

Computerized Adaptive Testing for Cognitively Based Multiple-Choice Data

Applied Psychological Measurement

1–14

© The Author(s) 2018

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0146621618798665

journals.sagepub.com/home/apm

Hulya D. Yigit¹, Miguel A. Sorrel²  and Jimmy de la Torre³

Abstract

Cognitive diagnosis models (CDMs) are latent class models that hold great promise for providing diagnostic information about student knowledge profiles. The increasing use of computers in classrooms enhances the advantages of CDMs for more efficient diagnostic testing by using adaptive algorithms, referred to as cognitive diagnosis computerized adaptive testing (CD-CAT). When multiple-choice items are involved, CD-CAT can be further improved by using polytomous scoring (i.e., considering the specific options students choose), instead of dichotomous scoring (i.e., marking answers as either right or wrong). In this study, the authors propose and evaluate the performance of the Jensen–Shannon divergence (JSD) index as an item selection method for the multiple-choice deterministic inputs, noisy “and” gate (MC-DINA) model. Attribute classification accuracy and item usage are evaluated under different conditions of item quality and test termination rule. The proposed approach is compared with the random selection method and an approximate approach based on dichotomized responses. The results show that under the MC-DINA model, JSD improves the attribute classification accuracy significantly by considering the information from distractors, even with a very short test length. This result has important implications in practical classroom settings as it can allow for dramatically reduced testing times, thus resulting in more targeted learning opportunities.

Keywords

cognitive diagnosis models, computerized adaptive testing, MC-DINA, G-DINA, item selection methods, JSD, GDI

Recent studies in the area of education emphasize the value of formative assessments (e.g., de la Torre & Minchen, 2014). Formative assessments allow for an efficient and individualized guidance by targeting specific learning gaps. In this regard, in contrast with traditional item response theory (IRT) based assessments that usually provide an overall unidimensional score, cognitive diagnosis models (CDMs), used in conjunction with formative assessments, can offer rich, fine-grained information about students’ learning progress (Tjoe & de la Torre, 2013). In addition, computerized adaptive testing (CAT) can enhance the advantages of CDMs over

¹University of Illinois at Urbana–Champaign, USA

²Universidad Autónoma de Madrid, Spain

³The University of Hong Kong, Hong Kong

Corresponding Author:

Miguel A. Sorrel, Department of Social Psychology and Methodology, Faculty of Psychology, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain.

E-mail: miguel.sorrel@uam.es

paper-pencil (P-P) tests by providing efficient test settings in classrooms such as shorter and flexible testing times and faster scoring. By utilizing items that are more appropriate for each individual, a more accurate classification rate can be obtained with shorter test lengths. To this end, item selection methods play an essential role. This is indeed a relatively new area of research referred to as cognitive diagnosis computerized adaptive testing (CD-CAT) that is receiving growing interest in the recent literature (for an overview, see Huebner, 2010). Although item selection methods for dichotomous data still hold considerable attention in CD-CAT (e.g., Cheng, 2009; Kaplan, de la Torre, & Barrada, 2015; C. Wang, 2013), no study has been conducted on item selection methods for polytomous data. This is surprising given that assessments typically include multiple-choice (MC) items.

To capitalize on information available in distractors, de la Torre (2009) proposed a CDM for cognitively based MC options, namely, the *multiple-choice deterministic inputs, noisy “and” gate* (MC-DINA) model. In this work, de la Torre showed that the attribute classification accuracy can be substantially increased when the distractors are used as an additional source of information. Given this result, the accuracy can be further improved when this model is applied in the CD-CAT context. Thus, the goal of this study is to examine the efficiency of MC data in CD-CAT administrations. Based on this goal, Jensen–Shannon divergence (JSD) is introduced as a new item selection method for MC data. To investigate the relative efficiency of the proposed method, two heuristic approaches were considered, namely methods based on random selection and the generalized deterministic inputs, noisy “and” gate (G-DINA; de la Torre, 2011) model discrimination index (GDI; Kaplan et al., 2015) applied to dichotomized polytomous data. It should be noted that both methods are considered heuristic, and therefore suboptimal, because they do not capitalize on all available information in the item selection process—GDI does not account for information that can be found in the distractors, whereas random selection does not take the item properties nor the examinees’ previous responses into consideration.

The rest of the article is structured as follows. First, a detailed overview on the CDMs and item selection methods used in the present study is provided. Second, the design of the simulation study is described, and the results from the different item selection methods under the different conditions are presented. Finally, in the “Discussion” section, several implications and limitations of this study are discussed, and possible future directions are provided.

MC-DINA Model

A number of CDMs with different assumptions and constraints on the data such as the deterministic inputs, noisy “and” gate model (DINA; Junker & Sijtsma, 2001), the deterministic inputs, noisy “or” gate model (DINO; Templin & Henson, 2006), and the MC-DINA model (de la Torre, 2009) have been proposed. Although easier to interpret and more practicable, relaxing the constraints of reduced CDMs and enabling more flexible parameterizations can result in better model data fit. To this end, more general CDMs, such as the general diagnostic model (GDM; von Davier, 2005), the G-DINA model (de la Torre, 2011), and the generalized diagnostic classification models for multiple choice option-based scoring (GDCM-MC; DiBello, Henson, & Stout, 2015) have been proposed. Despite for differences in their assumptions, the overall purpose of all of the CDMs is to ascertain whether the examinees have mastered attributes being measured in a certain domain. In this regard, all CDMs have several common components. A binary vector, which can denoted by $\alpha_i = \{\alpha_{ik}\}$ for $k = 1, 2, 3, \dots, K$, with K denoting the number of attributes being assessed, is constructed to specify mastery status for each examinee by coding each entry as 1 if the examinee possesses the attribute considered, otherwise 0. The response pattern for examinee i can be denoted by $\mathbf{X}_i = \{x_{ij}\}$ for an assessment composed of J items. An item-to-attribute mapping is specified in the so-called Q-matrix

(Tatsuoka, 1983). The Q-matrix is a $J \times K$ matrix where each entry can be 1 or 0 depending on whether the item requires this specific attribute or not, respectively.

One of the simplest CDMs is the above-mentioned DINA model. For each item j , the DINA model partitions examinees into two latent groups, those examinees mastering all the K_j^* attributes required by the item, and those examinees lacking at least one. For each of these latent groups, a common probability of success is estimated. The latent vector $\boldsymbol{\eta}_i = \{\eta_{ij}\}$ is used to represent examinee's latent group. This latent vector can only have two possible values. Group $\eta_{ij} = 0$ consists of examinees lacking at least one of the required attributes for this item, whereas group $\eta_{ij} = 1$ consists of examinees who have all of them required attributes for this item. All examinees in the same group will have the same probability of success. There have also been several studies modifying DINA for specific purposes. For instance, to allow for evaluating the possible different strategies that an examinee might use, de la Torre and Douglas (2008) developed the multiple-strategy DINA model. However, in all these models, responses are only considered as either correct or incorrect, regardless of which item types are used (e.g., multiple-choice, open-ended question). This may limit the diagnostic information that can be obtained.

Recently, several models to overcome this limitation including the MC-DINA model (de la Torre, 2009), three "structured" DINA models for MC items (Ozaki, 2015), and the GDCM-MC model (DiBello et al., 2015) have been proposed. Among these three, the GDCM-MC model can be seen as a general latent class model for MC items whereas the models proposed by Ozaki (2015) are restricted models similar to MC-DINA that can still take into account of more than two possible answer choices, but have less parameters than the MC-DINA model. Although the GDCM-MC and MC-DINA models have not been previously compared in terms of classification accuracy and model fit, some differences can be anticipated. DiBello et al. (2015) indicated that GDCM-MC performs worse as the item complexity increases, especially with shorter test lengths. As will be indicated in the "Method" section, the simulation study included a large number of items with more than two coded options. Considering the usual CAT length and the item complexity in our study, GDCM-MC might be expected to perform inauspiciously in terms of the classification accuracy. In addition, Ozaki (2015) compared three new versions of the MC-DINA model with the MC-DINA and the DINA models. The three new versions proposed by Ozaki (2015) always outperformed the DINA model, and sometimes produced better classification accuracy than the MC-DINA model. Importantly, when the test is short, the MC-DINA model provided a better classification accuracy. This result is of special interest for the purposes of the present study given that one of the goals of adaptive testing is to identify an examinee's proficiency with fewer items than that required in conventional tests. Ozaki (2015) also evaluated model fit indices and found that when the number of coded options is large (i.e., three or more), MC-DINA tended to produce better model fit indices. In de la Torre's (2009) work, the MC-DINA model was also compared with the DINA model, and dramatically better correct classification rates with the MC-DINA model were obtained.

Considering all above, in the present study the MC-DINA model was chosen in the data generation process. In depth, under the MC-DINA framework, a MC item could have H number of options. One of these options is the correct one, whereas the others are commonly referred to as distractors. Besides, some of these distractors might be constructed in such a way that examinees who have specific subset of required attributes will have a higher probability of choosing a particular distractor. In other words, some of the distractors might require mastery of some of the required attributes, which are referred to as coded distractors. If the distractor is not required any attributes, it is referred to as a noncoded or noncognitively based distractor (de la Torre, 2009). The response option that requires the largest number of attributes is considered as the correct option for an item.

The MC-DINA model partitions the examinees into $H_j^* + 1$ groups, where H_j^* denotes the number of coded options in item j . In this regard, g_{ij} is the latent group of the examinee i for item j defined by

$$g_{ij} = \arg \max_{h'} \left\{ \alpha_i' \mathbf{q}_{jh'} \mid \alpha_i' \mathbf{q}_{jh'} = \mathbf{q}_{jh'}' \mathbf{q}_{jh'} \right\}, \quad (1)$$

and $g_{ij} = 0$ is the latent group for examinee with a latent class that is not included in any of the coded options. Examinee who does not have any of the attribute patterns coded in the item will answer the item randomly. Examinee whose attribute pattern is included in at least one of the coded distractors will select that distractor with a higher probability. Specifically, the probability of choosing option h for examinee i on item j is given by

$$P_{jh}(\alpha_i) = P(X_{ij} = h \mid \alpha_i) = P(X_{ij} = h \mid g_{ij} = g) = P_j(h \mid g). \quad (2)$$

For each item, the sum of the probabilities of choosing an option for each latent group equals one, $\sum_{h=1}^H P_j(h \mid g) = 1$, and the total number of parameters is equal to $H(H_j^* + 1)$. DINA is a special case of MC-DINA when the number of coded options equals one.

Item Selection Methods for CD-CAT

The demand for CAT has increased recently due to the convenience of determining the examinee's ability profile accurately in the shortest possible time (Cheng, 2009). In CAT, the test for each examinee may include different features such as a different item set, different administration time, or different test length (Kaplan et al., 2015). This is because items are selected adaptively based on the examinee's responses, which enables individualizing the test according to the current estimate of the examinee's proficiency status. The item selection method has been generally considered the most relevant component of CAT. When transferring CAT from the IRT to CDM context, this component requires special attention. Most item selection methods in the IRT context are based on Fisher information. However, because latent variables are discrete in CDM, this index cannot be applied in CD-CAT settings.

Following this, different item selection methods for CD-CAT have been proposed. Initially, X. Xu, Chang, and Douglas (2003) evaluated two item selection methods based on Shannon entropy (SHE) and Kullback–Leibler (KL) information. Cheng (2009) then introduced two modifications of the KL method, namely hybrid HL (HKL) and posterior-weighted KL (PWKL). Later, Kaplan et al. (2015) proposed two methods in CD-CAT: the modification of Cheng's PWKL, namely the modified PWKL (MPWKL) index and the GDI. Although the results showed that the classification differences between the two methods were negligible, they both outperformed PWKL. In addition, GDI yielded shorter test administration times compared with MPWKL.

Furthermore, C. Wang (2013) proposed the mutual information (MI) index (Weissman, 2007) in CD-CAT by simplifying its mathematical formula, which facilitates a real-time application of this method. The results showed that MI was superior to SHE, KL, and PWKL. Zheng and Chang (2016) modified the CDM discrimination index (CDI; Henson & Douglas, 2005) and attribute level CDI (ACDI; Henson, Roussos, Douglas, & He, 2008). These two methods always outperformed the PWKL, ACDI, and CDI methods; generally outperformed MI; and MPWKL with few exceptions. Finally, G. Xu, Wang, and Shang (2016) addressed some issues related to optimal test design of finite items and evaluated the performance of SHE, PWKL, MPWKL, and GDI. Consistent with the previous studies, SHE, MPWKL, and GDI behaved quite similarly. Although some of these and other item selection methods proposed in recent

literature have not been exhaustively compared in a simulation study, the results of Xu et al. indicate that the current methods can be expected to perform quite similarly. The present study considers the GDI method proposed in Kaplan et al. (2015), considering its efficiency and that the G-DINA model is one of the models employed.

Goal of the Study

When dealing with continuous data where two or more probability distributions were compared, Minchen and de la Torre (2016) applied the JSD index as item selection method. They showed that JSD gave a significantly better classification rate of examinee parameters and shorter test length than the random selection method. The present study aims to assess the performance of JSD as an item selection method under the MC-DINA model. As showed in the online appendix, JSD is equivalent to MI in the context of this article (i.e., when multinomial distributions are involved). Thus, like MI, JSD can be interpreted as a measure of the relative entropy between the joint distribution of two random variables and the product of their marginal distributions. As such, JSD also measures the reduction of the uncertainty in one of the random variables due to the existence of the other random variable. In addition, Lin (1991) also showed, JSD can be used to compare more than one probability distribution, which makes it suitable for comparing the different probability distributions generated by the MC-DINA model for the different item options.

Method

Data Generation

The MC-DINA model was used in the generation of polytomous data. Each item has one correct option and four distractors. The MC-DINA model used the options the examinees selected in the analysis. These polytomous data were dichotomized so they can be used with the G-DINA model. Specifically, the option associated with the correct response was coded as 1 and all the distractors were coded as 0. The impact of the two main components of item quality (i.e., item discrimination and variance) was investigated. The coded option attached to the correct response for item j can be denoted as h_j^* . Examinees possessing all the attributes required by item j (i.e., $g_{ij} = h_j^*$) are expected to choose this option. The authors varied these probabilities of success considering two levels of item discrimination (high discrimination [HD] and low discrimination [LD]) and variance (high variance [HV] and low variance [LV]). The item parameters were generated from uniform distributions under the different conditions that are provided in Table 1. Specific values were chosen considering the simulation designs of other studies (e.g., Kaplan et al., 2015; Sorrel, de la Torre, Abad, & Olea, 2017). The mean of the uniform distribution indicates the overall quality of the item pool, whereas the variance affects the overall quality of the set of administered items. Examinees who do not possess any of the attributes required by item j can arrive at the correct response by guessing, and this probability will be defined by $1 / H_j$, where H_j is the number of response options for item j .

In addition, the MC-DINA model allows coding the distractors by associating them with a certain \mathbf{q} -vector specification. The structure has to be hierarchical within the MC-DINA model. Distractors with a simpler \mathbf{Q} -matrix specification (e.g., option C of item j requires Attribute 1) are nested within the distractors with a more complex \mathbf{q} -matrix specification (e.g., option B of item j requires both Attributes 1 and 2). In addition, all the distractors are nested within the correct response option (e.g., option A of item j , which is the correct one, requires Attributes 1, 2, and 3). Item parameters in the HD conditions included in the simulation design were similar to

Table 1. Item Quality and Variance Conditions in the Simulation Design.

Item quality	$P_j(h = h_j^* g = h_j^*)$
HD–HV	$U(0.750, 0.850)$
HD–LV	$U(0.775, 0.825)$
LD–HV	$U(0.550, 0.650)$
LD–LV	$U(0.575, 0.625)$

Note. $P_j(h = h_j^* | g = h_j^*)$ = probability of success for examinees possessing all the attributes required by the item; HD = high discrimination; HV = high variance; U = uniform distribution; LV = low variance; LD = low discrimination.

the ones considered by de la Torre (2009). Within each item, the probability of choosing each one of the possible response options varied as a function of the latent class. To this end, a different set of item parameters was used for the distractors, reflecting that in reality it is plausible that less information can be extracted from the distractors compared with the correct answer. In this sense, the probability of choosing the category measuring the set of attributes mastered by the examinee decreased by .05 for each “step down.” With the aim of representing only a small decrement and considering the number of item options, .05 was chosen. An illustrative example is included in the online appendix. Note that only one item bank per condition was involved, and that it is based on the MC-DINA model. However, two set of item parameters were used—the generating MC-DINA model parameters used in conjunction with polytomous data, and the G-DINA model parameters used in conjunction with the dichotomized data. G-DINA model parameters were specified by considering only the probabilities of success for the correct response options.

The number of attributes was set to 5. To design a more efficient simulation study, only a subset of the possible latent vectors was generated. Specifically, 1,000 alpha patterns were generated for each of the following latent classes: $\alpha_0 = (0, 0, 0, 0, 0)$, $\alpha_1 = (1, 0, 0, 0, 0)$, $\alpha_2 = (1, 1, 0, 0, 0)$, $\alpha_3 = (1, 1, 1, 0, 0)$, $\alpha_4 = (1, 1, 1, 1, 0)$, and $\alpha_5 = (1, 1, 1, 1, 1)$. This sampling design has been shown to provide almost identical results as a sampling design wherein a larger number of alpha patterns are generated uniformly from the possible latent classes (Kaplan et al., 2015). Thus, the number of examinees in the present study was 6,000.

Item Pool and Item Selection Methods

Six hundred items were generated in the MC format as having five options requiring up to five attributes. The item pool was also equally distributed based on the number of attributes that the correct option measured. There were 120 one-attribute items, 120 two-attribute items, and so on. As noted before, each item can have several coded options hierarchically associated with a latent class, which is a subset of the latent class represented by the correct answer. For example, if the correct option of an item had three attributes, this item had two coded distractors, each checking a subset of these three attributes (i.e., two and one attribute, respectively).

Regarding the item selection methods, three different procedures were considered—JSD, GDI, and random. Data were generated considering that each item may have more than one probability distribution depending on the attributes required. The JSD method computes the item discrimination, taking into account these probability distributions. The next item to be administered for an examinee is the one that maximizes JSD. JSD for item j is computed as

$$JSD_j = S(\mathbf{P}_j \times \boldsymbol{\pi}') - \sum_c^{2^K} \pi_c S(\mathbf{P}_{jc}), \quad (3)$$

where \mathbf{P}_j denotes a $H \times 2^K$ matrix where the c column indicates the probability of choosing each of the H item options for examinees belonging to latent class c for $c = 1, \dots, K_j$; $S(\bullet)$ represents SHE computed as $S(\mathbf{P}_{jc}) = E[-\ln(\mathbf{P}_{jc})]$; $\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots, \pi_{2^K}\}$ denotes posterior distribution weights, and $\pi_c \equiv \pi(\alpha_c)$. Note that the original 2^K latent classes are classified into $H_j^* + 1$ latent groups within the MC-DINA model. For all the latent classes included in a certain latent group, the probability of choosing each of the options is identical. As the test proceeds, the examinee's posterior distribution is updated after each item, and this posterior distribution is used to adjust the JSD values.

In contrast with JSD, the GDI computation of each item is allowed to have only one probability distribution resulting from two possible responses, that is, the examinee fails or correctly answers the item. Thus, as previously mentioned, the original data were dichotomized. In this item selection method, the discrimination value of an item is calculated based on how much the probability of success associated with a specific latent class spreads out from the mean of the success probabilities. Therefore, only the latent group probabilities related to the correct option were considered when computing GDI. Within the G-DINA model, the probability of success of examinees with the reduced attribute pattern α_{cj}^* is defined as

$$P(\alpha_{cj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk} \alpha_{ck} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'} \alpha_{ck} \alpha_{ck'} + \dots + \delta_{j12\dots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ck}, \quad (4)$$

where δ_{j0} is the intercept for item j , δ_{jk} is the main effect due to α_{ck} , $\delta_{jkk'}$ is the interaction effect due to α_{ck} and $\alpha_{ck'}$, and $\delta_{j12\dots K_j^*}$ is the interaction effect due to $\alpha_1, \dots, \alpha_{K_j^*}$.

In the context of adaptive testing, this index considers the current examinee's posterior distribution (Kaplan et al., 2015). The probability of this reduced attribute pattern α_{cj}^* is denoted $\pi(\alpha_{cj}^*)$. Let define $P(X_{ij} = 1 | \alpha_{cj}^*)$ as the conditional probability of success on item j given the reduced attribute pattern α_{cj}^* . The general formulation of GDI can be defined as

$$s_j^2 = \sum_{c=1}^{2^{K_j^*}} \pi(\alpha_{cj}^*) \left[P(X_{ij} = 1 | \alpha_{cj}^*) - \bar{P}_j \right]^2, \quad (5)$$

where the average success probability is $\bar{P}_j = \sum_{c=1}^{2^{K_j^*}} \pi(\alpha_{cj}^*) P(X_{ij} = 1 | \alpha_{cj}^*)$. In addition, $\pi(\alpha_{cj}^*)$ was replaced with $\pi^{(i)}(\alpha_{cj}^*)$ to enable GDI to be updated based on the examinee's responses.

The following text describes the other components of the CATs. During both adaptive testing procedures, the first item to be administered to each examinee was always randomly selected among the medium discrimination items in the item bank. Regarding the estimation method, the examinee's latent class was estimated using the *maximum a posteriori* (MAP) method with a flat prior distribution. The models used in the estimation of the person parameters were MC-DINA for JSD and the random selection method, and G-DINA for GDI. Although this might lead to a possible confound between item selection procedure and CDM used in the estimation, this design allows exploring if the approach that is proposed in this article leads to better results compared with the most common practice in CD-CAT, in which data are analyzed at the dichotomous level (i.e., correct–incorrect). Two different termination rules, fixed test length and minimum precision (minimum variance), were considered (Hsu, Wang, & Chen, 2013). In the fixed length condition, each examinee took a certain prespecified number of items; otherwise, the test was not terminated unless a predetermined precision criteria is satisfied, which could create different test lengths for each examinee. Three different fixed lengths with small increments (i.e.,

5, 10, and 20 items) were administered. In the minimum-precision condition, the posterior probability (i.e., $P(\alpha_i | \mathbf{X}_i)$) was aggregated across the latent classes to compute marginal attribute probabilities $P(\alpha_{jk} | \mathbf{X}_i)$. This estimation method, referred to as *expected a posteriori* (EAP) estimation in the CDM literature, can be expected to provide higher marginal correct classification rate for each attribute compared with the MAP estimator (Huebner & Wang, 2011). Four different minimum threshold criteria for the marginal attribute probabilities (i.e., .65, .75, .85, and .95) corresponding to four minimum precisions (i.e., .23, .19, .13, and .05, respectively) were used. In this case, the test was terminated only when the probability of being assigned to one of the two latent classes for each attribute (i.e., mastery or nonmastery) was higher than one of these cutoff values as indicated by the EAP examinee's estimates. In addition, a maximum test length was arbitrarily fixed to 20 items. This was implemented considering the fact that in practical settings testing time is typically limited.

To evaluate the performance of the different item selection methods, several criterion variables were considered. To assess the attribute classification accuracy, the proportion of correctly classified attribute vectors (PCV) and proportion of correctly classified attributes (PCA) were examined in the fixed test length conditions. Efficiency was evaluated in the variable test length conditions. To this end, descriptive statistics related to the final test length for the item selection methods were compared. In addition, item usage was also examined.

Results

Results for the PCV rates from different item selection methods are presented in Table 2. In every condition, JSD always led to better results than the random selection method, and the GDI item selection method. The PCV rates for JSD were generally high. Even with a very short test length (i.e., $J = 5$), the correct attribute pattern was specified 50% of the time when the item quality was high. As expected, as the item quality dropped, the estimation accuracy was negatively affected under every item selection method. However, even with in worst item quality condition (i.e., LD–LV), JSD could maintain adequate classification accuracy (i.e., .72) when the test length was long enough (i.e., $J = 20$). Moreover, it is worth noting that the random selection method for polytomous data produced slightly better results than the GDI method in the five-item test length conditions except for the HV–HD item pool. This is an expected result for short test lengths given that random selection still capitalized on information found in the distractors in estimating the attributes, whereas GDI solely relied on the key. For longer CAT lengths, GDI was always better than the random selection method. This is reflecting the fact that GDI is selecting appropriate items based on the examinee's posterior distribution that is updated at each step of the CAT. Due to the space constraints and the similarity to the PCV rates, the results for the PCA rates are not presented in the article. The pattern on the results for the PCA rates was very similar to that of the PCV rates, with the exception that GDI always yielded slightly higher results than the random selection method.

The results for the proportion of the overall item usage are presented in Table 3. As can be seen from table, JSD tended to use two-, three-, and four-attribute items whereas GDI used one- or two-attribute items more frequently. Moreover, for JSD the frequency of the usage of four-attribute items slightly decreased when the item discrimination was low regardless of the level of variance, except for the 20-item test length condition. In addition, for JSD with the HD items, as the test became longer, the usage of two-attribute items slightly increased whereas the usage of those with three- and four-attributes slightly decreased. However, with the LD items, the usage of two-attribute items increased substantially compared with the usage of the HD items. In contrast, GDI tended to use two-attribute items regardless of item quality.

Table 2. Vector-Level Classification Accuracy for the Different Tests.

IQ	<i>J</i>	Item selection method		
		JSD	GDI	Random
HD–HV	5	0.52	0.39	0.35
	10	0.79	0.71	0.52
	20	0.96	0.93	0.72
HD–LV	5	0.51	0.34	0.35
	10	0.76	0.65	0.52
	20	0.94	0.89	0.72
LD–HV	5	0.29	0.18	0.21
	10	0.52	0.40	0.31
	20	0.76	0.67	0.47
LD–LV	5	0.26	0.18	0.20
	10	0.47	0.34	0.30
	20	0.72	0.60	0.46

Note. Maximum value in each row is shown in bold. IQ = item quality; *J* = CAT length; JSD = Jensen–Shannon divergence; GDI = G-DINA model discrimination index; HD = high discrimination; LV = low variance; HV = high variance; LD = low discrimination; CAT = computerized adaptive testing; G-DINA = generalized deterministic inputs, noisy “and” gate.

Table 3. The Proportion of Overall Item Usage.

Item quality	<i>J</i>	Item selection method									
		JSD					GDI				
		Number of required attributes					Number of required attributes				
		1	2	3	4	5	1	2	3	4	5
HD–HV	5	0.02	0.22	0.41	0.28	0.07	0.34	0.46	0.11	0.08	0.01
	10	0.02	0.30	0.35	0.25	0.07	0.27	0.50	0.10	0.10	0.04
	20	0.07	0.37	0.28	0.20	0.08	0.25	0.45	0.12	0.13	0.05
HD–LV	5	0.00	0.15	0.50	0.25	0.10	0.35	0.45	0.11	0.08	0.01
	10	0.01	0.24	0.45	0.21	0.09	0.27	0.49	0.11	0.08	0.04
	20	0.05	0.34	0.36	0.16	0.08	0.25	0.45	0.13	0.10	0.06
LD–HV	5	0.05	0.41	0.33	0.17	0.04	0.48	0.36	0.09	0.07	0.00
	10	0.03	0.39	0.34	0.20	0.04	0.39	0.43	0.09	0.07	0.01
	20	0.05	0.39	0.31	0.20	0.05	0.33	0.42	0.12	0.10	0.03
LD–LV	5	0.04	0.35	0.43	0.15	0.03	0.41	0.43	0.09	0.07	0.00
	10	0.02	0.35	0.43	0.16	0.04	0.34	0.48	0.12	0.05	0.01
	20	0.03	0.38	0.37	0.16	0.06	0.30	0.45	0.15	0.06	0.04

Note. Maximum numbers in each row and item selection method condition are shown in bold. 1, . . . , 5 refers to each individual attribute; *J* = CAT length; JSD = Jensen–Shannon divergence; GDI = G-DINA model discrimination index; HD = high discrimination; HV = high variance; LV = low variance; LD = low discrimination; CAT = computerized adaptive testing; G-DINA = generalized deterministic inputs, noisy “and” gate.

For the minimum-precision condition, descriptive statistics related to the test length were investigated. Due to space constraints, only the results from two item sets (i.e., highest and lowest quality) are included in Table 4. In the condition with low-quality items, the prespecified precision levels could not be satisfied for a high proportion of the tests whereas with high-

Table 4. Descriptive Statistics of the Test Lengths.

Item quality		Item selection method									
		JSD					GDI				
		Descriptive statistics			Proportion of times that test length > 20 ^a		Descriptive statistics			Proportion of times that test length > 20 ^a	
Threshold criterion	Minimum	Maximum	M	CV	Minimum	Maximum	Minimum	Maximum	M	CV	
HD–HV	0.65	1	20	7.00	72.71	0.02	2	20	7.76	62.61	0.03
	0.75	2	20	7.63	69.15	0.04	2	20	8.85	55.32	0.04
	0.85	2	20	8.34	66.04	0.07	3	20	10.37	47.03	0.08
	0.95	3	20	9.46	58.19	0.11	3	20	11.26	43.72	0.13
LD–LV	0.65	2	20	12.92	45.95	0.2	3	20	14.35	36.33	0.19
	0.75	2	20	14.01	41.25	0.32	3	20	16.17	27.3	0.36
	0.85	3	20	14.84	36.58	0.39	4	20	17.33	22.69	0.54
	0.95	4	20	15.68	32.23	0.47	5	20	18.23	18.72	0.71

Note. JSD = Jensen–Shannon divergence; GDI= G-DINA model discrimination index; CV = coefficient of variation; HD = high discrimination; HV = high variance; LD = low discrimination; LV = low variance; CAT = computerized adaptive testing; G-DINA = generalized deterministic inputs, noisy “and” gate.

^aThe CAT stopped when the probability of being assigned to one of the two latent classes for each attribute (i.e., mastery or nonmastery) was higher than the threshold criterion or the number of items was 20.

quality items, only a small proportion of the tests was not terminated before the precision levels were satisfied. For example, even in the strictest condition (i.e., 0.95), these proportions were 0.11 and 0.13 for JSD and GDI, respectively. However, when the low-quality items were considered, as expected, the proportion of the tests that were terminated before the prespecified precision levels increased dramatically (e.g., 0.47 and 0.71 for JSD and GDI, respectively, in the 0.95 condition). JSD with the feature of accommodating polytomously scored data was more efficient than GDI performed only with dichotomous response data in terms of reaching the precision level using fewer items. Moreover, this efficiency became clearer as the targeted accuracy rate was stricter. This fact was also observed from the mean of the test length. On average, JSD administered two fewer items than GDI in most conditions. In addition, even for the strictest condition (i.e., 0.95) the mean test length for JSD was 9 to 10 and 15 to 16 items with the high-quality and low-quality item conditions, respectively. Moreover, regardless of the item selection methods, the coefficient of variation indicated that the number of administered items for each examinee varied more in the least strict condition (i.e., 0.65). It is important to note that maximum test length was always 20 items because of how the variable test length was designed, so the mean test lengths might be slightly biased. Finally, to obtain a deeper understanding of whether a specific pattern occurs as items are selected, the overall item usage on a specific attribute pattern was explored. The main findings are presented in the online appendix.

Discussion

In classroom settings, remedial instructions can be a useful tool to improve the effectiveness of teaching. To construct an efficient remedial instruction, one of the most important matters is to clarify the strengths and weaknesses of the students. CDMs have been proven to be useful to address this important issue (de la Torre & Minchen, 2014). However, time restrictions might limit the effectiveness of CDMs. In this regard, CD-CAT administrations hold promise for obtaining accurate results in the shortest possible time. Moreover, incorporating the distractors information makes the scoring process easier and faster, so that the time improvement can be magnified. However, although CD-CAT administrations typically include MC items, no item selection methods for polytomously scored MC items has been exhaustively studied before in the CD-CAT literature. This issue can hinder using distractors as an additional source of information about examinee's cognitive proficiency profile. In this article, the efficiency of administering polytomously scored MC items in the CD-CAT framework was investigated. A new item selection method, JSD, was used under MC-DINA, and its efficiency was evaluated against the random selection method and the common approach in CD-CAT consisting of using a dichotomous CDM (i.e., G-DINA) and an item selection rule suitable for dichotomous data (i.e., GDI). The results consistently showed that JSD always led to a better attribute classification accuracy by using information from the distractors compared with GDI, which uses information only from correct options.

Regarding the item usage, JSD used two- and three-attribute items more frequently whereas GDI tended to administer one- or two-attribute items. The pattern of item usage was consistent with previous studies using this index (Kaplan et al., 2015). Specifically, simpler items were preferred at the beginning of the test, and longer CAT lengths resulted in a higher use of complex items; one-attribute items were used more frequently under the LD conditions; and three- and four-, and five-attribute items were seldom used. These results are in line with those included in Kaplan et al. (2015) when data were generated with a noncompensatory model (i.e., DINA), which is equivalent to the dichotomous data set that was employed in this study. Compared with GDI, JSD tended to use items that are more complex. A possible reason for that

is that items with simple structures were nested within the more complex items and this information is available when using the MC-DINA model and the JSD index. The descriptive statistics related to the test length indicated that, mostly, JSD needed two fewer items on average than GDI to reach the targeted accuracy values. Overall, this study indicates the usefulness of MC items when scored polytomously in the CD-CAT context. These results have important implications in practical settings because CD-CAT combined with the MC-DINA model may allow dramatically reduced testing time and number of required items in item pools.

Accuracy and item usage results should be interpreted with care because JSD and GDI are used under different models, as in, the MC-DINA and G-DINA models, respectively. There is, thus, a possible confounding effect of the model used in the person parameter estimation. In any case, JSD in the context of MC tests is deemed more efficient because it considers the probabilities associated with the distractors when selecting the most suitable item given the examinee's current posterior distribution. Current selection methods in CD-CAT, which includes GDI, only consider the probabilities of success associated with the correct response option. Thus, JSD allows for a more optimal approach.

In spite of the promising results, some caveats are in order and should be considered in future studies. First, in the present study, only a subset of the possible attribute patterns was considered, and this subset of attribute patterns reflected an ordinal scale similar to a unidimensional IRT scale. A different design might include all the possible attribute patterns to explore in more detail the classification accuracy for each attribute pattern. Second, only the MC-DINA was employed in the data generating process. It should be remarked that this is a reduced CDM and therefore the conclusions here may not apply to a different reduced CDM (e.g., DINO, Templin & Henson, 2006). Compared with other polytomous CDMs such as the structured DINA models (Ozaki, 2015) for MC items and the GDCM-MC (DiBello et al., 2015), the MC-DINA model has the advantage that it is included in an R package (CDM; George, Robitzsch, Kiefer, Gross, & Uenlue, 2016). Nevertheless, future studies can explore the potential of these other CDMs to support adaptive testing in situations where MC-DINA model assumptions do not hold. Third, authors did not implement any item exposure method in the present study because it was out of the scope of the study. However, they noticed that some items were overexposed. Incorporating an item exposure method into the item selection algorithm could guarantee to limit item exposure rates as well as a more even use of the item pool. C. Wang, Chang, and Huebner (2011) and Zheng and Wang (2017) can be consulted for studies implementing item exposure control in the CD-CAT context. The present article focused on unrestricted algorithms and CAT designs. Some studies have explored the issue of attribute balancing and whether a combination of CAT and multistage adaptive testing designs can be implemented to improve the classification results (Cheng, 2010; S. Wang, Lin, Chang, & Douglas, 2016). Finally, it should be noted that in this study the item parameters for the distractors were generated in such a way that less information could be extracted from them, compared with the correct response. This reflects a plausible scenario but future studies might consider this a factor in the simulation design. Different studies considered the situation where the distractors are as informative as the correct response (e.g., de la Torre, 2009; Ozaki, 2015). It is also possible that distractors can be even more informative than the correct option; for example, this can occur when some of the attributes included in the subset of attributes measured by the distractors are easier to master. In these scenarios, the performance of JSD as an item selection method can be expected to be better given that the distractions would be more informative.


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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by Grants PSI2013-44300-P and PSI2017-85022-P (Ministerio de Economía y Competitividad and European Social Fund).

ORCID iD

Miguel A. Sorrel  <https://orcid.org/0000-0002-5234-5217>

Supplemental Material

Supplemental material is available for this article online.

References

- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74, 619-632.
- Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement*, 70, 902-913.
- de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement*, 33, 163-183.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.
- de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, 73, 595-624.
- de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, 20, 89-97.
- DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement*, 39, 62-79.
- George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Uenlue, A. (2016). The R Package CDM for cognitive diagnosis models. *Journal of Statistical Software*, 74, 1-24.
- Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, 29, 262-277.
- Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, 32, 275-288.
- Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, 37, 563-582.
- Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation*, 15, 1-7.
- Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, 71, 407-419.
- 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., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, 39, 167-188.
- Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37, 145-151.
- Minchen, N. D., & de la Torre, J. (2016, July). *The continuous G-DINA model and the Jensen-Shannon divergence*. Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.
- Ozaki, K. (2015). DINA models for multiple-choice items with few parameters: Considering incorrect answers. *Applied Psychological Measurement*, 39, 431-447.

- Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 13, 39-47.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287-305.
- Tjoe, H., & de la Torre, J. (2013). Designing cognitively-based proportional reasoning problems as an application of modern psychological measurement models. *Journal of Mathematics Education*, 6, 17-26.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RP-05-16). Princeton, NJ: Educational Testing Service.
- 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.
- Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, 48, 255-273.
- Wang, S., Lin, H., Chang, H. H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, 53, 45-62.
- Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, 67, 41-58.
- 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*. Annual meeting of the American Educational Research Association, Chicago, IL.
- Zheng, C., & Chang, H. H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, 40, 608-624.
- Zheng, C., & Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, 41, 561-576.