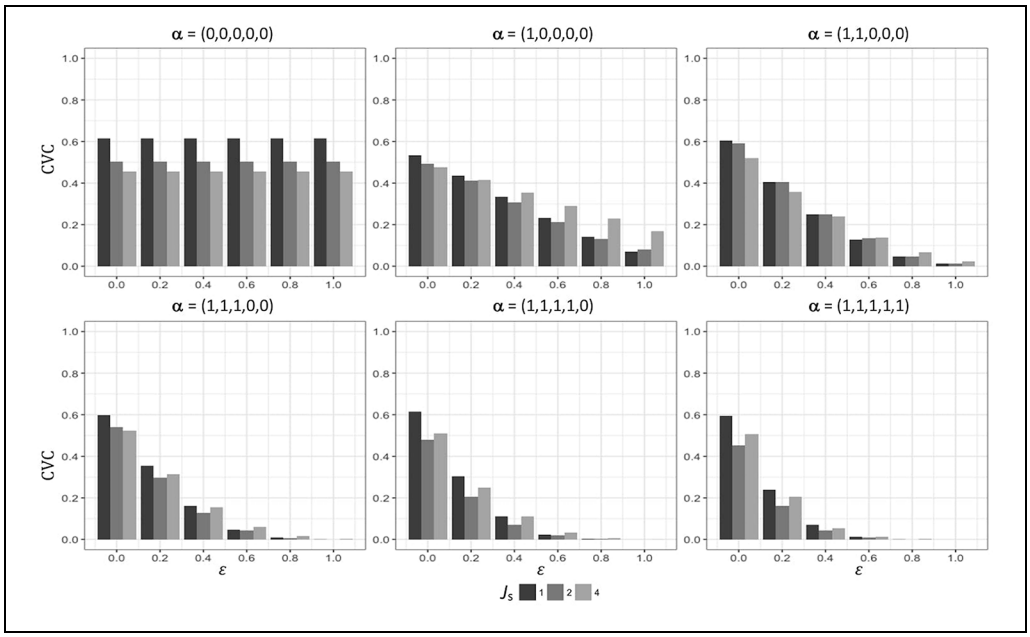

Figure 2. The CVC Rates for the Unconstrained, DINA, LQ, GDI, and $J = 8$.
Note. CVC = correct attribute vector classification; DINA = deterministic inputs, noisy "and" gate; LQ = low quality; GDI = G-DINA model discrimination index; ϵ = anxiety level; J_s = block size.

Table 3. The Unweighted and Weighted CVC Rates.

ϵ	α_{00000}	α_{10000}	α_{11000}	α_{11100}	α_{11110}	α_{11111}	W-CVC
0.5 (fixed)	0.80	0.40	0.26	0.13	0.08	0.04	0.22
0.6 to 0.4	0.80	0.42	0.24	0.12	0.08	0.03	0.22
0.8 to 0.2	0.80	0.44	0.23	0.10	0.07	0.02	0.21

Note. CVC = correct attribute vector classification; ϵ = anxiety level; W-CVC = weighted CVC.

sufficiently informative, as in, the test is short and of low quality, as well as when additional noise (i.e., nonzero ϵ) is involved.

In the second design, an additional simulation study was carried out using the unconstrained version of the GDI, DINA model, and a 16-item LQ test with nonconstant ϵ values across different blocks. Specifically, the CD-CAT administration was started with the large and moderate values of ϵ (i.e., assuming higher test anxiety with $\epsilon = 0.8$ and 0.6) and ended with the smaller values (i.e., assuming lower test anxiety with $\epsilon = 0.2$ and 0.4, respectively). The impact of the ϵ was linearly reduced from the beginning through the end of the test. For example, the ϵ values of .80, .71, .63, .54, .46, .37, .29, and .20 were used to calculate the new probability of success in the first, second, third, forth, fifth, sixth, seventh, and eighth block, respectively. Table 3 gives the results using both a constant ϵ value of 0.5 and nonconstant ϵ values. Different CVC rates were obtained for the six attribute vectors used in this study. The differences in unweighted and weighted CVC raters were negligible using constant and nonconstant ϵ values.

Discussion and Conclusion

Item review and answer change have several benefits for test takers such as reduced test anxiety, the opportunity to correct careless errors, and, most importantly, increased testing validity. However, this option have several drawbacks, including decreased testing efficiency and demand of more complicated item selection algorithms. In a blocked-design CAT, item review is allowed within a block of items, and several studies have shown that there was no significant difference in the accuracy of the ability estimated with limited review and no review procedures.

In this article, a new CD-CAT procedure was proposed to allow item review and answer change during test administration. In this procedure, a block of items was administered with and without a constraint on the q-vectors of the items. Based on the factors in the simulation study and compared to the PWKL, using the new procedure with the MPWKL and the GDI is promising for item review, particularly when HQ items and at least medium-test lengths were involved. In addition, the different blocking versions produced similar classification rates using the MPWKL and the GDI. In contrast, with the PWKL, different blocking versions produced dramatically different results—the constrained version had the best classification accuracy, whereas the unconstrained version had the worst classification accuracy regardless of the block size, test length, and item quality except for HQ items and long tests. The results of this study suggest several findings that are of practical value. First, it is not advisable to use the PWKL with the blocked-design CD-CAT, particularly with larger block sizes because of the substantial decreases in the classification rates across many conditions. However, this is not the case with the MPWKL and the GDI. Second, from this study, practitioners, so as to allow students to review and change their answers, can determine the tolerable level of loss in classification accuracy in deciding the appropriate block size to be used. For example, according to the results of this study, using the GDI with LQ items and long-test lengths, or HQ items with any test lengths, increasing the block size did not reduce the classification rates substantially. Third, item usage types found in this study can be helpful in test construction strategies in the context of cognitive diagnosis. For example, the findings of this study suggest that the use of single attribute items is very important to obtain more appropriate attribute vector estimates of examinees. Last, this study also shows that the classification rates can be improved to the extent that item review afforded by block design can reduce test anxiety.

Although this study showed promise with respect to item review for CD-CAT, more research must be conducted to determine the greater viability of the blocked-design CD-CAT. First, only a single constraint on the q-vectors was considered in the current study; however, it would be interesting to examine different possible constraints (e.g., hierarchical structures) on items. Second, further research needs to focus on possible cheating strategies for cognitive diagnosis. All previously mentioned cheating strategies are based on the difficulty parameter; however, there is no difficulty parameter to speak of for every relevant dimension in CDMs. Therefore, the applicability of those strategies in CDM is still doubtful. Third, the impact of the number of attributes and item pool size was not considered; these factors also affect the performance of the indices in real CAT applications. Last, the data sets were generated using a single reduced CDM. It would be more practical to examine the use of a more general model, which allows the item pool to be made up of various CDMs.

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.

ORCID iD


Mehmet Kaplan https://orcid.org/0000-0002-4175-3899

Supplemental Material

Supplemental material for this article is available online.

References

Benjamin, L. T., Cavell, T. A., & Shallenberger, W. R., III. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology*, *11*, 133-141.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*, 270-295.

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

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

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*, 227-249.

Han, K. T. (2013). Item pocket method to allow response review and change in computerized adaptive testing. *Applied Psychological Measurement*, *37*, 259-275.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191-210.

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

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

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurement*, *21*, 129-142.

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 (Statistical Methodology)*, *65*, 143-157.

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

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

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). Boston, MA: Kluwer.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, *35*, 328-345.

Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement*, *60*, 371-384.

Vispoel, W. P., Clough, S. J., & Bleiler, T. (2005). A closer look at using judgments of item difficulty to change answers on computerized adaptive tests. *Journal of Educational Measurement*, *42*, 331-350.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287-307.

von Davier, M., & Cheng, Y. (2014). Multistage testing using diagnostic models. In Y. Duanli, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 219-227). Boca Raton, FL: CRC Press.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, *73*, 1017-1035.

Wise, S. L. (1996, April). *A critical analysis of the arguments for and against item review in computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *69*, 291-315.

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

Yen, Y.-C., Ho, R.-G., Liao, W.-W., & Chen, L.-J. (2012). Reducing the impact of inappropriate items on reviewable computerized adaptive testing. *Journal of Educational Technology & Society*, *15*, 231-243.

Zheng, Y., & Chang, H.-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*, 104-118.