Journal of Computerized Adaptive Testing

Volume 6 Number 1

February 2018

Factors Affecting the Classification Accuracy and Average Length of a Variable-Length Cognitive Diagnostic Computerized Test

Alan Huebner, Matthew D. Finkelman, and Alexander Weissman

DOI 10.7333/1802-060101

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

Journal of Computerized Adaptive Testing

Volume 6, Number 1, February 2018

DOI 10.7333/1802-060101

ISSN 2165-6592

Factors Affecting the Classification Accuracy and Average Length of a Variable-Length Cognitive Diagnostic Computerized Test

Alan Huebner
University of Notre Dame

Matthew D. Finkelman
Tufts University, School of Dental Medicine

Alexander Weissman
Law School Admission Council

The aim of cognitive diagnosis models (CDMs) is to provide students and educators with individually tailored diagnostic results for students' mastery levels of a group of fine-grained skills, or attributes. The field of variable-length cognitive diagnostic computerized adaptive testing (CD-CAT) aims to deliver diagnostic assessments that accurately classify students using the fewest number of items possible. A crucial element of a CD-CAT is the Q-matrix, a 1-0 matrix mapping the skills required by each item. This paper describes a simulation study that systematically explored factors affecting the accuracy and average length of a variable-length CD-CAT, including composition of the Q-matrix, correlation among skills, item selection rule, and CDM. It was found that higher density Q-matrices (i.e., Q-matrices in which individual items tap many skills) yield longer and less accurate tests than lower density Q-matrices (i.e., Q-matrices in which individual items tap fewer skills). The two item selection rules examined—mutual information and modified posterior Kullback-Leibler information—performed very similarly. Higher correlation among skills tended to increase average test lengths and decrease accuracy noticeably when the Q-matrices were high density.

Keywords: adaptive testing, classification, cognitive diagnostic models, item selection, variable-length testing

Cognitive diagnostic models (CDMs) are discrete multivariate latent trait models that classify examinees according to whether they have mastered a set of K skills, or attributes. Some authors have suggested that "attributes" is a more general term than "skills." In particular, de la Torre (2009) stated that the term "attributes" subsumes skills and cognitive processes; he used "skills" and "attributes" synonymously in the context of assessing fraction subtraction. Similarly, this paper uses the two terms interchangeably. Specifically, each individual skill is regarded as a binary latent variable, with the two levels being "mastery" and "non-mastery." Thus, for an assessment diagnosing K skills, examinees are classified into one of 2^K possible skill mastery patterns.

CDMs are intended to provide timely and detailed diagnostic information to students, teachers, parents, and school administrators. It has been suggested that CDMs will be most effective in reaching these goals if the diagnostic assessments can be administered quickly and efficiently on a computer. For example, Jang (2008) elaborated upon the benefits of computerized diagnostic testing by describing a hypothetical scenario in which a language teacher completes a unit and creates a computerized diagnostic assessment to yield information about students' mastery on a set of fine-grained skills. Researchers are currently making progress to turn this hypothetical situation into reality.

The field of cognitive diagnostic computerized adaptive testing (CD-CAT) has seen methods proposed for selecting optimal items in diagnostic tests (Cheng, 2009, 2010; Kaplan, de la Torre, & Barrada, 2015; Wang, 2013; Xu, Chang, & Douglas, 2003) and for stopping a variable-length CD-CAT when a reliable classification can be made (Hsu, Wang, & Chen, 2013). Similar to item response theory (IRT)-based CAT, a CD-CAT is thus expected to be more efficient than a pencil-and-paper diagnostic test: At a given stage of the test, the next item is administered based upon the examinee's performance up to that point.

Previous simulation studies have focused attention on factors such as item quality and structure of the Q-matrix, a key element of CDM methodology that is described in more detail below. Briefly, a Q-matrix is a 1-0 mapping of the skills required by each test item and is said to be of higher or lower density depending on whether the individual items comprising it tend to require many or few skills, respectively. The aim of this study was to investigate the ability of a variable-length CD-CAT to accurately classify examinees using a minimal number of items under various testing conditions in which several factors were systematically examined, specifically, Q-matrix density, generating CDM, correlation among skills, and item selection method.

To the authors' knowledge, this is the first study to manipulate these factors simultaneously in the context of variable-length CD-CAT. Wang (2013) varied item quality and Q-matrix structure but in the context of fixed-length CD-CAT. Cheng (2010) focused on item selection for fixed-length tests but did not manipulate item quality or Q-matrix structure. Kaplan et al. (2015) conducted a simulation study that examined both fixed- and variable-length CD-CAT, but the Q-matrix was fixed and correlation among skills was not considered. It is hoped that the findings in this paper will spur discussion and the further development of these models for use in practical operational settings.

Method

Cognitive Diagnostic Models

There are currently many types of models and approaches in the psychometrics literature designed to provide fine-grained diagnostic score reports. For taxonomies, lists, and descriptions of methods, see DiBello, Roussos, and Stout (2007); Rupp and Templin (2008); or Rupp, Templin, and Henson (2010). Although the details differ from one approach to another, one commonality among many of the models is the positing of a vector of latent attributes representing skill masteries, $\boldsymbol{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK})'$, where for $k = 1, \dots, K$,

$$\alpha_{ik} = \begin{cases} 1 & if\ examinee\ i\ has\ mastered\ skill\ k \\ 0 & otherwise. \end{cases} \tag{1}$$

Many CDMs are special cases of restricted latent class models that use an item response function (IRF) to assign probabilities of a correct response conditionally upon the item parameters and $\boldsymbol{\alpha_i}$. The IRF takes the general form $P\left(X_{ij} = 1 \middle| \boldsymbol{\alpha_i}\right)$, where X_{ij} is the random variable denoting the response to item j by examinee i, and

$$X_{ij} = \begin{cases} 1 & if\ examinee\ i\ has\ responded\ correctly\ to\ item\ j \\ 0 & otherwise. \end{cases} \tag{2}$$

In addition, for an assessment with J total items, let $\boldsymbol{X_i}$ be the vector of responses for examinee i to all J items, or $\boldsymbol{X_i} = \left(X_{i1}, \dots, X_{iJ}\right)'$. Though different CDMs will have different IRFs, all depend upon a structure called the Q-matrix. The Q-matrix is a J by K matrix in which the element in the j^{th} row and k^{th} column is defined by

$$q_{jk} = \begin{cases} 1 & item\ j\ taps\ skill\ k \\ 0 & otherwise. \end{cases} \tag{3}$$

Then the entire Q-matrix row for item j, indicating all the skills tapped by that item, is denoted simply as $\boldsymbol{q_j}$. Although the basic concept is relatively simple, the construction, validation, and refinement of a Q-matrix for an operational diagnostic assessment is currently a topic of much discussion (see Chiu, 2013; de la Torre, 2008; and Li & Suen, 2013).

The particular CDMs used in this study are the deterministic inputs, noisy "and" gate model (DINA; Junker & Sijtsma, 2001) and the reduced re-parameterized unified model (RRUM; Hartz, 2002). The DINA is considered one of the simplest CDMs because it allows each item only two parameters, and for a given item the DINA IRF yields only two possible values. This is due to the use of the binary indicator

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}. \tag{4}$$

If examinee i has mastered all the required skills for item j, then $\eta_{ij} = 1$; otherwise, $\eta_{ij} = 0$. The DINA IRF for examinee i responding to item j is defined as

$$P(X_{ij} = 1|\boldsymbol{\alpha_i}) = (1 - s_j)^{\eta_{ij}} g_j{}^{(1-\eta_{ij})}, \tag{5}$$

where $s_j = P(X_{ij} = 0|\eta_{ij} = 1)$ and $g_j = P(X_{ij} = 1|\eta_{ij} = 0)$. Informally, s_j is the probability that an examinee mastering all required skills for item j suffers a careless "slip" and responds incorrectly, while g_j is the probability that an examinee not mastering at least one required skill makes a "lucky guess."

The RRUM, has an IRF given by

$$P(X_{ij} = 1|\boldsymbol{\alpha_i}) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_k)q_{jk}}. \tag{6}$$

Here, π_j^* is the probability of a correct response from an examinee who has mastered all skills tapped by item j. The r_{jk}^* parameters can be thought of as penalty parameters reducing the probability of a correct response when a tapped skill is not mastered. Specifically, if item j taps skill k and it has not been mastered by the examinee, the probability of a correct response is reduced by a factor of r_{jk}^*. In other words, smaller values of r_{jk}^* result in stricter penalties, i.e., more sharply reduced probabilities of a correct response, when skill k is not mastered. Thus, smaller values of r_{jk}^* indicate a stronger relation between the item and the latent skill it taps. For items tapping k skills, the RRUM requires the estimation of $(k + 1)$ parameters; therefore, it is more complex than the DINA model.

Examinee Classification

CDM item parameters can be estimated using an expectation-maximization (EM) algorithm (de la Torre, 2009, 2011) or Markov chain Monte-Carlo (MCMC) techniques (de la Torre & Douglas, 2004; Henson, Templin, & Willse, 2009). Examinees are then classified into one of the 2^K possible classes, or skill mastery patterns. Different methods of classification have been discussed in the literature. De la Torre (2008) stated that classification can be performed via maximum likelihood estimation (MLE) or expected a posteriori (EAP); Rupp et al. (2010) described EAP and maximum a posteriori (MAP) classification; and Huebner and Wang (2011) conducted a simulation study comparing the accuracy of the three methods under different testing conditions. Examinee classification is briefly reviewed here, as the general concepts are relevant to CD-CAT termination rules described below.

A key element for examinee classification is the likelihood function for responses $\boldsymbol{X_i}$, given by

$$L(\boldsymbol{X_i}; \boldsymbol{\alpha}) = \prod_{j=1}^{J} [P(X_{ij} = 1|\boldsymbol{\alpha})]^{X_{ij}} [1 - P(X_{ij} = 1|\boldsymbol{\alpha})]^{1-X_{ij}}. \tag{7}$$

Let c be an index for the classes, and c' used to denote a particular class. Then, $P(\boldsymbol{\alpha}_{c'})$ is the prior probability of belonging to the particular class c'. If the priors are assumed to be known for all 2^K classes, the classification can incorporate them by applying Bayes' Rule to compute the posterior distributions:

$$\pi_i(\boldsymbol{\alpha}_{c'}) = P(\boldsymbol{\alpha}_{c'}|\boldsymbol{X}_{\boldsymbol{i}}) = \frac{L(\boldsymbol{X}_{\boldsymbol{i}}|\boldsymbol{\alpha}_{c'})P(\boldsymbol{\alpha}_{c'})}{\sum_{c=1}^{2^K} L(\boldsymbol{X}_{\boldsymbol{i}}|\boldsymbol{\alpha}_{\boldsymbol{c}})\, P(\boldsymbol{\alpha}_{\boldsymbol{c}})} \tag{8}$$

Then, $\widehat{\boldsymbol{\alpha}}_{MAP}$, the MAP estimate of $\boldsymbol{\alpha}$, is given by

$$\widehat{\boldsymbol{\alpha}}_{MAP} = \arg\max_{c}\{P(\boldsymbol{\alpha}_{\boldsymbol{c}}|\boldsymbol{X}_{\boldsymbol{i}})\}. \tag{9}$$

The MLE method can be viewed as the MAP method when using a prior distribution in which all $P(\boldsymbol{\alpha}_c)$ are equal. This estimate is denoted as $\widehat{\boldsymbol{\alpha}}_{MLE}$.

CD-CAT

A large part of the research in the field of CD-CAT has been dedicated to item selection criteria. One of the earliest papers to propose an item selection method was by Xu, Chang, and Douglas (2003). They demonstrated for fixed-length CD-CAT that choosing items to maximize a criterion based upon Kullback-Leibler (KL) information or to minimize the Shannon Entropy (SHE) resulted in higher rates of correctly classifying examinees into their true mastery patterns than random item selection. Cheng (2009) proposed the posterior-weighted KL index (PWKL), which was shown to select items leading to higher correct classification rates than the original KL method. Cheng (2010) also proposed a modified version of the KL index that balanced the number of times each skill was tapped in a CD-CAT; this method was termed the maximum modified global discrimination index (MMGDI). Wang (2013) proposed mutual information (MUINF) methods of item selection for CD-CAT that were shown to lead to more accurate classifications than the PWKL in simulated testing conditions. Kaplan et al. (2015) proposed a modified PWKL index (MPWKL) and the generalized DINA discrimination index. The present study compared the MPWL and MUINF criteria for the first time, and thus both will be reviewed briefly.

After t items have been administered to examinee i, i.e., stage t of the test, the MPWKL criterion selects the next item for stage $(t+1)$ as the one maximizing the quantity

$$MPWKL_{ij} = \sum_{d=1}^{2^K}\left\{\sum_{c=1}^{2^K}\left[\sum_{x=0}^{1} P\left(X_j = x\middle|\boldsymbol{\alpha}_d\right)\log\left(\frac{P\left(X_j = x\middle|\boldsymbol{\alpha}_d\right)}{P\left(X_j = x\middle|\boldsymbol{\alpha}_c\right)}\right)\pi_i^{(t)}\left(\boldsymbol{\alpha}_c\right)\right]\pi_i^{(t)}\left(\boldsymbol{\alpha}_d\right)\right\}, \tag{10}$$

where at stage t, $\pi_i^{(t)}(\boldsymbol{\alpha}_c)$ is the estimated latent class posterior probability for attribute vector $\boldsymbol{\alpha}_c$, and $\pi_i^{(t)}(\boldsymbol{\alpha}_d)$ is the estimated latent class posterior probability for attribute vector $\boldsymbol{\alpha}_d$. It is noteworthy that the MPWKL does not depend on $\widehat{\boldsymbol{\alpha}}_i^{(t)}$, the interim classification at stage t, but rather on the entire posterior distribution of the latent classes. This innovation enables the MPWKL to be more informative than the more basic PWKL and KL criteria.

The MUINF criterion also has this property; its formula is given by

$$MUINF_{ij} = \sum_{c=1}^{2^K} \pi_i^{(t)}(\boldsymbol{\alpha}_c) \sum_{x=0}^{1} P(X_j = x|\boldsymbol{\alpha}_c)\, log\left(\frac{P(X_j = x|\boldsymbol{\alpha}_c)}{P(X_j = x)}\right), \tag{11}$$

where $P(X_j = x)$ is the marginal probability of response x. For $x = 1$, this probability is given by

$$P(X_j = 1) = \sum_{c=1}^{2^K} \pi_i^{(t)}(\boldsymbol{\alpha}_c)P(X = 1|\boldsymbol{\alpha_c}). \tag{12}$$

Then, for $x = 0$,

$$P(X_j = 0) = 1 - P(X_j = 1). \tag{13}$$

The only original work on termination rules for variable-length CD-CAT is by Hsu et al. (2013), to the best of the authors' knowledge. Hsu et al. (2013) defined two different criteria:

1. Terminate the test at stage t when the largest $\pi_i^{(t)}(\boldsymbol{\alpha}_c)$ is greater than or equal to some pre-specified probability threshold.
2. Terminate the test when the largest $\pi_i^{(t)}(\boldsymbol{\alpha}_c)$ is greater than or equal to some pre-specified probability threshold and the second largest $\pi_i^{(t)}(\boldsymbol{\alpha}_c)$ is less than or equal to a second smaller probability threshold.

The simulation study described below used the first termination criterion. For the sake of simplicity, the second criterion was not considered further. Moreover, Hsu et al.'s (2010) simulation study demonstrated that the two criteria performed very similarly unless the second probability threshold was set close to zero.

Simulation Study

A simulation study was performed to examine the classification accuracy and average length of a variable-length CD-CAT diagnosing $K = 6$ skills under various testing conditions. Within each condition, $N = 10{,}000$ examinees were administered tests with a maximum length of 30 items. The termination threshold was set to 0.80, and a uniform discrete prior distribution was used for all conditions. Examinees whose tests were not stopped early via the termination rule were classified according to their MAP estimate after the 30th item. Factors manipulated in the study included the generating model (DINA or RRUM), Q-matrix composition (low-, medium-, or high-density), correlation between skills ($\rho = 0.3$ or $\rho = 0.7$), and item selection method (MPWKL or MUINF). These factors were fully crossed, resulting in 2 × 3 × 2 × 2 = 24 total conditions.

Item Banks

Because the RRUM and DINA utilize different parameterizations, two item banks were generated, each consisting of 300 items. For all conditions in which the RRUM was used, the r_{jk}^* parameters were generated following Feng, Habing, and Huebner (2014), i.e., from the Uniform(0.05, 0.40) distribution. The π_j^* parameters were generated from the Uniform(0.75, 0.95) distribution, used by Wang, Chang, and Huebner (2011) for the π_j^* parameters in the Fusion model, which has an IRF similar to the RRUM with π_j^* having the same interpretation. For conditions in which the DINA was used, the slip and guess parameters were generated from the Uniform(0.05, 0.25) distribution, as was done in the simulation by Cheng (2010).

Q-Matrices

To vary the Q-matrix composition, low-, medium-, and high-density Q-matrices were generated so that each skill had a 20%, 40%, and 60% chance, respectively, of being tapped by a given item, and items were required to tap at least one skill. This method of randomly constructing the Q-matrix was also used in Cheng's (2010) simulation study; but that study was designed so that there was only a 20% chance for a skill to be tapped by an item. To induce correlation among skills, a K-length multivariate normal vector was generated for each examinee with

$$\boldsymbol{\mu} = \mathbf{0} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{14}$$

Then, for the i^{th} simulated examinee, the binary elements of $\boldsymbol{\alpha_i}$ were created by setting negative and positive values to 0 and 1, respectively; and the two levels of ρ used were 0.3 and 0.7.

Evaluation Criteria

The results yielded by each condition were evaluated in terms of average test length (ATL) and classification accuracy. Specifically, classification accuracy was measured by the proportion of examinees classified correctly on all six skills (PCC), as well as by the frequency counts of examinees with one, two, three, and four or more skills misclassified.

Results

Statistics summarizing classification accuracy and test length for all conditions are reported in Table 1. Some trends are readily apparent. First, regardless of model, item selection, or skill correlation, the high-density $\boldsymbol{Q}$-matrices produced a sharp drop in PCC and an increase in ATL compared to the low- and medium-density $\boldsymbol{Q}$-matrices. The relation between the low- and medium-density $\boldsymbol{Q}$-matrices is less clear; the medium settings tended to produce higher PCCs as well as higher ATLs than the low density. There was little difference between the PCCs and ATLs yielded by the different values of skill correlation when the $\boldsymbol{Q}$-matrix density was low or medium. However, for high-density $\boldsymbol{Q}$-matrices, the lower correlation value yielded noticeably lower ATLs (approximately one- to one-and-a-half-fewer items) and higher PCCs. For a given combination of generating model, $\boldsymbol{Q}$-matrix density, and skill correlation, the MPWKL and MUINF selection criteria performed very similarly. Figure 1 displays these trends visually for conditions 7–12 and 19–24, i.e., for conditions using selection via the MUINF criterion. Plots for the MPWKL are omitted due to the great similarity with the MUINF.

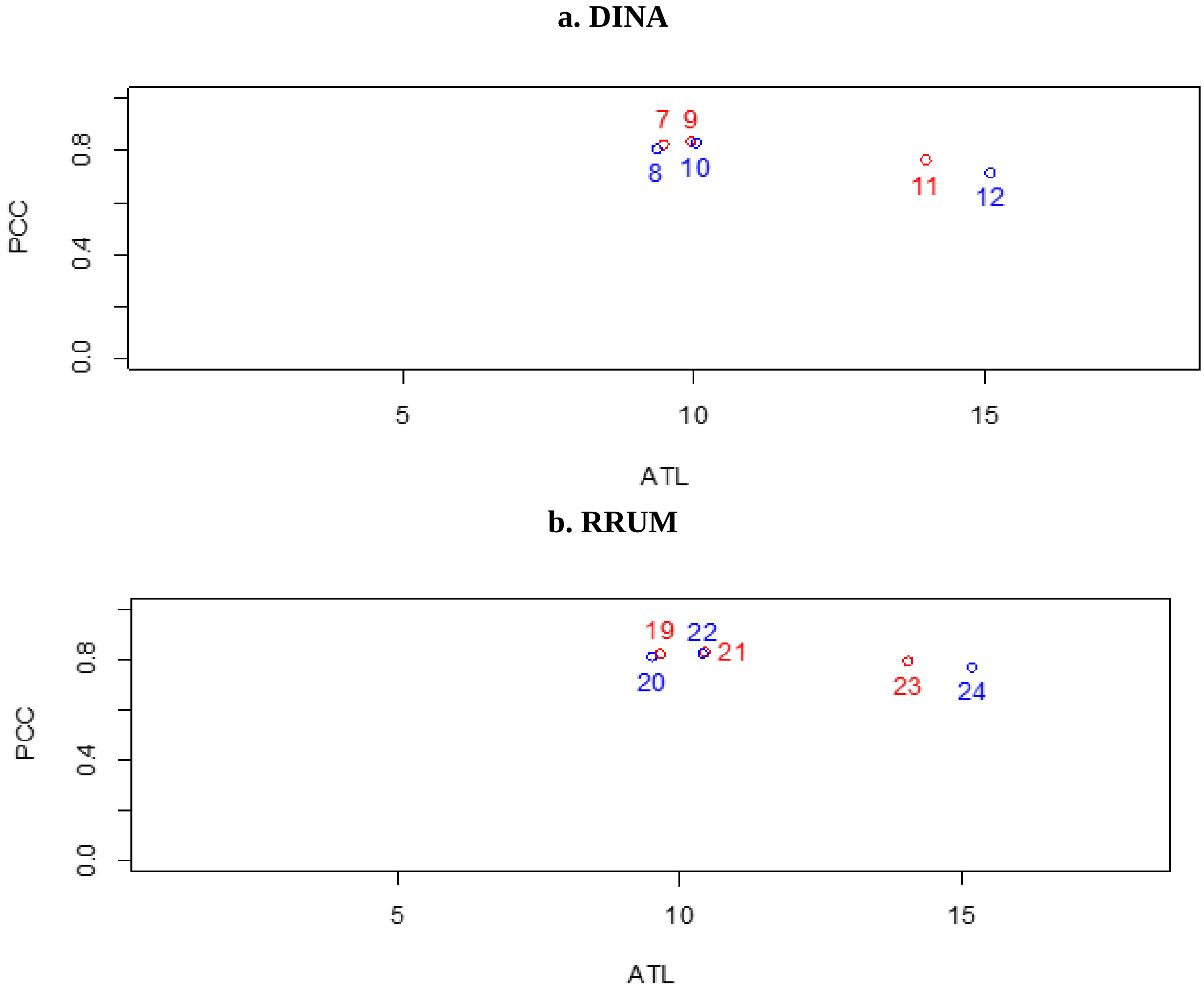
The PCCs in Table 1 indicate the rate at which examinees were classified correctly on all skills, and Table 2 provides insight into the misclassified examinees. Specifically, Table 2 presents the proportion of examinees classified incorrectly on one, two, three, and four or more skills. Again, the high-density Q-matrices fared poorly, as they resulted in far more examinees being classified incorrectly on three or four skills than did the low- and medium-density Q-matrices.

Table 1. Full Results for the Simulation Study

Condition	Model	Selection	Density	ρ	ATL	PCC
1	DINA	MPWKL	Low	0.3	9.55	0.819
2	DINA	MPWKL	Low	0.7	9.30	0.802
3	DINA	MPWKL	Medium	0.3	9.94	0.838
4	DINA	MPWKL	Medium	0.7	10.05	0.835
5	DINA	MPWKL	High	0.3	13.71	0.774
6	DINA	MPWKL	High	0.7	15.34	0.719
7	DINA	MUINF	Low	0.3	9.49	0.823
8	DINA	MUINF	Low	0.7	9.35	0.806
9	DINA	MUINF	Medium	0.3	9.93	0.834
10	DINA	MUINF	Medium	0.7	10.02	0.831
11	DINA	MUINF	High	0.3	13.97	0.761
12	DINA	MUINF	High	0.7	15.08	0.717
13	RRUM	MPWKL	Low	0.3	9.68	0.830
14	RRUM	MPWKL	Low	0.7	9.60	0.820
15	RRUM	MPWKL	Medium	0.3	10.51	0.837
16	RRUM	MPWKL	Medium	0.7	10.61	0.821
17	RRUM	MPWKL	High	0.3	13.89	0.800
18	RRUM	MPWKL	High	0.7	15.15	0.773
19	RRUM	MUINF	Low	0.3	9.64	0.825
20	RRUM	MUINF	Low	0.7	9.51	0.809
21	RRUM	MUINF	Medium	0.3	10.43	0.828
22	RRUM	MUINF	Medium	0.7	10.40	0.822
23	RRUM	MUINF	High	0.3	14.04	0.795
24	RRUM	MUINF	High	0.7	15.16	0.771

Figure 2 illustrates the average number of skills tapped per item at each stage of the test for each condition. At the beginning of the tests, the average number of skills tapped per item increased for all conditions and peaked at some point between the 5th and 10th item administered for most conditions. Around Stage 10, the lines separated into two groups—conditions with low- or medium-density Q-matrices and those with high-density matrices. The average number of skills tapped per item generally decreased throughout the later test stages for the low- and medium-density Q-matrices. The pattern was a bit more complex for the high-density Q-matrices. For those conditions, there was an initial dip in average skills tapped per item; then there was a steady increase throughout the later stages of the tests.

Figure 1. PCC vs. ATL for DINA and RRUM Using MUINF



a. DINA

b. RRUM

Note. Numbered points correspond to condition numbers in Table 1. Conditions with skill correlation $\rho = 0.3$ and 0.7 appear as red and blue, respectively.

Discussion and Conclusions

To understand the increase in average test length and decrease in classification accuracy incurred for denser Q-matrices, it is instructive to explore the effect of the Q-matrix on item information using a simple numerical example. Consider a situation in which $K = 3$ skills are being assessed in which there are $2^3 = 8$ possible skill patterns:$\{0,0,0\},\{1,0,0\},\ldots,\{1,1,1\}$. Furthermore, consider seven items with Q-matrix rows corresponding to the skill patterns, except $\{0,0,0\}$. To remove the influence of differing item parameters, assume all seven items have the same parameters. Specifically, in the case of the RRUM, $r_{jk}^* = 0.10$ and $\pi_j^* = 0.90$ for all items j and skills k; and for the DINA, $g_j = s_j = 0.10$.

Table 2. Proportion of Examinees with One, Two, Three, and Four or More Skills Misclassified, by Condition

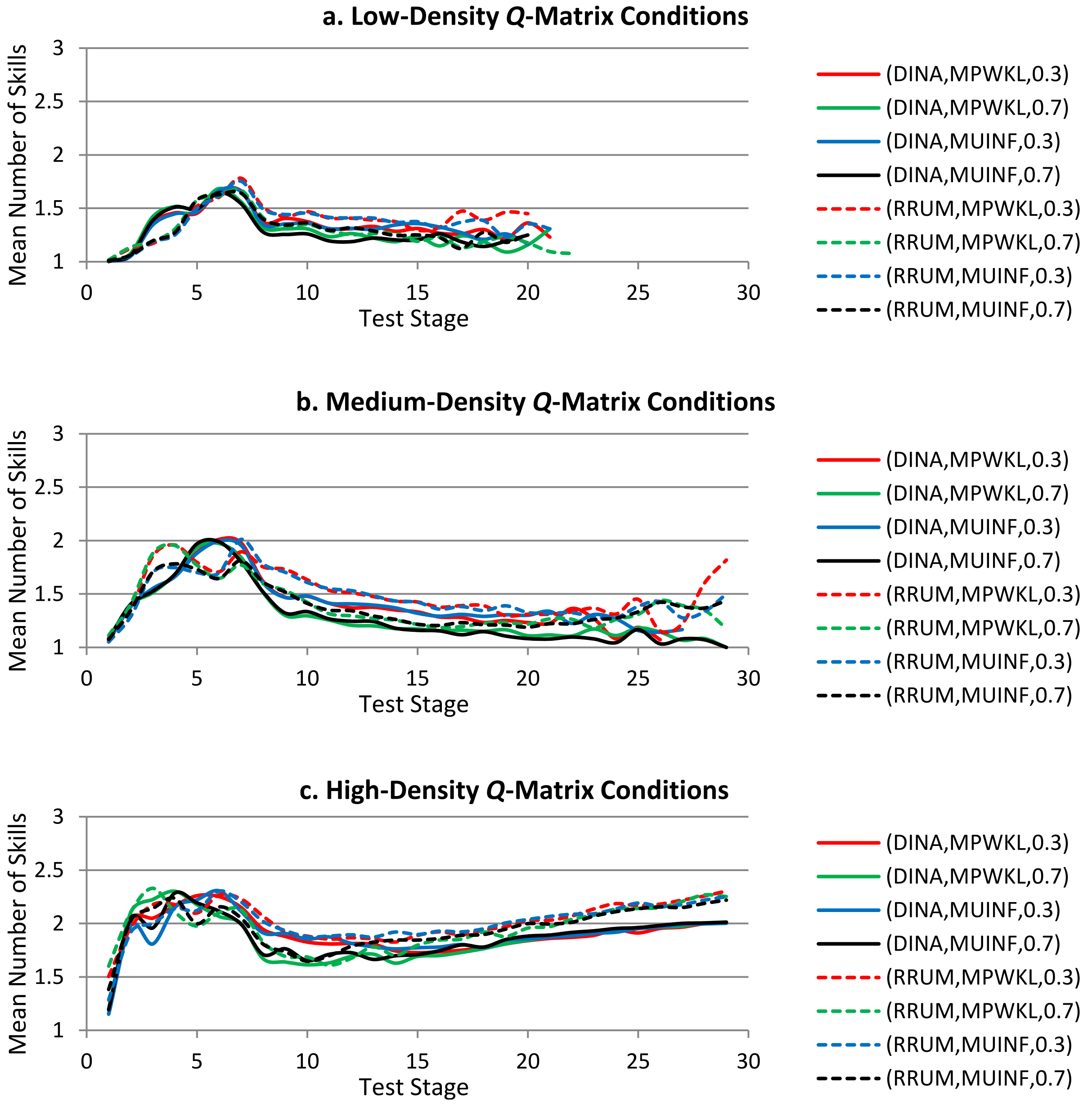
					Proportion with Number of Misclassified Skills			
Condition	Model	Selection	Density	ρ	1	2	3	≥ 4
1	DINA	MPWKL	Low	0.3	0.1651	0.0146	0.0014	0.0001
2	DINA	MPWKL	Low	0.7	0.1802	0.0166	0.0008	0.0002
3	DINA	MPWKL	Medium	0.3	0.1441	0.0162	0.0017	0.0003
4	DINA	MPWKL	Medium	0.7	0.1465	0.0167	0.0017	0.0002
5	DINA	MPWKL	High	0.3	0.1785	0.0361	0.0095	0.0023
6	DINA	MPWKL	High	0.7	0.2222	0.0500	0.0072	0.0013
7	DINA	MUINF	Low	0.3	0.1601	0.0157	0.0012	0.0001
8	DINA	MUINF	Low	0.7	0.1734	0.0196	0.0012	0.0000
9	DINA	MUINF	Medium	0.3	0.1488	0.0149	0.0019	0.0003
10	DINA	MUINF	Medium	0.7	0.1521	0.0152	0.0017	0.0004
11	DINA	MUINF	High	0.3	0.1837	0.0428	0.0112	0.0018
12	DINA	MUINF	High	0.7	0.2236	0.0513	0.0063	0.0019
13	RRUM	MPWKL	Low	0.3	0.1558	0.0126	0.0011	0.0001
14	RRUM	MPWKL	Low	0.7	0.1670	0.0124	0.0006	0.0000
15	RRUM	MPWKL	Medium	0.3	0.1457	0.0161	0.0013	0.0000
16	RRUM	MPWKL	Medium	0.7	0.1622	0.0159	0.0010	0.0001
17	RRUM	MPWKL	High	0.3	0.1647	0.0311	0.0038	0.0005
18	RRUM	MPWKL	High	0.7	0.2024	0.0223	0.0021	0.0001
19	RRUM	MUINF	Low	0.3	0.1651	0.0146	0.0014	0.0001
20	RRUM	MUINF	Low	0.7	0.1802	0.0166	0.0008	0.0002
21	RRUM	MUINF	Medium	0.3	0.1441	0.0162	0.0017	0.0003
22	RRUM	MUINF	Medium	0.7	0.1465	0.0167	0.0017	0.0002
23	RRUM	MUINF	High	0.3	0.1785	0.0361	0.0095	0.0023
24	RRUM	MUINF	High	0.7	0.2222	0.0500	0.0072	0.0013

Also, assume that the prior probability of each skill pattern was set to $1/8$ for each pattern. Thus, the example is designed so that the only difference between the items is the number of skills they tap. Table 3 shows the amount of information (both MUINF and MPWKL) in each item for each skill pattern. Clearly, under both information criteria, the items tapping fewer skills have more information.

Of course, these calculations are merely illustrative. A formal proof that item information decreases for any CDM and all K values as more skills are tapped is beyond the scope of this paper. However, the results of the simulation with $K = 6$ are consistent with Table 3, i.e., the longer ATLs and smaller PCCs incurred under conditions with dense Q-matrices were due to those items being less informative than those in the low- or medium-density conditions. Some intuition for the mutual information (MI) criterion can be gained from adapting the explanation of Weissman (2007) to the context of CDMs. The symmetric nature of MI can allow it to be interpreted both as

(1) the reduction in uncertainty in $\boldsymbol{\alpha_i}$ if item j were administered and (2) the reduction in uncertainty of predicting item response X_{ij} if item j were administered, given current knowledge of $\boldsymbol{\alpha_i}$. The latter interpretation seems to be helpful for the present case. It is reasonable to consider that as more skills are tapped by an item, the response to that item becomes more uncertain, as the successful application of each individual skill is a random event itself in the CDM framework.

Figure 2. Mean Number of Skills per Item by Test Stage and Testing Condition



a. Low-Density *Q*-Matrix Conditions

b. Medium-Density *Q*-Matrix Conditions

c. High-Density *Q*-Matrix Conditions

Note. Each line in each plot corresponds to a different testing condition, broken down by Q-matrix density. The *x*-axis of the plot begins at Stage 2 (second item administered), since the first item is drawn randomly. The *y*-axis represents the mean number of skills tapped per item at each test stage.

Table 3. MUINF and MPWKL under the DINA and RRUM by Q-Matrix Row for $K = 3$ Example

	DINA		RRUM	
Row	MUINF	MPWKL	MUINF	MPWKL
{1,0,0}	0.368	0.879	0.379	0.913
{0,1,0}	0.368	0.879	0.379	0.913
{0,0,1}	0.368	0.879	0.379	0.913
{1,1,0}	0.286	0.659	0.340	0.865
{0,1,1}	0.286	0.659	0.340	0.865
{1,0,1}	0.286	0.659	0.340	0.865
{1,1,1}	0.175	0.385	0.248	0.636

It is worthwhile to emphasize that the calculations displayed in Table 3 were performed assuming each skill pattern is equally likely. During the actual CD-CAT, however, the posterior distribution of each pattern was updated after each item was administered; the probability of each pattern was not equal, of course. In actual testing, sometimes an item tapping more skills might be more informative than one tapping fewer skills, depending on the posterior distribution of the skill patterns as well as the item parameters.

Although it is difficult to make direct comparisons due to differences in study designs, it is interesting to note that the patterns observed in Figure 2 for the number of skills tapped with the test stage follow similar trends as those reported by Kaplan et al. (2015). In particular, when employing the DINA model along with the GDI, Kaplan et al. (2015) show in their Figure 4 that for a fixed K = 5 true attribute vector {11100}, items measuring only a single skill were more frequently selected at the earlier stages of the CD-CAT (i.e., between one and five items administered), after which the mean number of skills increased to approximately 1.65 for the remaining stages (i.e., between 6 and 20 items administered).

This general trend can also be observed in Figure 2, where items measuring fewer skills were also selected more frequently at the earlier stages of the CD-CAT, keeping in mind that Figure 2 summarizes results across all examinees' attribute patterns, whereas Kaplan et al.'s (2015) Figure 4 is for a single attribute pattern.

Conclusions

This study systematically varied a number of factors in a variable-length CD-CAT in order to examine their effects on average test length and classification accuracy, specifically, the generating model, the denseness of the Q-matrix, the item selection method, and the correlation between skills. It is hoped that the results will further the understanding of the fundamentals of CD-CAT as well as point out issues to those working to implement CD-CAT in an actual classroom setting. In particular, although CDMs have been touted for their multidimensional nature (Rupp & Templin, 2008; Rupp et al., 2010), there have previously been little to no recommendations as to how Q-matrices should be constructed to ensure efficient CD-CATs.

This appears to be the first study to vary the density of the Q-matrix for a variable-length CD-CAT. The results suggest that low- to medium-density Q-matrices result in higher PCCs and shorter average ATLs, as compared to high-density Q-matrices. This seems to imply that, in practice, items should be written to tap as few skills as possible. Depending on the subject matter, however, it might not always be possible to write items that tap only one skill at a time. For example, items on an assessment covering a complex topic, such as advanced mathematics, might each tap at least several skills. Also, if the number of skills K is very large, it might be impractical to assess one skill at a time.

No single simulation study can examine effectively and coherently all possible factors affecting the classification accuracy and average test length of a CD-CAT. In future studies, researchers might wish to vary design elements such as K, generating CDM, item bank size, and test length. Although this study has looked at only a relatively small number of testing conditions, it has broached this practical subject for future studies to consider, as well. Finally, this study emphasizes the importance of the Q-matrix to CDM methodology in general. A clearer understanding of Q-matrix construction should help in designing and maintaining operational CD-CATs.

References

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619-632. CrossRef

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. CrossRef

Chiu, C. Y. (2013). Statistical refinement of the Q-Matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618. CrossRef

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353. CrossRef

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343–362. CrossRef

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115-130. CrossRef

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199. CrossRef

DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R Rao & S. Sinharay (Eds.), *Handbook of Statistics*, *26*, (pp. 979-1030). Amsterdam, The Netherlands: Elsevier B.V.

Feng, Y., Habing, B., & Huebner, A. (2014). Parameter estimation of the reduced RUM using the EM algorithm. *Applied Psychological Measurement*, *38*(2), 137–150. CrossRef

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

Henson, R.A., Templin J.L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. CrossRef

Hsu, C., Wang, W., & Chen, S. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, *37*(7), 563-582. CrossRef

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*(2), 407–419. CrossRef

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

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272. CrossRef

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. CrossRef

Li, H. & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading rest. *Educational Assessment*, *18*(1), 1–25. CrossRef

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219-262. CrossRef

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

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

Wang, C., Chang, H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *48*(3), 255-273. CrossRef

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*(1), 41-58. CrossRef

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

Author's Address

Alan Huebner, University of Notre Dame, 153 Hurley Hall, Notre Dame, IN 46556.
Email: Alan.Huebner.10@nd.edu