Applied Psychological Measurement

A New Method to Balance Measurement Accuracy and Attribute Coverage in Cognitive Diagnostic Computerized Adaptive Testing

Journal:	*Applied Psychological Measurement*
Manuscript ID	APM-19-02-031.R3
Manuscript Type:	Manuscripts
Keywords:	Cognitive diagnostic computerized adaptive testing, attribute coverage, measurement accuracy, the ratio of test length to the number of attributes

A New Method to Balance Measurement Accuracy and Attribute Coverage in Cognitive Diagnostic Computerized Adaptive Testing

Abstract

As one of the important research areas of cognitive diagnosis assessment, cognitive diagnostic computerized adaptive testing (CD-CAT) has received much attention in recent years. Measurement accuracy is the major theme in CD-CAT and both the item selection method and the attribute coverage have a crucial effect on measurement accuracy. A new attribute coverage index, the ratio of test length to the number of attributes (RTA), is introduced in the current study. RTA is appropriate when the item pool comprises many items that measure multiple attributes where it can both produce acceptable measurement accuracy and balance the attribute coverage. With simulations, the new index is compared to the original item selection method (ORI) and the attribute balance index (ABI), which have been proposed in previous studies. The results show that: (1) the RTA method produces comparable measurement accuracy to the ORI method under most item selection methods; (2) the RTA method produces higher measurement accuracy than the ABI method for most item selection methods, with the exception of the mutual information item selection method; (3) the RTA method prefers items that measure multiple attributes, compared to the ORI and ABI methods, while the ABI prefers items that measure a single attribute; and (4) the RTA method performs better than the ORI method with respect to attribute coverage, while it performs worse than the ABI with long tests.

Keywords

Cognitive diagnostic computerized adaptive testing, the ratio of test length to the number of attributes, measurement accuracy, attribute coverage

Introduction

Cognitive diagnosis assessment (CDA) has recently received much attention in educational and psychological assessment (Rupp & Templin, 2008). Compared to classical test theory and item response theory (IRT), which only provide an overall score to indicate the information about the position of one individual relative to others on one specific latent trait (de la Torre & Chiu, 2016), CDA can provide detailed information about the strengths and weaknesses of individuals for specific content domains. Consequently, efficient remediation can be conducted based on the fine-grained information available about individuals (Gierl, Leighton, & Hunka, 2007; Lim & Drasgow, 2017; Sawaki, Kim, & Gentile, 2009).

One important research area in CDA is cognitive diagnostic computerized adaptive testing (CD-CAT; Cheng, 2009; McGlohen & Chang, 2008; X. Xu, Chang, & Douglas, 2003). CD-CAT combines a cognitive diagnostic model (CDM) and computer technology to improve testing efficiency and measurement accuracy. Like IRT-based CAT, CD-CAT has compelling advantages over traditional paper-and-pencil (P&P) tests. For example, the performance of individuals can be estimated immediately after they provide a response to each item (Cheng & Chang, 2009). CD-CAT can also provide equivalent or higher accuracy in the measurement of an individual's latent skills, with reductions in test length.

The primary goal of CD-CAT is to improve the measurement accuracy of individuals (Zheng & Chang, 2016) and the item selection method is one of the most important keys to this. Numerous item selection methods have been proposed, such as the Kullback-Leibler method (KL; X. Xu et al., 2003), the Shannon Entropy method (Tatsuoka, 2002), the posterior-weighted KL method (PWKL; Cheng, 2009), the mutual information method (MI; Wang, 2013), and the modified PWKL method (MPWKL; Kaplan, de la Torre, & Barrada, 2015). Recently, Zheng and Chang (2016) developed two new item selection methods designed for short-length tests: the posterior-weighted cognitive diagnostic index (PWCDI) and the attribute-level discrimination index (PWADI), based on previous work by Henson & Douglas (2005) and Henson, Roussos, Douglas, & He (2008).

In addition to the item selection method, the coverage for each attribute can also impact the measurement accuracy. Cheng (2010) indicated that attribute coverage influences both measurement accuracy and reliability, and it is important to make sure that each attribute is measured adequately to ensure the validity of the inferences based on the test. Therefore, she used the modified maximum global discrimination index (MMGDI) method, first used in IRT-based CAT by Cheng and Chang (2009), to balance the attribute coverage and improve measurement accuracy. The simulation study showed that, compared with the original KL method, the MMGDI method produced a relatively higher attribute correct classification rate (ACCR) and pattern correct classification rate (PCCR).

When the minimum number of items that measure each attribute is not satisfied, the attribute balance index (ABI) used in Cheng (2010) tends to select items with a single attribute (Mao & Xin, 2013), which means that the ABI is suitable when the item pool is composed of many items that measure a single attribute. Measurement accuracy would however be lower if the item pool is comprised of many items that measure multiple attributes. Although a test with single-attribute items can produce high PCCR in the CDA framework (e.g. Madison & Bradshaw, 2015; Wang, 2013), it is difficult to construct such items because more than one attribute is required to successfully solve items in real testing situations (DeCarlo, 2011; Huang, 2018). An extreme case is when there are hierarchical relationships among attributes (Leighton, Gierl, & Hunka, 2004), where the ABI tends to produce low measurement accuracy. In addition, the ABI has only been used with the KL method and its performance with other item selection methods is unknown. Therefore, the current study proposes a new method — the modified ratio of test length to the number of attributes (RTA), influenced by the study conducted by Kuo, Pai, and de la Torre (2016) — to balance attribute coverage and improve measurement accuracy when the item pool comprises many multiple-attribute items. Furthermore, the study examines whether the RTA and ABI can be extended to more types of item selection methods.

The remainder of the paper is organized as follows: First, we will introduce the two CDMs used in the study and summarize the item selection methods used. After that, the ABI and RTA will be presented. Then, a simulation study is conducted to examine the RTA with respect to the correct classification rate conditional on several manipulated factors. Finally, the discussion and conclusions are presented.

Cognitive diagnostic models and item selection methods

Numerous CDMs have been proposed to deal with different test situations and with CD-CAT, the ‘‘Deterministic Input, Noisy ‘And’ Gate’’ (DINA) model (Junker & Sijtsma, 2001) and the Reduced Reparameterized Unified Model (RRUM; Hartz, 2002) are commonly used (e.g., Chen, Xin, Wang, & Chang, 2012; Cheng, 2010; Huebner, Finkelman, & Weissman, 2018; G. Xu, Wang, & Shang, 2016). Let α_{ik} denote the mastery of attribute k for individual i and q_{jk} denote if the attribute k is required to answer item j correctly. The item response function (IRF) of the DINA model is then

$$P\left(x_{ij}=1\mid\eta_{ij}\right)=\left(1-s_j\right)^{\eta_{ij}}\ g_j^{\ 1-\eta_{ij}},$$

where $\eta_{ij}=\prod_{k=1}^{K}(\alpha_{ik})^{q_{jk}}$ and s_j and g_j are item parameters. With the RRUM, the IRF is

$$P\left(x_{ij}=1\mid\boldsymbol{\alpha}_i\right)=\pi_j^*\prod_{k=1}^{K}r_{jk}^{*\,(1-\alpha_{ik})q_{jk}},$$

where π_j^* and r_{jk}^* are the item parameters. The item selection method plays an important role in CD-CAT and is the main determinant of ACCR and PCCR. This study uses the four item selection methods MI (Wang, 2013), MPWKL (Kaplan et al, 2015), PWCDI and PWADI (Zheng and Chang, 2016). For details on the interpretation of the cognitive diagnostic model parameters and the item selection methods, we refer to the supplementary material.

Attribute coverage indices

ABI. The ABI was proposed to make sure that each attribute was measured adequately to improve the correct classification rate (Cheng, 2010). It is defined as

$$ABI_j=\prod_{k=1}^{K}\left(\left(B_k-b_k\right)/B_k\right)^{q_{jk}},$$

where B_k is the minimum number of items that should measure the k^{th} attribute and b_k is the number of items that have already been selected to measure the k^{th} attribute.

RTA. Kuo et al. (2016) proposed the RTA to ensure that each attribute is adequately measured when constructing a P&P cognitive diagnostic test. In this paper, we extend this method to CD-CAT and modify it to balance the attribute coverage. The RTA in a CD-CAT context can be written as

$$RTA_j=\frac{1}{1+I\left(H\le B_k\right)\sum_{v=1}^{V}I\left(\boldsymbol{q}_j=\boldsymbol{q}_v^*\right)},\ H=\min\left(b_1,b_2,\ldots,b_K\right),$$

where V refers to the number of items that have already been selected; $I(\cdot)$ is the indicator function; and $\boldsymbol{q}_j$ and $\boldsymbol{q}_v^*$ are the q-vectors of items that have not been and have already been given to a specific person, respectively.

The term $I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ controls the usage of items that measure different numbers of attributes, and the relationship between H and B_k strives to ensure that each attribute is measured at least B_k times. If $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ is larger than 0 and H is no larger than B_k, then the value of $I\left(H \le B_k\right)\sum_{v=1}^{V} I\left(\boldsymbol{q}_j = \boldsymbol{q}_v^*\right)$ tends to be large. Consequently, the RTA becomes small and the j^{th} item will not be selected. Instead, items with different attribute patterns to the previously selected items will tend to be selected. On the other hand, when H is larger than B_k or $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ is 0, then the RTA is equal to 1. In such a case, RTA will not affect the item selection method and therefore the items will then be selected based on the original item selection method.

The RTA criterion balances the attribute coverage and prefers multiple-attribute items. On the contrary, the ABI criterion balances the attribute coverage and prefers single-attribute items. Note that the RTA is determined by both H and $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$, which means that, if H is larger than B_k (or $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ is 0), then $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ (or H) can be ignored. Therefore, the RTA criterion may not guarantee that each attribute is covered completely. In sum, we expect ABI to perform better than RTA regarding attribute coverage given a long enough test, with RTA performing better than ABI regarding measurement accuracy when the item pool contains many multiple-attribute items. Item selection methods that consider both the attribute coverage and the information that an item provides can be developed by multiplying the attribute coverage indices (ABI or RTA) and the original item selection methods, for example the MMGDI can be obtained by the multiplication ABI × KL.

Simulation study

The goals of the simulation study are to examine the performance of the new attribute coverage index and examine whether the RTA and ABI can be extended to other item selection methods. Several factors are manipulated: model type, number of attributes, Q-matrix structure, test length, attribute coverage index, and item selection method. In total there are 2 (model type) × 2 (number of attributes) × 2 (Q-matrix structure) × 3 (test length) × 3 (attribute coverage index) × 4 (item selection method) = 288 conditions in the study. The details of the simulation study are given in the following.

Model type. Both the DINA model and the RRUM will be used in the current study since these two CDMs are commonly used in CD-CAT (e.g., Cheng, 2010; Huebner et al., 2018; Mao & Xin, 2013; G. Xu et al., 2016).

Number of attributes. Wang (2013) and Zheng and Chang (2016) used five attributes in their studies, while Cheng (2010) used six attributes in her study. In the current study, both five and six attributes are considered to examine the performance of RTA and ABI.

Q-matrix structure. Two types of Q-matrix are generated in this study, namely simple structure and complex structure (Chen et al., 2012; Huang, 2018; Wang, 2013). For the simple structure Q-matrix, all items are unidimensional, meaning that each item measures a single attribute. This Q-matrix is generated based on a discrete uniform distribution with equal probability for all possible patterns. Meanwhile, for the complex structure Q-matrix between one and three attributes are measured by each item. The generation of the complex structure Q-matrix is based on Chen et al. (2012) and can be summarized as follows. First, three basic matrix units are generated. The first matrix unit is a K-by-K identity matrix, while the second and third matrix units are comprised of all possible q-vectors that measure two and three attributes, respectively. Second, the first matrix unit is replicated twenty times while the second and third matrix units are replicated ten times. This results in 100 items that each measure one, two, and three attributes, respectively. Third, the items are merged to create a 300-by-K matrix, and the rows of the 300-by-K matrix are randomly re-ordered.

Test length. Three different test lengths (10, 20, and 30 items) will be used in this study. We view these as short-length, moderate-length, and long-length tests, similar to previous research (e.g., Kuo et al., 2016).

Attribute coverage index (ACI). Three types of ACI will be used in the study. The first type is the original item selection method without attribute coverage control (abbreviated to ORI), which can be treated as the baseline. The second type is the ABI proposed by Cheng (2010) and the last type is the RTA which is proposed in the current study.

Item selection method. The item selection methods used in this study are the MI, MPWKL, PWADI, and PWCDI methods. All these methods can produce high correct classification rates even for short-length test.

Since the generation of the $\boldsymbol{\alpha}$-matrix for five and six attributes are the same, we will only describe the generation of the $\boldsymbol{\alpha}$-matrix for five attributes. A 1000-by-5 matrix is generated to represent the true attribute patterns ($\boldsymbol{\alpha}$-matrix). Each individual can master each attribute with probability equal to .5 and we assume independence among individuals and independence among attributes in the $\boldsymbol{\alpha}$-matrix. For the item parameters, both slipping and guessing parameters were generated from a uniform distribution $U(.05, .30)$ for the DINA model, and the baseline and penalty parameters were generated from $U(.65, .95)$ and $U(.05, .50)$, respectively, for the RRUM. During the item selection procedure, the minimum number of items that measure each attribute was set to 3 because previous studies demonstrated that each attribute should be measured at least three times in the CDA framework (e.g. Fang, Liu, & Ying, 2019; Gu & G. Xu, 2019; G. Xu, 2017). Finally, the expected a posteriori (EAP) method is used to estimate the attribute patterns. Twenty replications for each condition are used in current study.

The evaluation criteria used in this study are averaged ACCR (A-ACCR), PCCR, and the usage of *k*-attribute items (Kuo et al., 2016). These statistics are calculated by

$$PCCR = \sum_{i=1}^{N} I\left(\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i\right) \Big/ N,$$

$$A\text{-}ACCR = \sum_{i=1}^{N}\sum_{k=1}^{K} I\left(\hat{\alpha}_{ik} = \alpha_{ik}\right) \Big/ \left(N \times K\right), \text{ and}$$

$$Usage_k = \sum_{i=1}^{N}\sum_{j=1}^{J} I\left(\sum_{h=1}^{K} q^*_{ijh} = k\right) \Big/ \left(N \times J\right), k = 1, 2, \ldots, K,$$

where N and J are the number of individuals and test length, respectively; $I(\cdot)$ is the indicator function, which will be 1 if $\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i$ (or $\hat{\alpha}_{ik} = \alpha_{ik}$) is true, and vice versa; $\hat{\boldsymbol{\alpha}}_i$ and $\boldsymbol{\alpha}_i$ are the estimated and true values of an individual's attribute pattern, respectively; q^*_{ijh} is the h^{th} entry of q-vector for item j that has already been answered by individual i. In addition, the empirical standard errors (SEs) for PCCR and A-ACCR, $SE = \sqrt{\frac{1}{n_{sim} - 1}\sum_{i=1}^{n_{sim}}\left(\hat{\theta}_i - \bar{\theta}\right)^2}$ (where n_{sim} is the number of replications, $\hat{\theta}_i$ and $\bar{\theta}$ are the i^{th} estimation and the mean value of PCCR and ACCR, respectively), are calculated to evaluate the uncertainty of these two indices (Morris, White, & Crowther, 2019).

Table 1. Correct classification rate for the DINA model (K = 5)

Item selection method	Q matrix	Attribute coverage index	*J* = 10				*J* = 20				*J* = 30			
			PCCR		A-ACCR		PCCR		A-ACCR		PCCR		A-ACCR	
			Est	SE	Este	SE	Est	SE	Est	SE	Est	SE	Est	SE
MI	simple	ORI	.752	.015	.945	.003	.891	.013	.978	.003	.953	.007	.991	.001
		ABI	.704	.011	.932	.003	.920	.013	.984	.003	.958	.007	.992	.001
		RTA	.712	.015	.934	.004	.917	.008	.983	.002	.959	.006	.992	.001
	complex	ORI	.694	.014	.926	.004	.881	.012	.974	.003	.948	.005	.989	.001
		ABI	.699	.008	.931	.003	.901	.015	.979	.004	.959	.008	.991	.002
		RTA	.695	.013	.924	.004	.868	.011	.970	.003	.952	.006	.990	.001
MPWKL	simple	ORI	.844	.012	.966	.003	.986	.003	.997	.001	.999	.001	1.00	.000
		ABI	.835	.010	.964	.002	.980	.006	.996	.001	.999	.001	1.00	.000
		RTA	.838	.007	.965	.002	.987	.004	.997	.001	.999	.001	1.00	.000
	complex	ORI	.860	.006	.966	.002	.988	.003	.997	.001	.998	.001	1.00	.000
		ABI	.798	.014	.955	.003	.982	.004	.996	.001	.998	.001	1.00	.000
		RTA	.852	.014	.963	.004	.986	.004	.997	.001	.998	.002	1.00	.000
PWADI	simple	ORI	.847	.015	.967	.003	.987	.004	.997	.001	.999	.001	1.00	.000
		ABI	.831	.011	.964	.002	.979	.004	.996	.001	.999	.001	1.00	.000
		RTA	.845	.013	.966	.003	.985	.005	.997	.001	.999	.001	1.00	.000
	complex	ORI	.833	.011	.954	.004	.981	.004	.995	.001	.997	.001	.999	.000
		ABI	.789	.014	.952	.003	.982	.004	.996	.001	.998	.002	.999	.000
		RTA	.827	.014	.951	.005	.980	.004	.995	.001	.998	.001	1.00	.000
PWCDI	simple	ORI	.843	.014	.966	.003	.989	.003	.998	.001	.999	.001	1.00	.000
		ABI	.824	.013	.962	.003	.980	.003	.996	.001	.999	.001	1.00	.000
		RTA	.843	.013	.966	.003	.988	.004	.997	.001	.999	.001	1.00	.000
	complex	ORI	.858	.011	.965	.003	.985	.004	.997	.001	.998	.001	1.00	.000
		ABI	.803	.011	.956	.003	.983	.002	.996	.001	.998	.002	1.00	.000
		RTA	.846	.016	.960	.004	.984	.003	.996	.001	.998	.001	1.00	.000

Note. MI refers to mutual information method; MPWKL refers to modified posterior-weighted Kullback-Leibler method; PWADI refers to posterior-weighted attribute-level discrimination index; and PWCDI refers to posterior-weighted cognitive diagnostic index; ORI refers to original item selection method without attribute coverage control; ABI refers to Cheng's (2010) method; RTA refers to the ratio of test length to the number of attributes; PCCR refers to pattern correct classification rate; A-ACCR refers to averaged attribute correct classification rate; Est refers to the estimate, SE is standard error.

Results

Correct classification rate

Tables 1 and 2 present the correct classification rates and the corresponding empirical standard errors (SEs) for the DINA model and the RRUM, respectively, conditional on five attributes. Table 1 shows that all three attribute coverage indices produce similar PCCRs and A-ACCRs for the long-length test (*J* = 30). When test lengths are short (*J* = 10) and moderate

(J = 20), some differences are found between ORI, ABI, and RTA. The ABI, in general, produces higher PCCRs and A-ACCRs than ORI and RTA for moderate- and long-length tests with the MI method, while the RTA performs as well as or even better than ABI with other three item selection methods regardless of test length and Q-matrix structure. To examine which factors (attribute coverage index, test length, and Q-matrix structure) have a significant effect on the measurement accuracy, four repeated measures ANOVAs are conducted for the item selection methods, respectively. Results show that all main effects, second- and third-order interaction effects are statistically significant for the MI method, the partial etas (η_p^2) range from .085 (attribute coverage index) to .996 (test length), and the ABI performs significantly better than RTA for complex-structure Q-matrix and moderate- and long-length tests. For short-length tests, RTA produces significantly higher PCCR than ABI for a simple-structure Q-matrix, while RTA produces relatively lower PCCR than ABI for a complex-structure Q-matrix. With MPWKL and PWADI, all main effects, second- and third-order interaction effects are statistically significant, with the exception of main effect of Q-matrix structure and second-order interaction effect between test length and Q-matrix structure. The η_p^2 range from .101 (interaction effect between attribute coverage index and Q-matrix structure with PWADI method) to .998 (test length with MPWKL method) for the significant effects, and the RTA performs significantly better than ABI for complex-structure Q-matrix and short- and moderate-length tests with both MPWKL and PWADI. Similar to the MI method, all effects are significant for the PWCDI method, and the η_p^2 range from .052 (interaction effect between attribute coverage index and Q-matrix structure) to .997 (test length), and the RTA performs significantly better than ABI for complex-structure Q-matrix and short-length tests and for a simple-structure Q-matrix and moderate-length tests. In addition, the empirical SEs are small for all conditions, indicating that the estimates of PCCRs and A-ACCRs are stable.

Table 2. Correct classification rate for the RRUM (K = 5)

Item selection method	Q matrix	Attribute coverage index	$J = 10$				$J = 20$				$J = 30$			
			PCCR		A-ACCR		PCCR		A-ACCR		PCCR		A-ACCR	
			Est	SE	Este	SE	Est	SE	Est	SE	Est	SE	Est	SE
MI	simple	ORI	.745	.014	.943	.003	.892	.014	.977	.003	.955	.007	.991	.002
		ABI	.697	.012	.931	.003	.909	.008	.981	.002	.958	.005	.991	.001
		RTA	.698	.015	.930	.003	.912	.009	.982	.002	.957	.005	.991	.001
	complex	ORI	.705	.013	.930	.004	.872	.009	.971	.002	.943	.008	.988	.001
		ABI	.689	.012	.928	.003	.884	.012	.975	.003	.943	.009	.988	.002
		RTA	.697	.016	.927	.004	.862	.011	.969	.003	.938	.007	.987	.002
MPWKL	simple	ORI	.836	.013	.965	.003	.984	.004	.997	.001	.998	.001	1.00	.000
		ABI	.824	.009	.962	.002	.977	.005	.995	.001	.998	.001	1.00	.000
		RTA	.835	.011	.964	.003	.984	.004	.997	.001	.998	.001	1.00	.000
	complex	ORI	.849	.010	.965	.003	.980	.004	.996	.001	.997	.002	.999	.000
		ABI	.784	.016	.951	.004	.976	.005	.995	.001	.996	.002	.999	.000
		RTA	.841	.010	.962	.002	.976	.006	.995	.001	.996	.002	.999	.000
PWADI	simple	ORI	.831	.011	.963	.002	.985	.003	.997	.001	.998	.001	1.00	.000
		ABI	.825	.013	.962	.003	.978	.004	.995	.001	.998	.001	1.00	.000
		RTA	.832	.012	.964	.003	.983	.005	.997	.001	.998	.001	1.00	.000
	complex	ORI	.836	.011	.959	.003	.980	.004	.995	.001	.996	.002	.999	.000
		ABI	.768	.011	.948	.003	.973	.006	.994	.002	.996	.002	.999	.000
		RTA	.831	.012	.959	.003	.978	.003	.995	.001	.995	.002	.999	.001
PWCDI	simple	ORI	.839	.013	.965	.003	.984	.004	.997	.001	.998	.002	1.00	.000
		ABI	.826	.011	.962	.003	.977	.005	.995	.001	.998	.001	1.00	.000
		RTA	.829	.012	.963	.003	.982	.004	.996	.001	.998	.001	1.00	.000
	complex	ORI	.842	.009	.963	.002	.981	.005	.996	.001	.997	.001	.999	.000
		ABI	.780	.012	.951	.003	.972	.005	.994	.001	.995	.003	.999	.001
		RTA	.835	.011	.961	.003	.975	.004	.995	.001	.995	.003	.999	.001

The results in Table 2 exhibit a similar pattern to that observed with the RRUM model: the ABI performs better than ORI and RTA for moderate- and long-length tests for the MI method while it performs worse for short-length tests. In addition, both RTA and ORI produce larger PCCRs than ABI for short- and moderate-length tests for the MPWKL, PWADI, and PWCDI methods. Moreover, all of these three attribute coverage indices produce very similar PCCRs when the test length is long. Furthermore, the RTA produces a lower A-ACCR than

ABI for the MI method, while it produces an identical or larger A-ACCR than ABI for most conditions. All main effects and second- and third-order interaction effects are statistically significant, with the exception of the second-order interaction effect between test length and Q-matrix structure for the MPWKL method, and the η_p^2 range from .046 (interaction effect between attribute coverage index and Q-matrix structure with MI method) to .998 (test length with PWADI method) for the significant effects. Finally, the empirical SEs are small and the corresponding estimates are stable.

The PCCR and A-ACCR for six attributes are presented in the supplementary material and the results can be summarized as follows: (1) The ABI, in general, produces higher PCCRs and A-ACCRs than RTA for the MI method; (2) the RTA and ORI methods produce higher PCCRs and A-ACCRs than ABI with the MWPKL, PWADI, and PWCDI methods regardless of Q-matrix structure and test length; (3) all the third-order interaction effects are significant, and the η_p^2 range from .251 (for RRUM and PWCDI method condition) to .534 (for DINA model and MWPKL method condition); (4) with the increase of test length, the SEs are decreased for all conditions.

The usage of items

Since all of the items in the simple-structure Q-matrix are single-attribute, all item selection methods select single-attribute items, which results in no differences in the usage of items that measure *k*-attributes for ORI, ABI, and RTA. Therefore, the details will not be presented. Table 3 presents the usage of items that measure *k*-attributes for five attributes and with the complex-structure Q-matrix. The usage of items that measure *k*-attributes for six attributes is consistent with the results with five attributes, and the details can be accessed in the supplementary material. Unsurprisingly, the RTA method selects the least items that measure a single attribute in most of the conditions, followed by the ORI method. The ABI method uses the most items that measure a single attribute. Specifically,

Table 3. The usage of items measures *k*-attribute for five attributes and complex Q-matrix

model	item selection method	attribute coverage index	$J = 10^a$			$J = 20^a$			$J = 30^a$		
			1-A	2-As	3-As	1-A	2-As	3-As	1-A	2-As	3-As
DINA	MI	ORI	.489	.381	.131	.507	.337	.156	.528	.306	.166
		ABI	.864	.063	.074	.738	.178	.084	.633	.245	.122
		RTA	.420	.433	.147	.436	.397	.167	.490	.339	.171
	MPWKL	ORI	.468	.366	.166	.400	.373	.227	.408	.358	.233
		ABI	.888	.077	.036	.671	.212	.117	.540	.291	.168
		RTA	.435	.396	.169	.377	.393	.230	.397	.369	.233
	PWADI	ORI	.368	.406	.226	.360	.388	.253	.386	.368	.246
		ABI	.833	.130	.037	.622	.243	.135	.513	.307	.180
		RTA	.359	.421	.220	.346	.402	.253	.379	.376	.245
	PWCDI	ORI	.431	.385	.184	.387	.377	.236	.404	.358	.238
		ABI	.882	.083	.035	.665	.215	.119	.538	.291	.171
		RTA	.405	.408	.187	.367	.395	.239	.394	.367	.239
RRUM	MI	ORI	.480	.395	.125	.443	.396	.161	.430	.392	.178
		ABI	.932	.032	.036	.729	.189	.082	.574	.299	.127
		RTA	.410	.411	.179	.377	.422	.200	.396	.406	.198
	MPWKL	ORI	.492	.355	.153	.427	.381	.192	.413	.381	.206
		ABI	.875	.090	.035	.662	.231	.107	.524	.316	.160
		RTA	.434	.390	.175	.385	.408	.207	.399	.391	.211
	PWADI	ORI	.434	.380	.186	.402	.392	.205	.400	.387	.213
		ABI	.829	.135	.036	.622	.260	.118	.504	.330	.166
		RTA	.406	.393	.201	.372	.408	.219	.392	.391	.217
	PWCDI	ORI	.475	.365	.160	.418	.385	.197	.409	.382	.208
		ABI	.866	.099	.036	.653	.236	.111	.519	.319	.163
		RTA	.427	.391	.182	.381	.408	.212	.396	.391	.213

Note. *k*-A(s) means items measure *k* attribute(s);

[a] 4-As and 5-As equal to 0 for all conditions.

the proportion of items that measure a single attribute ranges from .346 to .490, .360 to .528, and .504 to .930 for the RTA, ORI, and ABI criteria, respectively. In addition, among these three attribute coverage indices, the RTA method produces the largest proportions of items that measure two and three attributes, followed by the ORI method, and the ABI method yields the smallest proportions of items that measure two and three attributes. These results can be expected since the RTA criterion tends to choose items that measure different attributes to the already administered items. The ABI criteria, on the contrary, tends to penalize items that measure multiple attributes by taking the product of deviances for all attributes. Consequently, items that measure a single attribute tend to be selected by the ABI criteria.

Coverage of attributes

Table 4 lists the proportion of individuals who have been administered at least three times measuring each attribute for moderate- and long-length tests. The results are omitted for the short-length test (i.e. $J = 10$) because all three attribute coverage indices do not satisfy the attribute coverage requirement. The ABI can ensure that most of the tests satisfy the attribute coverage regardless of number of attributes, model type, Q-matrix structure, item selection method, and test length, while RTA performs worse than ABI but better than the ORI. Repeated measures ANOVAs are conducted to investigate the differences among ORI, ABI, and RTA. The results show that most main effects, second-, third- and fourth-order interaction effects are significant under the DINA model, and most of the η_p^2 are larger than .50. Although the differences between ABI and RTA are significant for some conditions, the η_p^2 range from .001 to .114, which indicates that stronger evidence is needed to support differences between ABI and RTA. For the RRUM, all main effects and second-, third- and fourth-order interaction effects are significant, and the η_p^2 range from .797 to .999. In addition, all of the main effects of attribute coverage index, test length, and number of attributes are significant for all item selection methods, and all of the η_p^2 are larger than .950. Although the fourth-order interaction effects are significant for all item selection methods, the partial etas are small and range from .002 to .065. Furthermore, the third-order interaction effects among attribute coverage

Table 4. Overall percentage for moderate- and long-length tests

Number of attributes	Model type	Q matrix	Attribute coverage index	*J* = 20				*J* = 30			
				MI	MPWKL	ADI	CDI	MI	MPWKL	ADI	CDI
K = 5	DINA	simple	ORI	.357	.846	.842	.851	.881	.997	.998	.997
			ABI	.986	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.892	.891	.892	1.00	.999	1.00	.999
		complex	ORI	.772	.940	.956	.945	.974	.998	.999	.998
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.965	.976	.969	1.00	1.00	1.00	1.00
	RRUM	simple	ORI	.326	.812	.813	.817	.867	.996	.996	.996
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.872	.876	.871	1.00	.999	.999	.999
		complex	ORI	.795	.910	.919	.910	.977	.996	.995	.995
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.981	.984	.982	1.00	1.00	1.00	1.00
K = 6	DINA	simple	ORI	.034	.468	.464	.463	.495	.983	.984	.985
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.516	.507	.507	1.00	.993	.959	.961
		complex	ORI	.714	.775	.823	.793	.971	.988	.988	.987
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	.935	.828	.860	.834	1.00	.995	.987	.983
	RRUM	simple	ORI	.020	.327	.318	.322	.430	.957	.953	.957
			ABI	1.00	1.00	1.00	1.00	.993	1.00	1.00	1.00
			RTA	1.00	.427	.357	.418	1.00	.985	.952	.922
		complex	ORI	.569	.721	.755	.733	.865	.974	.981	.977
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	.900	.804	.794	.805	1.00	.999	.986	.972

Note. The results are omitted for the short-length test (i.e. *J* = 10) because all three attribute coverage indices do not satisfy the attribute coverage requirement.

index, number of attributes, and test length are significant for all item selection methods, and the corresponding η_p^2 are at the range of .969 and .980, and the ABI performs better than the RTA at six attributes and moderate-length tests.

Discussion and conclusions

The goals of this study are to develop a new attribute coverage method, RTA, to deal with empirical situations when more than one attribute is involved in successfully solving a test item (DeCarlo, 2011; Huang, 2018) and to examine the performance of both ABI and RTA when different item selection methods are used. A simulation study is conducted to examine the performance of RTA and ABI, and promising results are produced.

The results show that the RTA produces lower PCCRs than ABI for moderate- and long-length tests with the MI method, especially with a complex structure Q-matrix. On the contrary, the RTA produces relatively high PCCRs than the ABI for short- and moderate-length tests with the MPWKL, PWADI, and PWCDI methods. A possible explanation is that both the MI method and the ABI criterion prefer single-attribute items, while the RTA and three other item selection methods tend to use fewer single-attribute items than ABI and MI method. As Madison and Bradshaw (2015) and Huebner et al. (2018) demonstrated, the more single-attribute items there are in a test, the higher the measurement accuracy is for long-length tests. Therefore, the RTA can be expected to produce lower measurement accuracy since fewer single-attribute items are used for the MI method. As for the MPWKL, PWADI, and PWCDI methods, the differences between the usage of items that measure one and two attributes are small, meaning that these item selection methods prefer items that measure either one or two attributes. Therefore, when the ABI criteria, which prefers the single-attribute items, is added to these three item selection methods, information provided by two-attribute items may be lost and, consequently, lower measurement accuracy is produced for the ABI compared to the ORI and RTA criteria. Meanwhile, a possible reason why the ABI performs worst in most conditions for short-length tests (J = 10) is that it is hard to satisfy the minimum number of items that measure each attribute when the test length is short. Although previous studies demonstrated that tests containing more single-attribute items tend to produce higher measurement accuracy (Huebner et al., 2018; Madison & Bradshaw, 2015), the prerequisite for a high measurement accuracy is that the test length is long enough.

Moreover, the results show that the ABI is not suitable for all item selection methods. In the current study, the ABI is suitable for the MI method, while it is unsuitable for the MPWKL, PWADI, and PWCDI methods. In the study of Cheng (2010), the combination between ABI and KL method (MMGDI) can produce higher measurement accuracy than the original KL method (MGDI). Since both the ABI criterion and KL/MI methods prefer single-attribute items rather than multiple-attribute items, using the ABI criterion further reinforces the tendency of the KL and MI methods to select single-attribute items. Hence, the combination between the ABI criterion and the original item selection methods would produce high measurement accuracy if the original item selection methods prefer single-attribute items. On the flipside, low measurement accuracy would be produced if more than one attribute is preferred by the original item selection methods (e.g. MPWKL, PWADI and PWCDI).

It's worth noting that, although the RTA criteria produces higher measurement accuracy than the ABI criteria with the MPWKL, PWADI, and PWCDI methods, this does not indicate that the RTA performs better than ABI for all situations. By examining the ABI and RTA criteria, the ABI tends to penalize items that measure multiple attributes, while the RTA tends to select items that measure multiple attributes. Therefore, it is reasonable to infer that the composition of items that measure different number of attributes in the item pool have an important influence on these two criteria. The RTA performs better than ABI if there is a large number of multiple-attribute items in the item pool. Meanwhile, the ABI performs better than RTA if there is a majority of single-attribute items, producing higher measurement accuracy than RTA for all conditions.

The results also show that the ABI performs better than the RTA for moderate- and long-length tests concerning the attribute coverage, which coincides with our expectation. As stated previously, the formulation of the RTA is determined by two components. One is used to control the usage of items that measure different numbers of attributes and the other is used to control the attribute coverage. When one of the components is satisfied, the other component is ignored. For instance, when the summation of the first component is zero, the component that controls the attribute coverage is ignored and consequently the attribute coverage will not be satisfied.

In conclusion, the new attribute coverage control method—RTA—is suitable for controlling the attribute coverage and producing acceptable measurement accuracy when the item pool is comprised of a large number of items that measure multiple attributes, which is a common phenomenon in empirical testing situations (DeCarlo, 2011; Huang, 2018). The ABI, on the other hand, is appropriate for test situations when the majority of an item pool is comprised of single-attribute items. Furthermore, the ABI is suitable for item selection methods that prefer single-attribute items, such as the KL method (Cheng, 2010) and the MI method, but is not suitable for methods that prefer both single- and multiple- attributes items such as the MPWKL, PWADI, and PWCDI methods.

Although some promising results are found in the current study, several remaining open issues deserve further studies. First, we assume that the minimum number of items that measure each attribute are the same for all attributes. Considering that different attributes may carry different importance, this is not a necessary constraint and further studies can take the importance of each attribute into consideration to further investigate the performance of attribute coverage methods in CD-CAT. Second, fixed-length tests were used in the current study. Therefore, everyone was administered the same test length. Future studies can examine the performance of RTA when the test length is different for each individual (variable-length tests). Third, both the DINA model and the RRUM are specific CDMs and some constraints imposed on these specific CDMs are (a) only a single model is available across the entire test and (b) either compensatory or non-compensatory relationships is assumed for the test (Ravand, 2016). General CDMs relax these constraints and therefore a general CDM can be used in future studies.

Acknowledgment

The authors would like to thank the Editor in Chief, Dr. John R. Donoghue, the Associate Editor, Dr. Chun Wang, and two anonymous reviewers for their helpful comments on earlier drafts of this article.

Supplemental Material

Supplemental material for this article is available online.

References

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, 77, 201-222.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 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*(6), 902-913.

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 369-383.

DeCarlo, L, T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*(1), 8-26.

de la Torre, J., & Chiu, C. Y. (2016). General method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273.

Fang, G., Liu, J., & Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika, 84*(1), 19-40.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Gu, Y., & Xu, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika, 84*(2), 468-483.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practice.* Unpublished doctoral dissertation, University of Illinois at Urbana Champaign.

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.

Huang, H. Y. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *Journal of Classification, 35*(3), 437-465.

Huebner, A., Finkelman, M. D., & Weissman, A. (2018). Factors affecting the classification accuracy and average length of a variable-length cognitive diagnostic computerized test. *Journal of Computerized Adaptive Testing*, *6*(1). DOI: 10.7333/1802-060101.

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*(3), 167-188.

Kuo, B. C., Pai, H. S., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement*, *40*(5), 315-330.

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

Lim, Y. S., & Drasgow, F. (2017). Nonparametric calibration of item-by-attribute matrix in cognitive diagnosis. *Multivariate behavioral research*, *52*(5), 562-575.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491-511.

Mao, X., & Xin, T. (2013). The application of the Monte Carlo approach to cognitive diagnostic computerized adaptive testing with content constraints. *Applied Psychological Measurement*, *37*(6), 482-496.

McGlohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*(3), 808-821.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine, 38*(11), 2074-2102.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782-799.

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

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, *6*(3), 190-209.

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.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, *73*(6), 1017-1035.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics, 45*(2), 675-707.

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*(3), 291-315.

Xu, X., Chang, H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the 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*(8), 608-624.

Supplementary material

CDMs

The item response function (IRF) of the DINA model can be written as

$$P\left(x_{ij}=1\mid\eta_{ij}\right)=\left(1-s_j\right)^{\eta_{ij}}g_j^{\,1-\eta_{ij}},$$

where $\eta_{ij}=\prod_{k=1}^{K}(\alpha_{ik})^{q_{jk}}$ is the ideal response, which indicates whether the individual masters all the attributes that a specific item requires; s_j is the slip parameter, which indicates the probability of an individual who has mastered all the attributes that are required for item *j* to obtain an incorrect response and g_j is the guess parameter, which indicates the probability of an individual who has not mastered all the required attributes to obtain a correct response for item *j*.

The IRF of the RRUM can be expressed as

$$P\left(x_{ij}=1\mid\boldsymbol{\alpha}_i\right)=\pi_j^*\prod_{k=1}^{K}r_{jk}^{*\,(1-\alpha_{ik})q_{jk}},$$

where π_j^* is the baseline parameter, which refers to the probability of correct response to item *j* when individuals master all the attributes that item *j* requires. Meanwhile, r_{jk}^* is the penalty parameter, which indicates the reduction in the probability of correct response to item *j* when individuals lack attribute *k*. Both π_j^* and r_{jk}^* range from 0 to 1.

Item selection methods

Mutual information (MI) method. The MI method in CD-CAT has been proposed as an item selection method for short-length tests. It is defined as the expected KL divergence between the joint distribution of the posterior distribution of attribute pattern $\boldsymbol{\alpha}$ given the first *j−1* items, $\pi\left(\boldsymbol{\alpha} \mid \boldsymbol{X}_{j-1}\right)$, and the posterior predictive probability of the *j*th item given all previous *j−1* items, $P\left(X_{ij} = x \mid \boldsymbol{X}_{j-1}\right)$, and the product of the marginal distributions of $\pi\left(\boldsymbol{\alpha} \mid \boldsymbol{X}_{j-1}\right)$ and $P\left(X_{ij} = x \mid \boldsymbol{X}_{j-1}\right)$ (Wang, 2013). The MI index can be written as

$$MI_{ij} = \sum_{x=0}^{1} P\left(X_{ij} = x \mid \boldsymbol{X}_{j-1}\right)\left[\sum_{c=1}^{2^K} \pi\left(\boldsymbol{\alpha}_c \mid \boldsymbol{X}_{j-1}, \mathrm{X}_{ij} = x\right) \times \log\left(\frac{\pi\left(\boldsymbol{\alpha}_c \mid \boldsymbol{X}_{j-1}, \mathrm{X}_{ij} = x\right)}{\pi\left(\boldsymbol{\alpha}_c \mid \boldsymbol{X}_{j-1}\right)}\right)\right],$$

where $P\left(X_{ij} = x \mid \boldsymbol{X}_{j-1}\right)$ can be calculated as

$$P\left(X_{ij} = x \mid \boldsymbol{X}_{j-1}\right) = \sum_{c=1}^{2^K} P\left(X_{ij} = x \mid \boldsymbol{\alpha}_c\right)\pi\left(\boldsymbol{\alpha}_c \mid \boldsymbol{X}_{j-1}\right) = \frac{\sum_{c=1}^{2^K} P\left(\boldsymbol{X}_{j-1}, X_{ij} = x \mid \boldsymbol{\alpha}_c\right)\pi_0\left(\boldsymbol{\alpha}_c\right)}{\sum_{c=1}^{2^K} P\left(\boldsymbol{X}_{j-1} \mid \boldsymbol{\alpha}_c\right)\pi_0\left(\boldsymbol{\alpha}_c\right)},$$

and $\pi\left(\boldsymbol{\alpha}_c \mid \boldsymbol{X}_{j-1}, \mathrm{X}_{ij} = x\right)$ is the posterior probability conditional on the first *j* items.

Modified posterior-weighted Kullback-Leibler (MPWKL) method. The MPWKL method is a modification of the PWKL method. The PWKL method uses the point estimate to represent an individual's posterior probability of the attribute patterns given the response pattern. The MPWKL method, on the other hand, uses the entire rather than a single posterior distribution of attribute pattern(s) to represent the KL divergence between the current estimate of the attribute pattern and other attribute patterns; therefore, it can be expected that the MPWKL method can provide more information and produce smaller measurement error of the posterior probability about individuals than the PWKL method (Kaplan et al., 2015). The MPWKL index can be calculated as

$$MPWKL_{ij} = \sum_{d=1}^{2^K} \left\{ \sum_{c=1}^{2^K} \left[\sum_{x=0}^{1} \log \left(\frac{P\left(X_{ij} = x \mid \boldsymbol{\alpha}_d\right)}{P\left(X_{ij} = x \mid \boldsymbol{\alpha}_c\right)} \right) P\left(X_{ij} = x \mid \boldsymbol{\alpha}_d\right) \pi\left(\boldsymbol{\alpha}_c \mid \boldsymbol{X}_{n-1}\right) \right] \pi\left(\boldsymbol{\alpha}_d \mid \boldsymbol{X}_{n-1}\right) \right\}.$$

Posterior-weighted cognitive diagnostic index (PWCDI) and posterior-weighted attribute-level discrimination index (PWADI) methods. Henson and colleagues (2005, 2008) proposed CDI and ADI to construct cognitive diagnostic testing in a P&P context and Zheng and Chang (2016) extended them to CD-CAT. And based on the same logic as the PWKL method, Zheng and Chang (2016) proposed the PWCDI and PWADI methods, which can be written as

$$PWCDI_j = \frac{1}{\sum_{u \neq v} h\left(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v\right)^{-1}} \sum_{u \neq v} h\left(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v\right)^{-1} PWD_{juv}$$

and

$$PWADI_j = \frac{1}{2^K} \sum_{h\left(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v\right) \equiv 1} PWD_{juv},$$

$$PWD_{juv} = \pi\left(\boldsymbol{\alpha}_u\right) \times \pi\left(\boldsymbol{\alpha}_v\right) \times \sum_{x=0}^{1} P_{\boldsymbol{\alpha}_u}\left(X_j = x\right) \log \left(\frac{P_{\boldsymbol{\alpha}_u}\left(X_j = x\right)}{P_{\boldsymbol{\alpha}_v}\left(X_j = x\right)} \right)$$

where $h\left(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v\right) = \sum_{k=1}^{K} \left| \boldsymbol{\alpha}_{uk} - \boldsymbol{\alpha}_{vk} \right|$ is the Hamming distance between attribute patterns $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ $(u, v = 1, 2, \ldots, 2^K)$, and $h\left(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v\right) \equiv 1$ refers to any pair of attribute patterns $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$ with the Hamming distance equal to 1; $P_{\boldsymbol{\alpha}_u}\left(X_j\right)$ and $P_{\boldsymbol{\alpha}_v}\left(X_j\right)$ are either the IRFs of the DINA model or the RRUM, and $\pi\left(\boldsymbol{\alpha}\right)$ is the posterior probability of all attribute patterns (2^K).

Items with the largest value will be administered to an individual for those item selection methods mentioned above.

Results

Table A. Correct classification rate for the DINA model (K = 6)

Item selection method	Q matrix	Attribute coverage index	J = 10				J = 20				J = 30			
			PCCR		A-ACCR		PCCR		A-ACCR		PCCR		A-ACCR	
			Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
MI	simple	ORI	.672	.019	.936	.004	.818	.014	.967	.003	.905	.009	.984	.002
		ABI	.649	.014	.930	.003	.874	.007	.978	.001	.928	.009	.988	.002
		RTA	.649	.018	.930	.004	.872	.012	.978	.002	.928	.006	.988	.001
	complex	ORI	.538	.012	.893	.004	.792	.013	.959	.003	.896	.008	.981	.002
		ABI	.576	.011	.913	.003	.825	.010	.968	.002	.915	.008	.985	.001
		RTA	.556	.015	.899	.004	.775	.011	.955	.002	.894	.008	.980	.002
MPWKL	simple	ORI	.742	.013	.951	.003	.966	.005	.994	.001	.996	.003	.999	.000
		ABI	.734	.012	.950	.003	.925	.008	.987	.002	.995	.003	.999	.000
		RTA	.743	.015	.952	.003	.963	.005	.994	.001	.995	.002	.999	.000
	complex	ORI	.759	.013	.948	.003	.965	.004	.993	.001	.996	.003	.999	.000
		ABI	.693	.016	.940	.003	.939	.007	.988	.001	.995	.002	.999	.000
		RTA	.752	.011	.944	.003	.962	.006	.993	.001	.994	.002	.999	.000
PWADI	simple	ORI	.734	.011	.950	.002	.967	.004	.994	.001	.994	.003	.999	.000
		ABI	.735	.013	.950	.003	.928	.009	.988	.002	.995	.003	.999	.000
		RTA	.737	.014	.950	.003	.964	.006	.994	.001	.992	.003	.999	.000
	complex	ORI	.713	.016	.927	.004	.955	.006	.989	.001	.992	.002	.998	.001
		ABI	.688	.018	.936	.004	.942	.007	.988	.001	.994	.003	.999	.001
		RTA	.711	.010	.927	.003	.946	.009	.988	.002	.992	.004	.998	.001
PWCDI	simple	ORI	.735	.014	.950	.003	.965	.005	.994	.001	.995	.002	.999	.000
		ABI	.734	.013	.950	.003	.927	.006	.987	.001	.995	.002	.999	.000
		RTA	.736	.012	.950	.003	.966	.005	.994	.001	.992	.003	.999	.001
	complex	ORI	.750	.014	.943	.004	.968	.006	.994	.001	.995	.003	.999	.000
		ABI	.700	.015	.940	.003	.939	.006	.988	.001	.994	.002	.999	.000
		RTA	.747	.016	.943	.006	.961	.007	.992	.002	.993	.003	.999	.000

Table B. Correct classification rate for the RRUM (K = 6)

Item selection method	Q matrix	Attribute coverage index	J = 10				J = 20				J = 30			
			PCCR		A-ACCR		PCCR		A-ACCR		PCCR		A-ACCR	
			Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
MI	simple	ORI	.631	.016	.927	.004	.797	.014	.963	.003	.892	.008	.981	.001
		ABI	.601	.017	.919	.004	.851	.010	.974	.002	.894	.011	.982	.002
		RTA	.598	.015	.918	.004	.849	.013	.973	.002	.916	.006	.985	.001
	complex	ORI	.571	.014	.906	.003	.770	.010	.955	.002	.882	.009	.978	.002
		ABI	.557	.019	.907	.004	.765	.013	.956	.003	.848	.010	.971	.002
		RTA	.572	.016	.906	.004	.765	.013	.953	.003	.873	.011	.976	.002
MPWKL	simple	ORI	.704	.015	.943	.003	.954	.006	.992	.001	.991	.004	.998	.001
		ABI	.699	.015	.942	.003	.902	.009	.983	.002	.990	.003	.998	.001
		RTA	.701	.014	.942	.003	.948	.008	.991	.001	.989	.003	.998	.001
	complex	ORI	.728	.013	.942	.003	.954	.008	.991	.002	.990	.003	.998	.001
		ABI	.660	.014	.931	.003	.919	.009	.985	.002	.988	.004	.998	.001
		RTA	.712	.013	.937	.003	.943	.009	.989	.002	.990	.003	.998	.001
PWADI	simple	ORI	.706	.012	.943	.003	.951	.008	.992	.001	.992	.002	.999	.000
		ABI	.700	.012	.943	.003	.904	.011	.983	.002	.990	.003	.998	.001
		RTA	.706	.016	.943	.003	.943	.011	.990	.002	.987	.003	.998	.001
	complex	ORI	.704	.014	.932	.004	.947	.006	.989	.002	.989	.004	.998	.001
		ABI	.666	.014	.931	.004	.915	.007	.983	.002	.987	.003	.997	.001
		RTA	.700	.013	.931	.004	.942	.007	.988	.001	.988	.003	.998	.001
PWCDI	simple	ORI	.710	.013	.944	.003	.951	.006	.992	.001	.990	.003	.998	.001
		ABI	.698	.013	.942	.003	.900	.011	.982	.002	.989	.003	.998	.001
		RTA	.714	.014	.945	.003	.951	.008	.992	.001	.987	.003	.998	.001
	complex	ORI	.726	.014	.940	.003	.951	.007	.991	.001	.991	.002	.998	.000
		ABI	.665	.011	.932	.002	.917	.010	.984	.002	.988	.003	.998	.001
		RTA	.715	.015	.937	.003	.943	.008	.989	.002	.990	.003	.998	.001

Table C. The usage of items that measuring *k*-attribute for six attributes and complex Q-matrix

model	item selection method	attribute coverage index	J = 10[a]			J = 20[a]			J = 30[a]		
			1-A	2-As	3-As	1-A	2-As	3-As	1-A	2-As	3-As
DINA	MI	ORI	.450	.359	.191	.445	.343	.212	.465	.326	.209
		ABI	.935	.033	.032	.845	.086	.069	.650	.213	.137
		RTA	.417	.388	.195	.356	.416	.228	.418	.366	.216
	MPWKL	ORI	.426	.335	.238	.374	.356	.270	.382	.353	.266
		ABI	.875	.090	.035	.769	.132	.098	.578	.241	.181
		RTA	.393	.368	.239	.326	.394	.280	.361	.369	.270
	PWADI	ORI	.314	.377	.308	.312	.381	.307	.351	.365	.284
		ABI	.780	.177	.042	.680	.192	.128	.528	.272	.200
		RTA	.298	.398	.304	.287	.401	.312	.351	.365	.284
	PWCDI	ORI	.401	.343	.256	.357	.363	.280	.376	.353	.271
		ABI	.870	.094	.036	.765	.134	.100	.577	.241	.183
		RTA	.373	.369	.258	.315	.395	.290	.375	.355	.270
RRUM	MI	ORI	.479	.331	.190	.403	.407	.190	.378	.429	.193
		ABI	.935	.034	.031	.836	.113	.052	.432	.347	.221
		RTA	.441	.360	.200	.320	.451	.229	.331	.452	.218
	MPWKL	ORI	.443	.354	.204	.354	.426	.220	.342	.434	.224
		ABI	.798	.170	.032	.700	.211	.089	.522	.321	.158
		RTA	.416	.372	.212	.312	.445	.244	.318	.447	.235
	PWADI	ORI	.354	.401	.245	.318	.443	.239	.325	.443	.232
		ABI	.786	.179	.035	.683	.219	.097	.509	.328	.163
		RTA	.346	.405	.249	.303	.445	.251	.408	.391	.202
	PWCDI	ORI	.415	.364	.221	.345	.430	.226	.337	.435	.228
		ABI	.796	.171	.033	.694	.213	.094	.518	.321	.160
		RTA	.399	.376	.225	.306	.444	.250	.334	.436	.230

Note. *k*-A(s) means items measure *k* attribute(s);

[a] 5-As equal to 0 for all conditions.

APM: Final Manuscript Submission Checklist

Please note that all materials should, if possible, be submitted online, as a revision of your manuscript. This allows the manuscript's final files to proceed directly, via the Manuscript Central System, to Production after the Editor has completed his final check. It is not necessary any longer to send electronic files to us via e-mail.

(√) Author Contact Information for all authors if possible, and for the corresponding author at the absolute minimum, containing up-to-date contact information for each author, including:

- Institutional affiliation, with desired affiliation to be printed with the manuscript if different from current affiliation
- Work address
- Work phone number
- Work fax number
- Work e-mail address
- Home address*
- Home phone number*
- Other e-mail address (if any)*

(√) The final version of the manuscript (preferably in MS Word format; if submitting in LaTex/Tex format, please follow the LaTex/Tex submission instructions attached separately).

(√) Electronic files for all figures (camera-ready; see attachment for information on preparing camera-ready figures: "SAGE Artwork Submission Guidelines.")

(√) Abstract (should be part of the manuscript itself)

(√) Keywords (should be part of the manuscript itself)

(√) Tables (should be part of the manuscript itself). Also please make sure all the tables and equations are in their raw formats (word, excel, LaTex, etc.), not as images. We cannot accept tables or equations in image format because we are unable to edit them and they can become disfigured.

(√) References (should be part of the manuscript itself)

(√) Acknowledgements, if any (should be part of the manuscript itself)

(√) Appendices, if any (If the appendices are accepted as a part of the main article, they must be submitted as a part of the Main Document; if they are accepted as online supplements, they must be submitted as a separate Supplementary File.)

(√) Permission to use any reproduced or copyrighted material (if needed)

(√) This checklist, completed.

* Needed to facilitate communications in case of a change in institutional affiliation. Will not be used unless contact cannot be made using work information. (Since Sage Publishers is working during vacations, weekends, etc., they like to have alternative contact information available in case something urgent arises and you are not at work.)

A New Method to Balance Measurement Accuracy and Attribute Coverage in Cognitive Diagnostic Computerized Adaptive Testing

Abstract

As one of the important research areas of cognitive diagnosis assessment, cognitive diagnostic computerized adaptive testing (CD-CAT) has received much attention in recent years. Measurement accuracy is the major theme in CD-CAT and both the item selection method and the attribute coverage have a crucial effect on measurement accuracy. A new attribute coverage index, the ratio of test length to the number of attributes (RTA), is introduced in the current study. RTA is appropriate when the item pool comprises many items that measure multiple attributes where it can both produce acceptable measurement accuracy and balance the attribute coverage. With simulations, the new index is compared to the original item selection method (ORI) and the attribute balance index (ABI), which have been proposed in previous studies. The results show that: (1) the RTA method produces comparable measurement accuracy to the ORI method under most item selection methods; (2) the RTA method produces higher measurement accuracy than the ABI method for most item selection methods, with the exception of the mutual information item selection method; (3) the RTA method prefers items that measure multiple attributes, compared to the ORI and ABI methods, while the ABI prefers items that measure a single attribute; and (4) the RTA method performs better than the ORI method with respect to attribute coverage, while it performs worse than the ABI with long tests.

Keywords

Cognitive diagnostic computerized adaptive testing, the ratio of test length to the number of attributes, measurement accuracy, attribute coverage

Introduction

Cognitive diagnosis assessment (CDA) has recently received much attention in educational and psychological assessment (Rupp & Templin, 2008). Compared to classical test theory and item response theory (IRT), which only provide an overall score to indicate the information about the position of one individual relative to others on one specific latent trait (de la Torre & Chiu, 2016), CDA can provide detailed information about the strengths and weaknesses of individuals for specific content domains. Consequently, efficient remediation can be conducted based on the fine-grained information available about individuals (Gierl, Leighton, & Hunka, 2007; Lim & Drasgow, 2017; Sawaki, Kim, & Gentile, 2009).

One important research area in CDA is cognitive diagnostic computerized adaptive testing (CD-CAT; Cheng, 2009; McGlohen & Chang, 2008; X. Xu, Chang, & Douglas, 2003). CD-CAT combines a cognitive diagnostic model (CDM) and computer technology to improve testing efficiency and measurement accuracy. Like IRT-based CAT, CD-CAT has compelling advantages over traditional paper-and-pencil (P&P) tests. For example, the performance of individuals can be estimated immediately after they provide a response to each item (Cheng & Chang, 2009). CD-CAT can also provide equivalent or higher accuracy in the measurement of an individual's latent skills, with reductions in test length.

The primary goal of CD-CAT is to improve the measurement accuracy of individuals (Zheng & Chang, 2016) and the item selection method is one of the most important keys to this. Numerous item selection methods have been proposed, such as the Kullback-Leibler method (KL; X. Xu et al., 2003), the Shannon Entropy method (Tatsuoka, 2002), the posterior-weighted KL method (PWKL; Cheng, 2009), the mutual information method (MI; Wang, 2013), and the modified PWKL method (MPWKL; Kaplan, de la Torre, & Barrada, 2015). Recently, Zheng and Chang (2016) developed two new item selection methods designed for short-length tests: the posterior-weighted cognitive diagnostic index (PWCDI) and the attribute-level discrimination index (PWADI), based on previous work by Henson & Douglas (2005) and Henson, Roussos, Douglas, & He (2008).

In addition to the item selection method, the coverage for each attribute can also impact the measurement accuracy. Cheng (2010) indicated that attribute coverage influences both measurement accuracy and reliability, and it is important to make sure that each attribute is measured adequately to ensure the validity of the inferences based on the test. Therefore, she used the modified maximum global discrimination index (MMGDI) method, first used in IRT-based CAT by Cheng and Chang (2009), to balance the attribute coverage and improve measurement accuracy. The simulation study showed that, compared with the original KL method, the MMGDI method produced a relatively higher attribute correct classification rate (ACCR) and pattern correct classification rate (PCCR).

When the minimum number of items that measure each attribute is not satisfied, the attribute balance index (ABI) used in Cheng (2010) tends to select items with a single attribute (Mao & Xin, 2013), which means that the ABI is suitable when the item pool is composed of many items that measure a single attribute. Measurement accuracy would however be lower if the item pool is comprised of many items that measure multiple attributes. Although a test with single-attribute items can produce high PCCR in the CDA framework (e.g. Madison & Bradshaw, 2015; Wang, 2013), it is difficult to construct such items because more than one attribute is required to successfully solve items in real testing situations (DeCarlo, 2011; Huang, 2018). An extreme case is when there are hierarchical relationships among attributes (Leighton, Gierl, & Hunka, 2004), where the ABI tends to produce low measurement accuracy. In addition, the ABI has only been used with the KL method and its performance with other item selection methods is unknown. Therefore, the current study proposes a new method — the modified ratio of test length to the number of attributes (RTA), influenced by the study conducted by Kuo, Pai, and de la Torre (2016) — to balance attribute coverage and improve measurement accuracy when the item pool comprises many multiple-attribute items. Furthermore, the study examines whether the RTA and ABI can be extended to more types of item selection methods.

The remainder of the paper is organized as follows: First, we will introduce the two CDMs used in the study and summarize the item selection methods used. After that, the ABI and RTA will be presented. Then, a simulation study is conducted to examine the RTA with respect to the correct classification rate conditional on several manipulated factors. Finally, the discussion and conclusions are presented.

Cognitive diagnostic models and item selection methods

Numerous CDMs have been proposed to deal with different test situations and with CD-CAT, the ‘‘Deterministic Input, Noisy ‘And’ Gate’’ (DINA) model (Junker & Sijtsma, 2001) and the Reduced Reparameterized Unified Model (RRUM; Hartz, 2002) are commonly used (e.g., Chen, Xin, Wang, & Chang, 2012; Cheng, 2010; Huebner, Finkelman, & Weissman, 2018; G. Xu, Wang, & Shang, 2016). Let α_{ik} denote the mastery of attribute k for individual i and q_{jk} denote if the attribute k is required to answer item j correctly. The item response function (IRF) of the DINA model is then

$$P\left(x_{ij}=1\mid\eta_{ij}\right)=\left(1-s_j\right)^{\eta_{ij}}g_j^{1-\eta_{ij}},$$

where $\eta_{ij}=\prod_{k=1}^{K}\left(\alpha_{ik}\right)^{q_{jk}}$ and s_j and g_j are item parameters. With the RRUM, the IRF is

$$P\left(x_{ij}=1\mid\boldsymbol{\alpha}_i\right)=\pi_j^*\prod_{k=1}^{K}r_{jk}^{*\,(1-\alpha_{ik})q_{jk}},$$

where π_j^* and r_{jk}^* are the item parameters. The item selection method plays an important role in CD-CAT and is the main determinant of ACCR and PCCR. This study uses the four item selection methods MI (Wang, 2013), MPWKL (Kaplan et al, 2015), PWCDI and PWADI (Zheng and Chang, 2016). For details on the interpretation of the cognitive diagnostic model parameters and the item selection methods, we refer to the supplementary material.

Attribute coverage indices

ABI. The ABI was proposed to make sure that each attribute was measured adequately to improve the correct classification rate (Cheng, 2010). It is defined as

$$ABI_j=\prod_{k=1}^{K}\left(\left(B_k-b_k\right)/B_k\right)^{q_{jk}},$$

where B_k is the minimum number of items that should measure the k^{th} attribute and b_k is the number of items that have already been selected to measure the k^{th} attribute.

RTA. Kuo et al. (2016) proposed the RTA to ensure that each attribute is adequately measured when constructing a P&P cognitive diagnostic test. In this paper, we extend this method to CD-CAT and modify it to balance the attribute coverage. The RTA in a CD-CAT context can be written as

$$RTA_j=\frac{1}{1+I\left(H\le B_k\right)\sum_{v=1}^{V}I\left(\boldsymbol{q}_j=\boldsymbol{q}_v^*\right)},\ H=\min\left(b_1,b_2,\cdots,b_K\right),$$

where V refers to the number of items that have already been selected; $I(\cdot)$ is the indicator function; and $\boldsymbol{q}_j$ and $\boldsymbol{q}_v^*$ are the q-vectors of items that have not been and have already been given to a specific person, respectively.

The term $I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ controls the usage of items that measure different numbers of attributes, and the relationship between H and B_k strives to ensure that each attribute is measured at least B_k times. If $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ is larger than 0 and H is no larger than B_k, then the value of $I\left(H \le B_k\right)\sum_{v=1}^{V} I\left(\boldsymbol{q}_j = \boldsymbol{q}_v^*\right)$ tends to be large. Consequently, the RTA becomes small and the j^{th} item will not be selected. Instead, items with different attribute patterns to the previously selected items will tend to be selected. On the other hand, when H is larger than B_k or $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ is 0, then the RTA is equal to 1. In such a case, RTA will not affect the item selection method and therefore the items will then be selected based on the original item selection method.

The RTA criterion balances the attribute coverage and prefers multiple-attribute items. On the contrary, the ABI criterion balances the attribute coverage and prefers single-attribute items. Note that the RTA is determined by both H and $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$, which means that, if H is larger than B_k (or $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ is 0), then $\sum_{v=1}^{V} I(\boldsymbol{q}_{\mathrm{j}} = \boldsymbol{q}_v^*)$ (or H) can be ignored. Therefore, the RTA criterion may not guarantee that each attribute is covered completely. In sum, we expect ABI to perform better than RTA regarding attribute coverage given a long enough test, with RTA performing better than ABI regarding measurement accuracy when the item pool contains many multiple-attribute items. Item selection methods that consider both the attribute coverage and the information that an item provides can be developed by multiplying the attribute coverage indices (ABI or RTA) and the original item selection methods, for example the MMGDI can be obtained by the multiplication ABI × KL.

Simulation study

The goals of the simulation study are to examine the performance of the new attribute coverage index and examine whether the RTA and ABI can be extended to other item selection methods. Several factors are manipulated: model type, number of attributes, Q-matrix structure, test length, attribute coverage index, and item selection method. In total there are 2 (model type) × 2 (number of attributes) × 2 (Q-matrix structure) × 3 (test length) × 3 (attribute coverage index) × 4 (item selection method) = 288 conditions in the study. The details of the simulation study are given in the following.

Model type. Both the DINA model and the RRUM will be used in the current study since these two CDMs are commonly used in CD-CAT (e.g., Cheng, 2010; Huebner et al., 2018; Mao & Xin, 2013; G. Xu et al., 2016).

Number of attributes. Wang (2013) and Zheng and Chang (2016) used five attributes in their studies, while Cheng (2010) used six attributes in her study. In the current study, both five and six attributes are considered to examine the performance of RTA and ABI.

Q-matrix structure. Two types of Q-matrix are generated in this study, namely simple structure and complex structure (Chen et al., 2012; Huang, 2018; Wang, 2013). For the simple structure Q-matrix, all items are unidimensional, meaning that each item measures a single attribute. This Q-matrix is generated based on a discrete uniform distribution with equal probability for all possible patterns. Meanwhile, for the complex structure Q-matrix between one and three attributes are measured by each item. The generation of the complex structure Q-matrix is based on Chen et al. (2012) and can be summarized as follows. First, three basic matrix units are generated. The first matrix unit is a K-by-K identity matrix, while the second and third matrix units are comprised of all possible q-vectors that measure two and three attributes, respectively. Second, the first matrix unit is replicated twenty times while the second and third matrix units are replicated ten times. This results in 100 items that each measure one, two, and three attributes, respectively. Third, the items are merged to create a 300-by-K matrix, and the rows of the 300-by-K matrix are randomly re-ordered.

Test length. Three different test lengths (10, 20, and 30 items) will be used in this study. We view these as short-length, moderate-length, and long-length tests, similar to previous research (e.g., Kuo et al., 2016).

Attribute coverage index (ACI). Three types of ACI will be used in the study. The first type is the original item selection method without attribute coverage control (abbreviated to ORI), which can be treated as the baseline. The second type is the ABI proposed by Cheng (2010) and the last type is the RTA which is proposed in the current study.

Item selection method. The item selection methods used in this study are the MI, MPWKL, PWADI, and PWCDI methods. All these methods can produce high correct classification rates even for short-length test.

Since the generation of the $\boldsymbol{\alpha}$-matrix for five and six attributes are the same, we will only describe the generation of the $\boldsymbol{\alpha}$-matrix for five attributes. A 1000-by-5 matrix is generated to represent the true attribute patterns ($\boldsymbol{\alpha}$-matrix). Each individual can master each attribute with probability equal to .5 and we assume independence among individuals and independence among attributes in the $\boldsymbol{\alpha}$-matrix. For the item parameters, both slipping and guessing parameters were generated from a uniform distribution $U(.05, .30)$ for the DINA model, and the baseline and penalty parameters were generated from $U(.65, .95)$ and $U(.05, .50)$, respectively, for the RRUM. During the item selection procedure, the minimum number of items that measure each attribute was set to 3 because previous studies demonstrated that each attribute should be measured at least three times in the CDA framework (e.g. Fang, Liu, & Ying, 2019; Gu & G. Xu, 2019; G. Xu, 2017). Finally, the expected a posteriori (EAP) method is used to estimate the attribute patterns. Twenty replications for each condition are used in current study.

The evaluation criteria used in this study are averaged ACCR (A-ACCR), PCCR, and the usage of *k*-attribute items (Kuo et al., 2016). These statistics are calculated by

$$PCCR = \sum_{i=1}^{N} I\left(\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i\right) \Big/ N,$$

$$A\text{-}ACCR = \sum_{i=1}^{N}\sum_{k=1}^{K} I\left(\hat{\alpha}_{ik} = \alpha_{ik}\right) \Big/ \left(N \times K\right), \text{ and}$$

$$Usage_k = \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{J} I\left(\sum\nolimits_{h=1}^{K} q_{ijh}^{*} = k\right) \Big/ \left(N \times J\right), k = 1, 2, \cdots, K,$$

where N and J are the number of individuals and test length, respectively; $I(\cdot)$ is the indicator function, which will be 1 if $\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i$ (or $\hat{\alpha}_{ik} = \alpha_{ik}$) is true, and vice versa; $\hat{\boldsymbol{\alpha}}_i$ and $\boldsymbol{\alpha}_i$ are the estimated and true values of an individual's attribute pattern, respectively; q_{ijh}^{*} is the h^{th} entry of q-vector for item *j* that has already been answered by individual *i*. In addition, the empirical standard errors (SEs) for PCCR and A-ACCR, $SE = \sqrt{\frac{1}{n_{sim}-1}\sum_{i=1}^{n_{sim}}\left(\hat{\theta}_i - \bar{\theta}\right)^2}$ (where n_{sim} is the number of replications, $\hat{\theta}_i$ and $\bar{\theta}$ are the i^{th} estimation and the mean value of PCCR and ACCR, respectively), are calculated to evaluate the uncertainty of these two indices (Morris, White, & Crowther, 2019).

Table 1. Correct classification rate for the DINA model (K = 5)

Item selection method	Q matrix	Attribute coverage index	*J* = 10				*J* = 20				*J* = 30			
			PCCR		A-ACCR		PCCR		A-ACCR		PCCR		A-ACCR	
			Est	SE	Este	SE	Est	SE	Est	SE	Est	SE	Est	SE
MI	simple	ORI	.752	.015	.945	.003	.891	.013	.978	.003	.953	.007	.991	.001
		ABI	.704	.011	.932	.003	.920	.013	.984	.003	.958	.007	.992	.001
		RTA	.712	.015	.934	.004	.917	.008	.983	.002	.959	.006	.992	.001
	complex	ORI	.694	.014	.926	.004	.881	.012	.974	.003	.948	.005	.989	.001
		ABI	.699	.008	.931	.003	.901	.015	.979	.004	.959	.008	.991	.002
		RTA	.695	.013	.924	.004	.868	.011	.970	.003	.952	.006	.990	.001
MPWKL	simple	ORI	.844	.012	.966	.003	.986	.003	.997	.001	.999	.001	1.00	.000
		ABI	.835	.010	.964	.002	.980	.006	.996	.001	.999	.001	1.00	.000
		RTA	.838	.007	.965	.002	.987	.004	.997	.001	.999	.001	1.00	.000
	complex	ORI	.860	.006	.966	.002	.988	.003	.997	.001	.998	.001	1.00	.000
		ABI	.798	.014	.955	.003	.982	.004	.996	.001	.998	.001	1.00	.000
		RTA	.852	.014	.963	.004	.986	.004	.997	.001	.998	.002	1.00	.000
PWADI	simple	ORI	.847	.015	.967	.003	.987	.004	.997	.001	.999	.001	1.00	.000
		ABI	.831	.011	.964	.002	.979	.004	.996	.001	.999	.001	1.00	.000
		RTA	.845	.013	.966	.003	.985	.005	.997	.001	.999	.001	1.00	.000
	complex	ORI	.833	.011	.954	.004	.981	.004	.995	.001	.997	.001	.999	.000
		ABI	.789	.014	.952	.003	.982	.004	.996	.001	.998	.002	.999	.000
		RTA	.827	.014	.951	.005	.980	.004	.995	.001	.998	.001	1.00	.000
PWCDI	simple	ORI	.843	.014	.966	.003	.989	.003	.998	.001	.999	.001	1.00	.000
		ABI	.824	.013	.962	.003	.980	.003	.996	.001	.999	.001	1.00	.000
		RTA	.843	.013	.966	.003	.988	.004	.997	.001	.999	.001	1.00	.000
	complex	ORI	.858	.011	.965	.003	.985	.004	.997	.001	.998	.001	1.00	.000
		ABI	.803	.011	.956	.003	.983	.002	.996	.001	.998	.002	1.00	.000
		RTA	.846	.016	.960	.004	.984	.003	.996	.001	.998	.001	1.00	.000

Note. MI refers to mutual information method; MPWKL refers to modified posterior-weighted Kullback-Leibler method; PWADI refers to posterior-weighted attribute-level discrimination index; and PWCDI refers to posterior-weighted cognitive diagnostic index; ORI refers to original item selection method without attribute coverage control; ABI refers to Cheng's (2010) method; RTA refers to the ratio of test length to the number of attributes; PCCR refers to pattern correct classification rate; A-ACCR refers to averaged attribute correct classification rate; Est refers to the estimate, SE is standard error.

Results

Correct classification rate

Tables 1 and 2 present the correct classification rates and the corresponding empirical standard errors (SEs) for the DINA model and the RRUM, respectively, conditional on five attributes. Table 1 shows that all three attribute coverage indices produce similar PCCRs and A-ACCRs for the long-length test (*J* = 30). When test lengths are short (*J* = 10) and moderate

(J = 20), some differences are found between ORI, ABI, and RTA. The ABI, in general, produces higher PCCRs and A-ACCRs than ORI and RTA for moderate- and long-length tests with the MI method, while the RTA performs as well as or even better than ABI with other three item selection methods regardless of test length and Q-matrix structure. To examine which factors (attribute coverage index, test length, and Q-matrix structure) have a significant effect on the measurement accuracy, four repeated measures ANOVAs are conducted for the item selection methods, respectively. Results show that all main effects, second- and third-order interaction effects are statistically significant for the MI method, the partial etas (η_p^2) range from .085 (attribute coverage index) to .996 (test length), and the ABI performs significantly better than RTA for complex-structure Q-matrix and moderate- and long-length tests. For short-length tests, RTA produces significantly higher PCCR than ABI for a simple-structure Q-matrix, while RTA produces relatively lower PCCR than ABI for a complex-structure Q-matrix. With MPWKL and PWADI, all main effects, second- and third-order interaction effects are statistically significant, with the exception of main effect of Q-matrix structure and second-order interaction effect between test length and Q-matrix structure. The η_p^2 range from .101 (interaction effect between attribute coverage index and Q-matrix structure with PWADI method) to .998 (test length with MPWKL method) for the significant effects, and the RTA performs significantly better than ABI for complex-structure Q-matrix and short- and moderate-length tests with both MPWKL and PWADI. Similar to the MI method, all effects are significant for the PWCDI method, and the η_p^2 range from .052 (interaction effect between attribute coverage index and Q-matrix structure) to .997 (test length), and the RTA performs significantly better than ABI for complex-structure Q-matrix and short-length tests and for a simple-structure Q-matrix and moderate-length tests. In addition, the empirical SEs are small for all conditions, indicating that the estimates of PCCRs and A-ACCRs are stable.

Table 2. Correct classification rate for the RRUM (K = 5)

Item selection method	Q matrix	Attribute coverage index	*J* = 10				*J* = 20				*J* = 30			
			PCCR		A-ACCR		PCCR		A-ACCR		PCCR		A-ACCR	
			Est	SE	Este	SE	Est	SE	Est	SE	Est	SE	Est	SE
MI	simple	ORI	.745	.014	.943	.003	.892	.014	.977	.003	.955	.007	.991	.002
		ABI	.697	.012	.931	.003	.909	.008	.981	.002	.958	.005	.991	.001
		RTA	.698	.015	.930	.003	.912	.009	.982	.002	.957	.005	.991	.001
	complex	ORI	.705	.013	.930	.004	.872	.009	.971	.002	.943	.008	.988	.001
		ABI	.689	.012	.928	.003	.884	.012	.975	.003	.943	.009	.988	.002
		RTA	.697	.016	.927	.004	.862	.011	.969	.003	.938	.007	.987	.002
MPWKL	simple	ORI	.836	.013	.965	.003	.984	.004	.997	.001	.998	.001	1.00	.000
		ABI	.824	.009	.962	.002	.977	.005	.995	.001	.998	.001	1.00	.000
		RTA	.835	.011	.964	.003	.984	.004	.997	.001	.998	.001	1.00	.000
	complex	ORI	.849	.010	.965	.003	.980	.004	.996	.001	.997	.002	.999	.000
		ABI	.784	.016	.951	.004	.976	.005	.995	.001	.996	.002	.999	.000
		RTA	.841	.010	.962	.002	.976	.006	.995	.001	.996	.002	.999	.000
PWADI	simple	ORI	.831	.011	.963	.002	.985	.003	.997	.001	.998	.001	1.00	.000
		ABI	.825	.013	.962	.003	.978	.004	.995	.001	.998	.001	1.00	.000
		RTA	.832	.012	.964	.003	.983	.005	.997	.001	.998	.001	1.00	.000
	complex	ORI	.836	.011	.959	.003	.980	.004	.995	.001	.996	.002	.999	.000
		ABI	.768	.011	.948	.003	.973	.006	.994	.002	.996	.002	.999	.000
		RTA	.831	.012	.959	.003	.978	.003	.995	.001	.995	.002	.999	.001
PWCDI	simple	ORI	.839	.013	.965	.003	.984	.004	.997	.001	.998	.002	1.00	.000
		ABI	.826	.011	.962	.003	.977	.005	.995	.001	.998	.001	1.00	.000
		RTA	.829	.012	.963	.003	.982	.004	.996	.001	.998	.001	1.00	.000
	complex	ORI	.842	.009	.963	.002	.981	.005	.996	.001	.997	.001	.999	.000
		ABI	.780	.012	.951	.003	.972	.005	.994	.001	.995	.003	.999	.001
		RTA	.835	.011	.961	.003	.975	.004	.995	.001	.995	.003	.999	.001

The results in Table 2 exhibit a similar pattern to that observed with the RRUM model: the ABI performs better than ORI and RTA for moderate- and long-length tests for the MI method while it performs worse for short-length tests. In addition, both RTA and ORI produce larger PCCRs than ABI for short- and moderate-length tests for the MPWKL, PWADI, and PWCDI methods. Moreover, all of these three attribute coverage indices produce very similar PCCRs when the test length is long. Furthermore, the RTA produces a lower A-ACCR than

ABI for the MI method, while it produces an identical or larger A-ACCR than ABI for most conditions. All main effects and second- and third-order interaction effects are statistically significant, with the exception of the second-order interaction effect between test length and Q-matrix structure for the MPWKL method, and the η_p^2 range from .046 (interaction effect between attribute coverage index and Q-matrix structure with MI method) to .998 (test length with PWADI method) for the significant effects. Finally, the empirical SEs are small and the corresponding estimates are stable.

The PCCR and A-ACCR for six attributes are presented in the supplementary material and the results can be summarized as follows: (1) The ABI, in general, produces higher PCCRs and A-ACCRs than RTA for the MI method; (2) the RTA and ORI methods produce higher PCCRs and A-ACCRs than ABI with the MWPKL, PWADI, and PWCDI methods regardless of Q-matrix structure and test length; (3) all the third-order interaction effects are significant, and the η_p^2 range from .251 (for RRUM and PWCDI method condition) to .534 (for DINA model and MWPKL method condition); (4) with the increase of test length, the SEs are decreased for all conditions.

The usage of items

Since all of the items in the simple-structure Q-matrix are single-attribute, all item selection methods select single-attribute items, which results in no differences in the usage of items that measure *k*-attributes for ORI, ABI, and RTA. Therefore, the details will not be presented. Table 3 presents the usage of items that measure *k*-attributes for five attributes and with the complex-structure Q-matrix. The usage of items that measure *k*-attributes for six attributes is consistent with the results with five attributes, and the details can be accessed in the supplementary material. Unsurprisingly, the RTA method selects the least items that measure a single attribute in most of the conditions, followed by the ORI method. The ABI method uses the most items that measure a single attribute. Specifically,

Table 3. The usage of items measures *k*-attribute for five attributes and complex Q-matrix

model	item selection method	attribute coverage index	$J = 10^a$			$J = 20^a$			$J = 30^a$		
			1-A	2-As	3-As	1-A	2-As	3-As	1-A	2-As	3-As
DINA	MI	ORI	.489	.381	.131	.507	.337	.156	.528	.306	.166
		ABI	.864	.063	.074	.738	.178	.084	.633	.245	.122
		RTA	.420	.433	.147	.436	.397	.167	.490	.339	.171
	MPWKL	ORI	.468	.366	.166	.400	.373	.227	.408	.358	.233
		ABI	.888	.077	.036	.671	.212	.117	.540	.291	.168
		RTA	.435	.396	.169	.377	.393	.230	.397	.369	.233
	PWADI	ORI	.368	.406	.226	.360	.388	.253	.386	.368	.246
		ABI	.833	.130	.037	.622	.243	.135	.513	.307	.180
		RTA	.359	.421	.220	.346	.402	.253	.379	.376	.245
	PWCDI	ORI	.431	.385	.184	.387	.377	.236	.404	.358	.238
		ABI	.882	.083	.035	.665	.215	.119	.538	.291	.171
		RTA	.405	.408	.187	.367	.395	.239	.394	.367	.239
RRUM	MI	ORI	.480	.395	.125	.443	.396	.161	.430	.392	.178
		ABI	.932	.032	.036	.729	.189	.082	.574	.299	.127
		RTA	.410	.411	.179	.377	.422	.200	.396	.406	.198
	MPWKL	ORI	.492	.355	.153	.427	.381	.192	.413	.381	.206
		ABI	.875	.090	.035	.662	.231	.107	.524	.316	.160
		RTA	.434	.390	.175	.385	.408	.207	.399	.391	.211
	PWADI	ORI	.434	.380	.186	.402	.392	.205	.400	.387	.213
		ABI	.829	.135	.036	.622	.260	.118	.504	.330	.166
		RTA	.406	.393	.201	.372	.408	.219	.392	.391	.217
	PWCDI	ORI	.475	.365	.160	.418	.385	.197	.409	.382	.208
		ABI	.866	.099	.036	.653	.236	.111	.519	.319	.163
		RTA	.427	.391	.182	.381	.408	.212	.396	.391	.213

Note. *k*-A(s) means items measure *k* attribute(s);

[a] 4-As and 5-As equal to 0 for all conditions.

the proportion of items that measure a single attribute ranges from .346 to .490, .360 to .528, and .504 to .930 for the RTA, ORI, and ABI criteria, respectively. In addition, among these three attribute coverage indices, the RTA method produces the largest proportions of items that measure two and three attributes, followed by the ORI method, and the ABI method yields the smallest proportions of items that measure two and three attributes. These results can be expected since the RTA criterion tends to choose items that measure different attributes to the already administered items. The ABI criteria, on the contrary, tends to penalize items that measure multiple attributes by taking the product of deviances for all attributes. Consequently, items that measure a single attribute tend to be selected by the ABI criteria.

Coverage of attributes

Table 4 lists the proportion of individuals who have been administered at least three times measuring each attribute for moderate- and long-length tests. The results are omitted for the short-length test (i.e. J = 10) because all three attribute coverage indices do not satisfy the attribute coverage requirement. The ABI can ensure that most of the tests satisfy the attribute coverage regardless of number of attributes, model type, Q-matrix structure, item selection method, and test length, while RTA performs worse than ABI but better than the ORI. Repeated measures ANOVAs are conducted to investigate the differences among ORI, ABI, and RTA. The results show that most main effects, second-, third- and fourth-order interaction effects are significant under the DINA model, and most of the η_p^2 are larger than .50. Although the differences between ABI and RTA are significant for some conditions, the η_p^2 range from .001 to .114, which indicates that stronger evidence is needed to support differences between ABI and RTA. For the RRUM, all main effects and second-, third- and fourth-order interaction effects are significant, and the η_p^2 range from .797 to .999. In addition, all of the main effects of attribute coverage index, test length, and number of attributes are significant for all item selection methods, and all of the η_p^2 are larger than .950. Although the fourth-order interaction effects are significant for all item selection methods, the partial etas are small and range from .002 to .065. Furthermore, the third-order interaction effects among attribute coverage

Table 4. Overall percentage for moderate- and long-length tests

Number of attributes	Model type	Q matrix	Attribute coverage index	*J* = 20				*J* = 30			
				MI	MPWKL	ADI	CDI	MI	MPWKL	ADI	CDI
K = 5	DINA	simple	ORI	.357	.846	.842	.851	.881	.997	.998	.997
			ABI	.986	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.892	.891	.892	1.00	.999	1.00	.999
		complex	ORI	.772	.940	.956	.945	.974	.998	.999	.998
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.965	.976	.969	1.00	1.00	1.00	1.00
	RRUM	simple	ORI	.326	.812	.813	.817	.867	.996	.996	.996
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.872	.876	.871	1.00	.999	.999	.999
		complex	ORI	.795	.910	.919	.910	.977	.996	.995	.995
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.981	.984	.982	1.00	1.00	1.00	1.00
K = 6	DINA	simple	ORI	.034	.468	.464	.463	.495	.983	.984	.985
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	1.00	.516	.507	.507	1.00	.993	.959	.961
		complex	ORI	.714	.775	.823	.793	.971	.988	.988	.987
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	.935	.828	.860	.834	1.00	.995	.987	.983
	RRUM	simple	ORI	.020	.327	.318	.322	.430	.957	.953	.957
			ABI	1.00	1.00	1.00	1.00	.993	1.00	1.00	1.00
			RTA	1.00	.427	.357	.418	1.00	.985	.952	.922
		complex	ORI	.569	.721	.755	.733	.865	.974	.981	.977
			ABI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			RTA	.900	.804	.794	.805	1.00	.999	.986	.972

Note. The results are omitted for the short-length test (i.e. *J* = 10) because all three attribute coverage indices do not satisfy the attribute coverage requirement.

index, number of attributes, and test length are significant for all item selection methods, and the corresponding η_p^2 are at the range of .969 and .980, and the ABI performs better than the RTA at six attributes and moderate-length tests.

Discussion and conclusions

The goals of this study are to develop a new attribute coverage method, RTA, to deal with empirical situations when more than one attribute is involved in successfully solving a test item (DeCarlo, 2011; Huang, 2018) and to examine the performance of both ABI and RTA when different item selection methods are used. A simulation study is conducted to examine the performance of RTA and ABI, and promising results are produced.

The results show that the RTA produces lower PCCRs than ABI for moderate- and long-length tests with the MI method, especially with a complex structure Q-matrix. On the contrary, the RTA produces relatively high PCCRs than the ABI for short- and moderate-length tests with the MPWKL, PWADI, and PWCDI methods. A possible explanation is that both the MI method and the ABI criterion prefer single-attribute items, while the RTA and three other item selection methods tend to use fewer single-attribute items than ABI and MI method. As Madison and Bradshaw (2015) and Huebner et al. (2018) demonstrated, the more single-attribute items there are in a test, the higher the measurement accuracy is for long-length tests. Therefore, the RTA can be expected to produce lower measurement accuracy since fewer single-attribute items are used for the MI method. As for the MPWKL, PWADI, and PWCDI methods, the differences between the usage of items that measure one and two attributes are small, meaning that these item selection methods prefer items that measure either one or two attributes. Therefore, when the ABI criteria, which prefers the single-attribute items, is added to these three item selection methods, information provided by two-attribute items may be lost and, consequently, lower measurement accuracy is produced for the ABI compared to the ORI and RTA criteria. Meanwhile, a possible reason why the ABI performs worst in most conditions for short-length tests (J = 10) is that it is hard to satisfy the minimum number of items that measure each attribute when the test length is short. Although previous studies demonstrated that tests containing more single-attribute items tend to produce higher measurement accuracy (Huebner et al., 2018; Madison & Bradshaw, 2015), the prerequisite for a high measurement accuracy is that the test length is long enough.

Moreover, the results show that the ABI is not suitable for all item selection methods. In the current study, the ABI is suitable for the MI method, while it is unsuitable for the MPWKL, PWADI, and PWCDI methods. In the study of Cheng (2010), the combination between ABI and KL method (MMGDI) can produce higher measurement accuracy than the original KL method (MGDI). Since both the ABI criterion and KL/MI methods prefer single-attribute items rather than multiple-attribute items, using the ABI criterion further reinforces the tendency of the KL and MI methods to select single-attribute items. Hence, the combination between the ABI criterion and the original item selection methods would produce high measurement accuracy if the original item selection methods prefer single-attribute items. On the flipside, low measurement accuracy would be produced if more than one attribute is preferred by the original item selection methods (e.g. MPWKL, PWADI and PWCDI).

It's worth noting that, although the RTA criteria produces higher measurement accuracy than the ABI criteria with the MPWKL, PWADI, and PWCDI methods, this does not indicate that the RTA performs better than ABI for all situations. By examining the ABI and RTA criteria, the ABI tends to penalize items that measure multiple attributes, while the RTA tends to select items that measure multiple attributes. Therefore, it is reasonable to infer that the composition of items that measure different number of attributes in the item pool have an important influence on these two criteria. The RTA performs better than ABI if there is a large number of multiple-attribute items in the item pool. Meanwhile, the ABI performs better than RTA if there is a majority of single-attribute items, producing higher measurement accuracy than RTA for all conditions.

The results also show that the ABI performs better than the RTA for moderate- and long-length tests concerning the attribute coverage, which coincides with our expectation. As stated previously, the formulation of the RTA is determined by two components. One is used to control the usage of items that measure different numbers of attributes and the other is used to control the attribute coverage. When one of the components is satisfied, the other component is ignored. For instance, when the summation of the first component is zero, the component that controls the attribute coverage is ignored and consequently the attribute coverage will not be satisfied.

In conclusion, the new attribute coverage control method—RTA—is suitable for controlling the attribute coverage and producing acceptable measurement accuracy when the item pool is comprised of a large number of items that measure multiple attributes, which is a common phenomenon in empirical testing situations (DeCarlo, 2011; Huang, 2018). The ABI, on the other hand, is appropriate for test situations when the majority of an item pool is comprised of single-attribute items. Furthermore, the ABI is suitable for item selection methods that prefer single-attribute items, such as the KL method (Cheng, 2010) and the MI method, but is not suitable for methods that prefer both single- and multiple- attributes items such as the MPWKL, PWADI, and PWCDI methods.

Although some promising results are found in the current study, several remaining open issues deserve further studies. First, we assume that the minimum number of items that measure each attribute are the same for all attributes. Considering that different attributes may carry different importance, this is not a necessary constraint and further studies can take the importance of each attribute into consideration to further investigate the performance of attribute coverage methods in CD-CAT. Second, fixed-length tests were used in the current study. Therefore, everyone was administered the same test length. Future studies can examine the performance of RTA when the test length is different for each individual (variable-length tests). Third, both the DINA model and the RRUM are specific CDMs and some constraints imposed on these specific CDMs are (a) only a single model is available across the entire test and (b) either compensatory or non-compensatory relationships is assumed for the test (Ravand, 2016). General CDMs relax these constraints and therefore a general CDM can be used in future studies.

Acknowledgment

The authors would like to thank the Editor in Chief, Dr. John R. Donoghue, the Associate Editor, Dr. Chun Wang, and two anonymous reviewers for their helpful comments on earlier drafts of this article.

Supplemental Material

Supplemental material for this article is available online.

References

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, 77, 201-222.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 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*(6), 902-913.

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 369-383.

DeCarlo, L, T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*(1), 8-26.

de la Torre, J., & Chiu, C. Y. (2016). General method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273.

Fang, G., Liu, J., & Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika, 84*(1), 19-40.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Gu, Y., & Xu, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika, 84*(2), 468-483.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practice.* Unpublished doctoral dissertation, University of Illinois at Urbana Champaign.

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.

Huang, H. Y. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *Journal of Classification, 35*(3), 437-465.

Huebner, A., Finkelman, M. D., & Weissman, A. (2018). Factors affecting the classification accuracy and average length of a variable-length cognitive diagnostic computerized test. *Journal of Computerized Adaptive Testing*, *6*(1). DOI: 10.7333/1802-060101.

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*(3), 167-188.

Kuo, B. C., Pai, H. S., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement*, *40*(5), 315-330.

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

Lim, Y. S., & Drasgow, F. (2017). Nonparametric calibration of item-by-attribute matrix in cognitive diagnosis. *Multivariate behavioral research*, *52*(5), 562-575.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491-511.

Mao, X., & Xin, T. (2013). The application of the Monte Carlo approach to cognitive diagnostic computerized adaptive testing with content constraints. *Applied Psychological Measurement*, *37*(6), 482-496.

McGlohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*(3), 808-821.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine, 38*(11), 2074-2102.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 782-799.

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

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly*, *6*(3), 190-209.

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.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, *73*(6), 1017-1035.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics, 45*(2), 675-707.

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*(3), 291-315.

Xu, X., Chang, H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the 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*(8), 608-624.