ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 148-165

Investigating The Performance of Item Selection Algorithms in Cognitive Diagnosis Computerized Adaptive Testing

Semih AŞİRET * Seçil ÖMÜR SÜNBÜL**

Abstract

This study aimed to examine the performances of item selection algorithms in terms of measurement accuracy and computational time, using factors such as test length, number of attributes, and item quality in fixed-length CD-CAT and average test lengths and computational time, using factors such as number of attributes and item quality in variable-length CD-CAT. In the research, two different simulation studies were conducted for the fixed and variable-length tests. Item responses were generated according to the DINA model. Two item banks, which consisted of 480 items for 5 and 6 attributes, were generated, and the item banks were used for both the fixed and variable-length tests. Q-matrix was generated item by item and attribute by attribute. In the study, 3000 examinees were generated in such a way that each examinee had a 50% chance of achieving each attribute. The cognitive patterns of the examinees were estimated by using MAP. In the variable-length CD-CAT, the first-highest posterior probability threshold is 0.80, and the second-highest posterior probability threshold is 0.10. The CD-CAT administration and other analyses were conducted using R 3.6.1.At the end of the study in which the fixed-length CD-CAT was used, it was concluded that an increase in the number of attributes resulted in a decrease in the pattern recovery rates of item selection algorithms. Conversely, these rates improved with higher item quality and longer test lengths. The highest values in terms of pattern recovery rate were obtained from JSD and MPWKL algorithms. In the variable-length CD-CAT, it was concluded that the average test length increased with the number of attributes and decreased with higher item quality. Across all conditions, the JSD algorithm yielded the shortest average test length. Additionally, It has been determined that GDI algorithm had the shortest computation time in all scenarios, whereas the MPWKL algorithm exhibited the longest computation time.

Keywords: computerized adaptive testing, cognitive diagnosis models, item selection algorithms

Introduction

Monitoring students' learning situation and understanding their progress has a critical importance in educational sciences. Assessment is not only limited to measuring students' existing knowledge and skills; it also plays a vital role in guiding their learning processes and increasing their motivation. In this context, assessment should be recognized as an integral part of the educational processes. Stiggins (2002) also supports this perspective and emphasizes that assessment should not only reveal the current state of learning but also be used to improve learning. Assessment should present interpretative, diagnostic, highly informative, and predictive information (Pellegrino et al., 1999). However, in many studies (Bennett, 2011; Black & William, 2018; Heritage, 2010; William, 2011), it is reported that only the learning situation is supervised and the information that will facilitate the learning of examinees is not provided.

* PhD., Mersin University, Faculty of Education, Mersin-Turkey, semihasiret@gmail.com, ORCID ID: 0000-0002-0577-2603
** Assoc. Prof., Mersin University, Faculty of Education, Mersin-Turkey, secilomur@gmail.com, ORCID ID: 0000-0001-9442-1516

To cite this article:
Aşiret, S., & Ömür-Sübül, S. (2024). Investigating the performance of item selection algorithms in cognitive diagnosis computerized adaptive testing. *Journal of Measurement and Evaluation in Education and Psychology, 15*(2), 148-165. https://doi.org/10.21031/epod.1456094

Received: 25.03.2024
Accepted: 21.06.2024

Assessments based on Cognitive Diagnostic Models (CDM), which evaluate whether examinees have certain attributes, aim to provide descriptive feedback for each examinee rather than giving examinees subscale scores or summative scores. Therefore, diagnostic assessments provide detailed and useful information about each examinee's learning strengths and weaknesses and assess student achievement. This makes assessments based on cognitive diagnosis powerful and interesting, especially in areas where formative assessment is aimed and classroom assessments will be used.

CDMs are discrete latent variable models that enable the diagnosis of the operations required to solve a problem in a test or the presence or absence of many minor skills (de la Torre, 2009). Diagnoses obtained as a result of analysis with CDM can provide more details for a particular area and allow interventions that will bring solutions. With this model, a cognitive pattern can be produced for each examinee or group about whether the necessary skills or process steps are sufficient for a situation (Rupp & Templin, 2008).

CDMs are discrete latent variable models designed to diagnose the specific operations required to solve a problem in a test or to determine the presence or absence of various minor skills (de la Torre, 2009). Analyses conducted using CDMs can yield detailed diagnostic information for a particular domain and facilitate targeted interventions to address identified issues. This model allows for the generation of a cognitive profile for each examinee or group, indicating whether the necessary skills or process steps are sufficient for a given situation (Rupp & Templin, 2008).

The rise of computer technologies and their increasing accessibility for examinees has paved the way for the emergence of computerized adaptive testing (CAT) as a significant and popular research and application topic within psychometrics and education (Magis et al., 2017). Many CAT implementations position examinees along a latent continuum, but the need for individualized diagnostic feedback to examinees remains a challenge in this approach. Cognitive Diagnosis Computerized Adaptive Tests (CD-CAT) integrate Cognitive Diagnostic Models (CDM) with CAT methodologies. CD-CAT aims to classify examinees according to their latent status and apply latent class models to these latent classes (Cheng, 2009). More broadly, the primary aim of CD-CAT is to deliver individualized diagnostic feedback to examinees. Similar to CAT applications, CD-CAT encompasses the creation of an item bank, selecting the initial item to begin the test, estimating cognitive pattern, selecting subsequent items, terminating rules, estimating the final cognitive pattern, and reporting. However, item selection algorithms used in CAT are not suitable for CD-CAT. This is because CDMs operate with discrete latent variables, and algorithms such as Maximum Fisher Information (MFI) fail to make accurate predictions when the number of items is low. They are susceptible to the effects of chance success.

In recent years, numerous theories and algorithms related to CD-CAT applications have been developed (Cheng, 2009; Kaplan et al., 2015; McGlohen & Chang, 2008; Wang, 2013; Tatsuoka, 2002; Tatsuoka & Ferguson, 2003; Xu et al., 2003; Zheng & Chang, 2016). The item selection algorithms in the CD-CAT studies are primarily based on the Shannon Entropy (SHE) algorithm developed by Tatsuoka (2002) and Tatsuoka and Ferguson (2003), as well as the Kullback-Leibler (KL) information developed by Xu et al. (2003). However, Cheng (2009) utilized the Post-Weighted Kullback-Leibler information (PWKL) and Hybrid Kullback-Leibler information (HKL), while Wang (2013) employed Mutual information (MI) and Kaplan et al. (2015) used Modified Post-Weighted Kullback-Leibler information (MPWKL) and GDINA discrimination index (GDI). Additionally, Zheng and Chang (2016) developed the Post-Weighted Cognitive Discrimination Index (PWCDI) and the Post-Weighted Attribute-level Cognitive Discrimination Index (PWACDI), and Minchen and de la Torre (2016) introduced the Jensen-Shannon divergence (JSD) index.

The critical aspect of CD-CAT is the item selection algorithms (Cheng, 2009; Zheng, 2015; Zheng & Chang, 2016). Various item selection algorithms have been developed in CAT applications to cater to different needs. These algorithms are firmly established in CAT studies based on IRT. However, limited studies discuss item selection algorithms in the context of CD-CAT, as it is a relatively new field. Numerous factors, such as the number of attributes, structure of the Q matrix, item quality, termination rule, and estimation method, can influence the accuracy of results in these studies. Zheng (2015)

emphasizes that the primary goal of item selection algorithms is to achieve high measurement accuracy, which is also true for CD-CAT.

Many item selection algorithms have been developed for CD-CAT applications in recent years. However, most of these algorithms have not been evaluated under the same conditions. This study aims to examine various item selection algorithms in CD-CAT and compare them based on test length, number of attributes, item quality, and termination rule. By manipulating these factors at different levels in CD-CAT, the goal is to determine the item selection algorithms that provide the most accurate and maximum pattern recovery rates, computation time, and average test length.

Method

Factors that Manipulated in the Research

***Number of attributes*:** The number of attributes is one of the important factors affecting the accuracy of estimations in CD-CAT. Rupp and Templin (2008) stated that the number of attributes between 4 and 6 is moderate. In this study, since the number of attributes (K) was aimed at a medium level, K was manipulated to 5 and 6.

Test length: There are two ways to handle test length: fixed-length and variable-length tests. DiBello and Stout (2007), as well as Wang (2013), argue that tests should be short to avoid wasting class time, especially since CD-CAT is mostly used in low-stakes tests and classroom assessments. In addition, tests in classroom assessments should be answered during the course time after each item is administered. Therefore, test lengths were manipulated to 5, 10, 15, and 20 in this study for the fixed-length tests.

Item Selection Algorithms: Item selection algorithms play a crucial role in CD-CAT studies, according to Cheng (2009). Various item selection algorithms have been developed for CD-CAT studies. The fixed-length test for CD-CAT used item selection algorithms including KL (Xu et al., 2003), SHE (Tatsuoka, 2002), PWKL, and HKL (Cheng, 2009), MI (Wang, 2013), GDI, and MPWKL (Kaplan et al., 2015), PWCDI, PWACDI (Zheng & Chang, 2016), and JSD (Minchen & de la Torre, 2016). Meanwhile, the CD-CAT based on variable-length test used PWKL and HKL (Cheng, 2009), MI (Wang, 2013), GDI, and MPWKL (Kaplan et al., 2015), PWCDI, PWACDI (Zheng & Chang, 2016), and JSD (Minchen & de la Torre, 2016) item selection algorithms. Additionally, random selection was used as the base algorithm for all conditions to facilitate comparisons of other algorithms' performances.

Item quality: The quality of the items was determined according to the discrimination index. In this study, item parameter distributions by Kaplan et al. (2015) were used. Therefore, the item parameters were generated from a uniform distribution. These item quality parameters are given in Table 1.

Table 1.

Item parameters

Item quality	$\left[\mathbf{1}-\boldsymbol{P_j}(\mathbf{1})\right]$ & $\boldsymbol{P_j}(\mathbf{0})$
LD-LV	U (0.15, 0.25)
LD-HV	U (0.10, 0.30)
HD-LV	U (0.05, 0.15)
HD-HV	U (0.00, 0.20)

Note: LD= low discrimination, HD=high discrimination, LV= low variance, HV= high variance

Termination Rule: CD-CAT studies use fixed and variable test lengths as termination rules. In this study, a two-criterion termination rule, suggested by Hsu et al. (2013), was used for variable-length CD-CAT. The first highest posterior probability threshold value was set at 0.80, and the second highest posterior probability threshold was set at 0.10. As the number of attributes increased, the number of cognitive patterns required would also increase exponentially to ensure that all items in the item bank were used. For this reason, the maximum test length was set to 40.

Data Generation and Analysis

Within the scope of this study, R. 3.6.1 (R Core Team, 2020) carried out manipulating factors, data generation, and data analysis according to the levels of these factors.

Generating the item bank and examinees: Creating the item bank involves generating the Q matrix and the parameters of the DINA model. In this study, the longest test includes 20 items in the fixed-length CD-CAT. The maximum test length was set to 40 items in the study using the variable test length termination rule. Stocking (1994) suggested that the item bank should be at least 12 times the test length (Cheng, 2009). Therefore, two separate item banks with a total of 480 items, consisting of 5 and 6 attributes, which were used for both fixed and variable-length CD-CAT were created. The Q matrix was developed item-by-item and attribute-by-attribute. To ensure equal representation of each attribute in the item bank and to make it applicable to real-world scenarios, data were generated so that each item had a 30% chance of measuring each attribute, and each item was required to measure at least one attribute. The data were generated so that there was no correlation between the attributes. The Q matrix contains $2^K - 1$ cognitive patterns. 3000 examinees were generated, each with a 50% chance of mastering each attribute, and common examinees were used for both studies. Based on the estimated item parameters and the Q matrix, the item responses of 3000 examinees and the probability of each examinee answering each item correctly according to the DINA model were computed.

Table 2 shows the distribution of the number of items measuring each attribute and the number of examinees with each attribute according to the number of manipulated attributes.

Table 2.

Number of Items Measuring Each Attribute and the Number of Examinees with Each Attribute

	Attributes					
K=5	1	2	3	4	5	
Number of items (J=480)	174	184	170	175	169	
Number of examinees	1484	1462	1494	1454	1517	
K=6	1	2	3	4	5	6
Number of items (J=480)	171	167	160	164	161	162
Number of examinees	1494	1476	1535	1500	1495	1510

In Table 3, the number of items measuring the possible number of attributes for 5 and 6 attributes in the item bank consisting of 480 items and the number of examinees with each attribute are given. In producing the Q matrix, each item was created to measure 30% of the attributes on average to be close to the real situation.

Table 3.

The Number of Items Measuring the Possible Number of Attributes and the Number of the Examinees with the Attribute

Number of Attributes (K=5)	0	1	2	3	4	5	
Number of Items (J=480)	0	192	199	69	16	1	
Number of Examinees	108	476	948	923	455	90	
Number of Attributes (K=6)	0	1	2	3	4	5	6
Number of Items (J=480)	0	162	176	102	32	7	1
Number of Examinees	45	282	707	947	685	271	63

Analysis Model: The DINA model is frequently preferred in simulation studies based on CDM and in low-stake tests due to the ease of parameter estimation and interpretation (Cheng, 2009; de la Torre, 2011; DeCarlo, 2011). Therefore, in this study, the DINA model was used.

First item selection: CD-CAT starts with the first item selection. Within the scope of this study, the first item selection was made randomly and kept constant in other algorithms.

Estimating the cognitive pattern: The Maximum Likelihood Estimation (MLE) method cannot estimate when examinees answer all items correctly or incorrectly (de Ayala, 2010). A similar situation applies to CDM studies. Test lengths can be short (e.g., five items), as CD-CAT studies are frequently conducted for classroom assessment. In such short-length tests, examinees' item response patterns are highly likely to be either all 0s or all 1s. Therefore, in this study, the cognitive patterns of examinees were estimated using the Maximum a posteriori (MAP) estimation method.

Evaluation criteria: Within the scope of this study, the Pattern Recovery Rates (PRR) and computation time were used to evaluate the item selection algorithms for the fixed-length CD-CAT. For the variable test-length CD-CAT, PRR, computation time, and average test length were used to evaluate the performance of item selection algorithms.

For fixed-length CD-CAT, PRR is the rate of all correctly defined attribute patterns (Zheng & Chang, 2016). It refers to the proportion of examinees within the sample whose estimated cognitive pattern, $\hat{\alpha}_i$, is identical to their true cognitive pattern, α_i, across all attributes. The higher PRR indicates greater classification accuracy. PRR is calculated by Equation 1.

$$\mathrm{PRR_k} = \frac{\sum_{i=1}^{N} R_i}{N} = \frac{\sum_{i=1}^{N}\left(I_{\hat{\alpha}_i,\alpha_i}\right)}{N}, \qquad (k=1, 2, \ldots., K) \qquad (1)$$

The computation times for the item selection algorithms were measured in seconds from the start of the process to estimate the first examinee's cognitive pattern until all examinees' cognitive patterns were estimated. The "*tictoc*" package (Izrailev, 2021) was used for this purpose. After calculating the time taken by all examinees, the total time was divided by the total number of examinees (in seconds) to get the average computation time for each examinee. This value was multiplied by 1000 for easier interpretation and reported as milliseconds per examinee. To calculate the relative average computation time of the item selection algorithms, it was divided by the computation time of the algorithm with the lowest average computation time by the computation time of the other algorithms.

For variable-length CD-CAT, the posterior probability of the cognitive pattern was used as the termination criterion instead of the fixed test length. After each selected item was administered to each examinee, the posterior probabilities of the cognitive patterns were estimated. In addition to the criterion that the highest posterior probability value is greater than 0.80 and the second highest posterior probability is less than 0.10 when the maximum number of items administered is 40, the test was terminated even if the posterior probability estimated for the examinee could not exceed 0.80. Therefore, these examinees were retained as examinees who did not complete the test. The estimated cognitive patterns of the examinees and the items used were recorded in the loop. After the loop was completed, minimum, maximum, and average statistics of the number of items used for each examinee were recorded for each item selection algorithm. In addition, the total number of examinees who could not complete the test was calculated. After these processes, the item selection algorithms' attribute and pattern recovery rates and average computation times were calculated. Finally, tables and graphs were produced using R 3.6.1. The "*ggplot2*" package (Wickham, 2016) was used to produce and edit the graphics.

Findings

Fixed-length CD-CAT

Pattern recovery rates of item selection algorithms: The results of the pattern recovery rates of the item selection algorithms for fixed-length CD-CAT across various test lengths, item qualities, and number of attributes are presented in Table 4. These results are also graphically represented in Figure 1.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 148-165

Table 4.
Pattern Recovery Rates of Item Selection Algorithms in Fixed-Length CD-CAT

			Item Selection Algorithms										
K	TL	IQ	Random	GDI	HKL	JSD	KL	MPWKL	MI	PWACDI	PWCDI	PWKL	SHE
5	5	LD-LV	0,142	0,323	0,274	0,403	0,179	0,402	0,328	0,343	0,370	0,273	0,282
		LD-HV	0,158	0,385	0,318	0,533	0,183	0,533	0,385	0,407	0,430	0,319	0,374
		HD-LV	0,192	0,544	0,423	0,730	0,263	0,730	0,549	0,595	0,672	0,423	0,544
		HD-HV	0,270	0,662	0,458	0,916	0,321	0,860	0,661	0,656	0,721	0,450	0,655
	10	LD-LV	0,242	0,621	0,599	0,641	0,300	0,640	0,619	0,595	0,612	0,584	0,608
		LD-HV	0,322	0,727	0,713	0,759	0,307	0,758	0,733	0,690	0,724	0,711	0,729
		HD-LV	0,405	0,902	0,881	0,920	0,480	0,921	0,904	0,888	0,906	0,881	0,908
		HD-HV	0,500	0,986	0,956	0,990	0,688	0,991	0,985	0,974	0,983	0,959	0,988
	15	LD-LV	0,344	0,798	0,772	0,820	0,419	0,812	0,802	0,776	0,798	0,772	0,805
		LD-HV	0,429	0,895	0,874	0,911	0,493	0,899	0,900	0,859	0,888	0,875	0,889
		HD-LV	0,515	0,980	0,975	0,985	0,645	0,984	0,984	0,970	0,983	0,975	0,982
		HD-HV	0,665	0,999	0,998	0,999	0,793	0,999	0,998	0,998	1,000	0,997	0,999
	20	LD-LV	0,462	0,899	0,881	0,912	0,511	0,909	0,906	0,867	0,897	0,887	0,895
		LD-HV	0,529	0,957	0,949	0,956	0,627	0,961	0,953	0,938	0,954	0,948	0,959
		HD-LV	0,568	0,996	0,997	0,998	0,756	0,997	0,998	0,993	0,996	0,997	0,997
		HD-HV	0,775	1,00	1,00	1,00	0,885	1,00	1,00	0,999	1,00	1,000	1,00
6	5	LD-LV	0,088	0,189	0,164	0,204	0,108	0,206	0,189	0,201	0,184	0,161	0,188
		LD-HV	0,086	0,222	0,210	0,261	0,116	0,254	0,222	0,225	0,235	0,199	0,210
		HD-LV	0,131	0,341	0,237	0,360	0,140	0,362	0,339	0,321	0,353	0,237	0,34
		HD-HV	0,134	0,398	0,259	0,435	0,269	0,426	0,413	0,376	0,424	0,391	0,388
	10	LD-LV	0,160	0,482	0,463	0,493	0,180	0,499	0,483	0,457	0,483	0,464	0,475
		LD-HV	0,183	0,610	0,583	0,631	0,212	0,628	0,604	0,565	0,605	0,579	0,593
		HD-LV	0,233	0,843	0,79	0,847	0,321	0,851	0,841	0,777	0,821	0,799	0,839
		HD-HV	0,335	0,956	0,907	0,969	0,413	0,963	0,965	0,916	0,933	0,938	0,964
	15	LD-LV	0,255	0,687	0,656	0,688	0,276	0,691	0,686	0,645	0,669	0,658	0,674
		LD-HV	0,292	0,806	0,778	0,821	0,346	0,807	0,802	0,744	0,788	0,777	0,786
		HD-LV	0,389	0,959	0,942	0,960	0,478	0,962	0,955	0,923	0,951	0,943	0,962
		HD-HV	0,495	0,997	0,992	0,996	0,673	0,997	0,997	0,992	0,994	0,994	0,996
	20	LD-LV	0,335	0,817	0,795	0,825	0,376	0,829	0,817	0,765	0,809	0,802	0,811
		LD-HV	0,392	0,902	0,888	0,910	0,458	0,914	0,901	0,859	0,893	0,893	0,894
		HD-LV	0,496	0,989	0,988	0,989	0,617	0,990	0,987	0,98	0,988	0,987	0,991
		HD-HV	0,638	1,00	0,998	1,00	0,771	1,00	0,999	0,999	1,00	0,999	0,998

* PhD., Mersin University, Faculty of Education, Mersin-Turkey, semihasiret@gmail.com, ORCID ID: 0000-0002-0577-2603
** Assoc. Prof., Mersin University, Faculty of Education, Mersin-Turkey, secilomur@gmail.com, ORCID ID: 0000-0001-9442-1516

To cite this article:
Aşiret, S., Ömür Sübül, S. (2024). Investigating the performance of item selection algorithms in cognitive diagnosis computerized adaptive testing. Journal of Measurement and Evaluation in Education and Psychology, *15*(2), 148-165. https://doi.org/10.21031/epod.1456094

Received: 25.03.2024
Accepted: 21.06.2024

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(2); 148-165

Figure 1.

PRR of Item Selection Algorithms in Fixed-Length CD-CAT

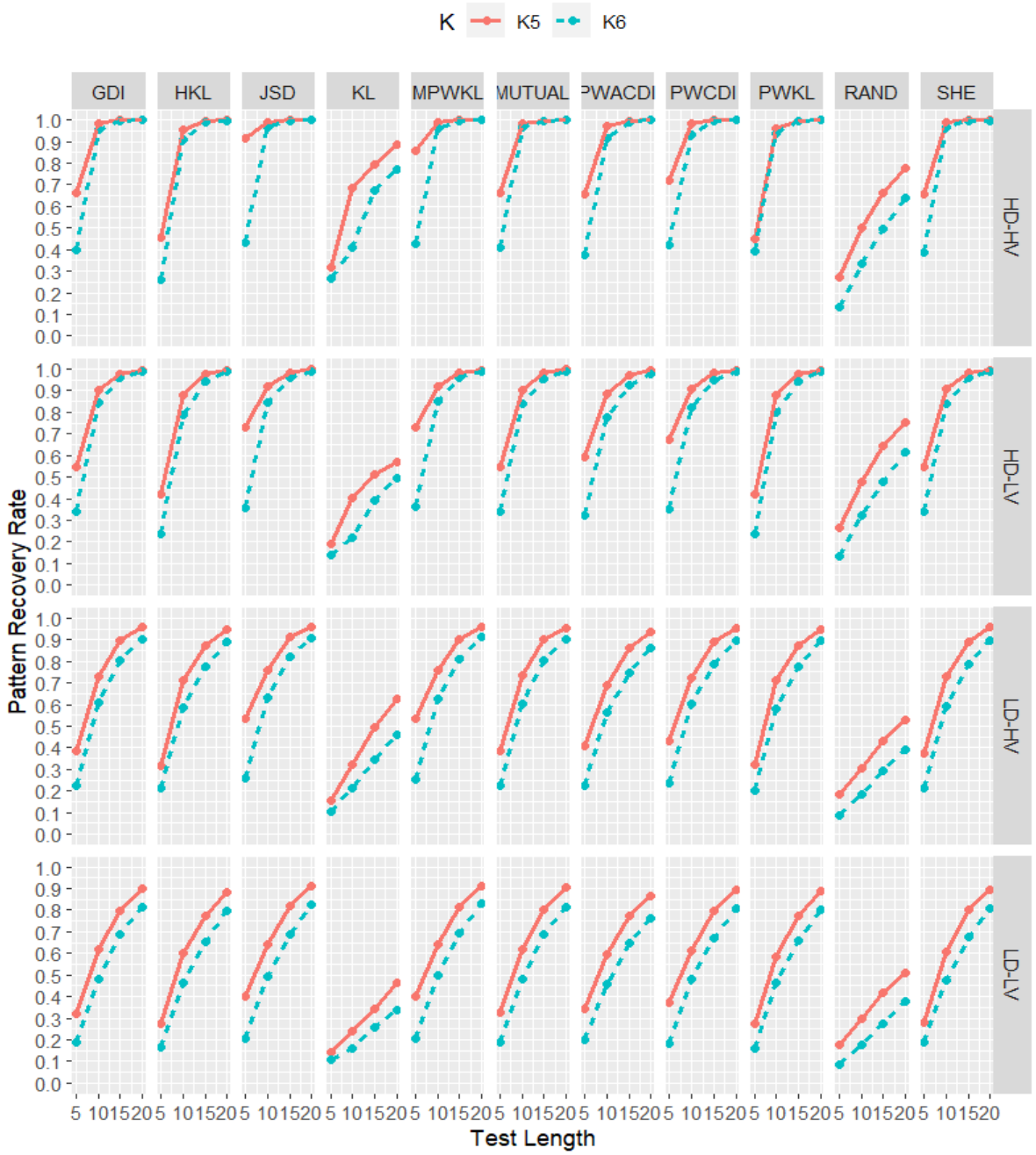


Table 4 and Figure 1 show that PRR for item selection algorithms increases significantly with increasing test length and item discrimination-item variance. Meanwhile, they decrease with an increasing number of attributes. Analysis of Figure 1 indicates that the increase in PRR is most pronounced when the test length is increased from 5 to 10, compared to other test lengths. At test lengths of 15 and 20, the rates for high item quality (HD-LV and HD-HV) are very close between 5 and 6 attributes, whereas, for low

* PhD., Mersin University, Faculty of Education, Mersin-Turkey, semihasiret@gmail.com, ORCID ID: 0000-0002-0577-2603
** Assoc. Prof., Mersin University, Faculty of Education, Mersin-Turkey, secilomur@gmail.com, ORCID ID: 0000-0001-9442-1516

To cite this article:
Aşiret, S., Ömür Sübül, S. (2024). Investigating the performance of item selection algorithms in cognitive diagnosis computerized adaptive testing. Journal of Measurement and Evaluation in Education and Psychology, *15*(2), 148-165. https://doi.org/10.21031/epod.1456094

Received: 25.03.2024
Accepted: 21.06.2024

item quality (LD-LV and LD-HV), the rates for 5 attributes are higher than those for 6 attributes. Additionally, at 20 test lengths with high item discrimination (HD-HV and HD-LV), the PRR of item selection algorithms is very close to 1 (0.99-1.00).

The lowest PRR was obtained by random selection, followed by the KL algorithm. Figure 1 further supports that these two algorithms performed worse than others. The highest PRR among item selection algorithms, under conditions of low item quality, 6 attributes, and a test length of 5, is 0.206 (MPWKL). For test length 5, the highest PRR was obtained with the JSD and MPWKL algorithms, with minimal differences. Generally, the highest PRR was achieved with the JSD and MPWKL across most conditions. Specifically, these algorithms outperformed others in short tests (5) with 5 attributes. For the HD-HV item quality level, the JSD algorithm's PRR is 0.916 for 5 test lengths and 5 attributes. The PWCDI algorithm follows JSD and MPWKL in the PRR for 5 test lengths and 5 attributes. However, the performance of PWCDI decreased with increasing test length, except for the HD-HV item quality.

The PWACDI algorithm consistently had a lower PRR after KL and random selection, except for test length 5. Similar results were observed for the GDI, SHE, and MI algorithms. In short tests with 5 attributes, MI performed better than SHE, whereas both gave similar results in longer tests. For 6 attributes, MI and GDI outperformed SHE. The PWKL and HKL generally provided similar results across different conditions, but their PRR was lower than those of MPWKL, JSD, GDI, MI, SHE, and PWCDI algorithms in most conditions.

Average Computation Times of Item Selection Algorithms: The computation times of various item selection algorithms, considering different item qualities and numbers of attributes for 10 test lengths, were measured separately for each algorithm. These calculations were performed in milliseconds for a single examinee on a computer with an i7-7700HQ processor. The average computation times are presented in Table 5. Furthermore, Figures 2 and 3 show the relative average computation times of the item selection algorithms compared to the GDI algorithm for five and six attributes, respectively. The other algorithms' relative average computation times were calculated compared to the GDI because, after random selection, it consistently had the lowest average computation time under all conditions. Given the substantially lower PRR values of the random selection compared to other algorithms, it was excluded from consideration as a reference algorithm.

Table 5.

Average Computation Time of Item Selection Algorithms for an Examinee at Fixed-Length CD-CAT (10 items, milliseconds)

K	Item Quality	Item Selection Algorithms										
		Random	GDI	HKL	JSD	KL	MPWKL	MI	PWACDI	PWCDI	PWKL	SHE
5	LD-LV	2,49	19,4	53,81	524,34	43,83	980,06	55,04	75,28	77,76	46,51	60,27
	LD-HV	2,57	20,92	57	514,09	46,73	984,6	60,12	84,39	84,32	49,49	63,54
	HD-LV	2,54	20,91	56,79	510,72	46,39	979,08	60,14	85,38	84,63	49,15	63,57
	HD-HV	2,72	20,81	56,49	518,86	46,25	980,41	60	84,41	84,2	49,39	62,99
6	LD-LV	4,03	26,56	75,85	901,22	63,49	1673,60	81,4	238,62	235,13	65,46	85,24
	LD-HV	4,21	26,68	76,52	894,85	64	1692,10	83,1	249,64	249,23	66,86	85,31
	HD-LV	3,98	26,58	75,75	892,6	63,46	1689,31	82,62	248,14	248,15	66,28	85,2
	HD-HV	4,32	27,29	77,37	899,6	64,78	1692,96	84,34	256,54	259,03	68,38	86,71

Table 5 shows that the GDI algorithm (19.4-27.29 ms) has a shorter average computation time than other algorithms, except for random selection. Figures 2 and 3 demonstrate that, at both attribute levels, the average computation time of the MPWKL and JSD algorithms is significantly higher than that of the other algorithms. The algorithm with the highest average computation time is MPWKL (980.06-1692.96 ms). When Figure 2 and Figure 3 are examined, the algorithms with the lowest relative average computation times at both quality levels are GDI, KL, PWKL, HKL, MI, SHE, PWACDI, JSD, and MPWKL, respectively.

Figure 2.

Rates of Average Computation Times of Item Selection Algorithms to Average Computation Time of GDI for an Examinee with K= 5 for Fixed-Length CD-CAT

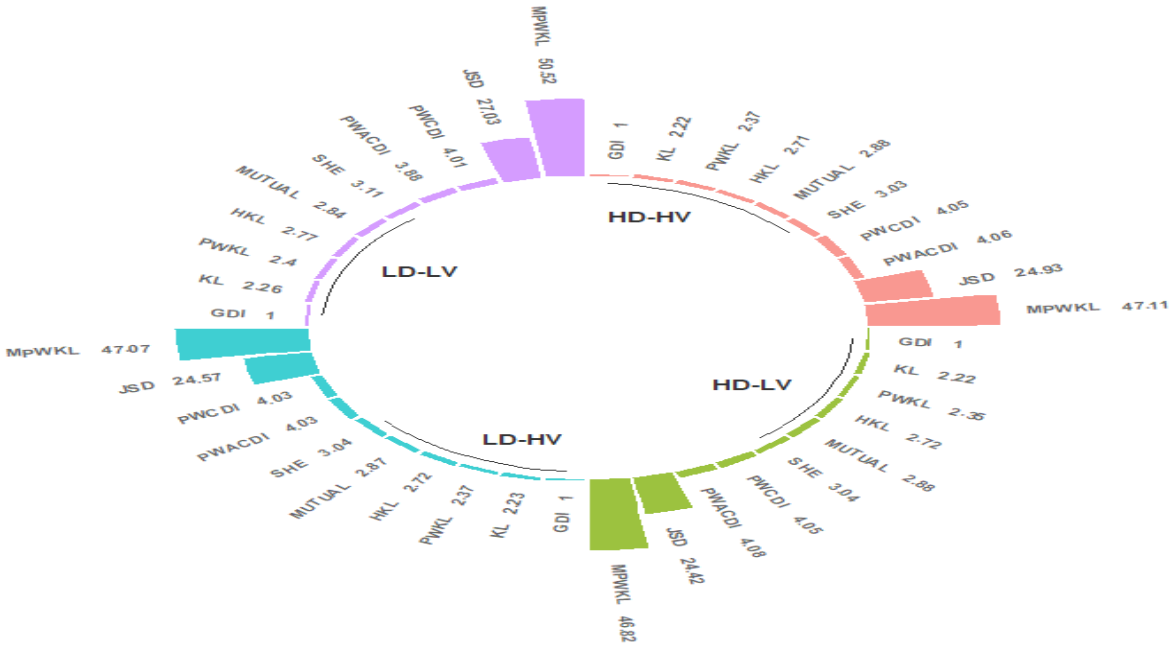
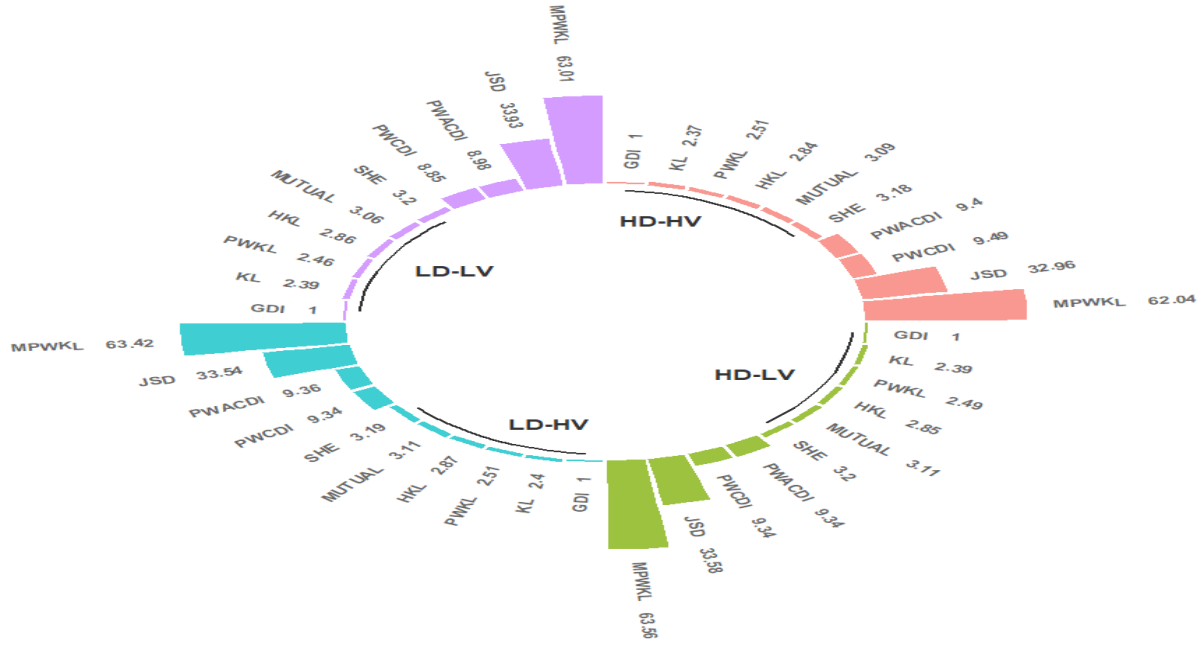


Figure 3.

Rates of Average Computation Times of Item Selection Algorithms to Average Computation Time of GDI for an Examinee with K=6 for Fixed-Length CD-CAT



According to Table 4, the GDI has the smallest proportional increase in computation time as the number of attributes increases. The relative average computation times for PWKL, HKL, KL, MI, and SHE compared to GDI show very small increases as the number of attributes rises. Specifically, the relative average computation times for JSD and MPWKL increased by approximately 1.35 times with the number of attributes, while PWCDI and PWACDI showed an increase of approximately 2.30 times. It can be said that PWCDI and PWACDI are more significantly affected by the increase in the number of attributes compared to other algorithms.

Variable-Length CD-CAT

Average test length of item selection algorithms: Descriptive statistics for item selection algorithms at various item quality levels are given in Table 6, and average test lengths are graphically represented in Figure 4. In addition, the number of examinees who could not complete the test at different item quality levels is shown in Table 7.

Table 6.

Descriptive Statistics of Item Selection Algorithms in the Variable-Length CD-CAT (p_1=0.80; p_2=0.10)

K	Item Quality		Descriptive Statistics / Test Length							
			GDI	HKL	JSD	MPWKL	MI	PWACDI	PWCDI	PWKL
5	LD-LV	Min.	6	4	6	6	6	6	6	4
		Max.	40	40	40	40	40	40	40	40
		Average	13,27	13,81	12,83	13,19	13,18	14,23	13,39	13,87
	LD-HV	Min.	6	4	6	6	6	6	6	4
		Max.	34	35	31	34	33	38	30	36
		Average	11,65	12,06	11,24	11,67	11,54	11,94	11,58	12,14
	HD-LV	Min.	5	3	5	4	5	5	5	3
		Max.	20	17	21	23	23	32	21	18
		Average	7,25	7,58	6,74	7,12	7,16	7,60	7,11	7,61
	HD-HV	Min.	4	2	5	4	4	5	5	2
		Max.	12	19	5	12	9	15	11	20
		Average	5,49	7,18	5,00	5,97	5,49	6,34	6,21	6,99
6	LD-LV	Min.	7	6	7	7	7	6	6	7
		Max.	40	40	40	40	40	40	40	40
		Average	16,70	17,22	16,40	16,69	16,73	17,88	16,78	17,24
	LD-HV	Min.	7	5	6	7	7	5	6	6
		Max.	40	40	40	38	39	40	40	39
		Average	13,87	14,01	13,51	13,87	13,84	14,65	13,85	14,19
	HD-LV	Min.	6	4	6	6	6	4	5	4
		Max.	22	24	22	23	25	28	24	25
		Average	8,62	9,30	8,31	8,60	8,55	9,29	8,71	9,38
	HD-HV	Min.	6	3	6	6	6	4	5	5
		Max.	16	24	15	16	15	21	15	18
		Average	6,82	8,08	6,47	7,06	6,80	7,56	7,33	7,89

Figure 4.

Average Test Length of Item Selection Algorithms for Variable-length CD-CAT

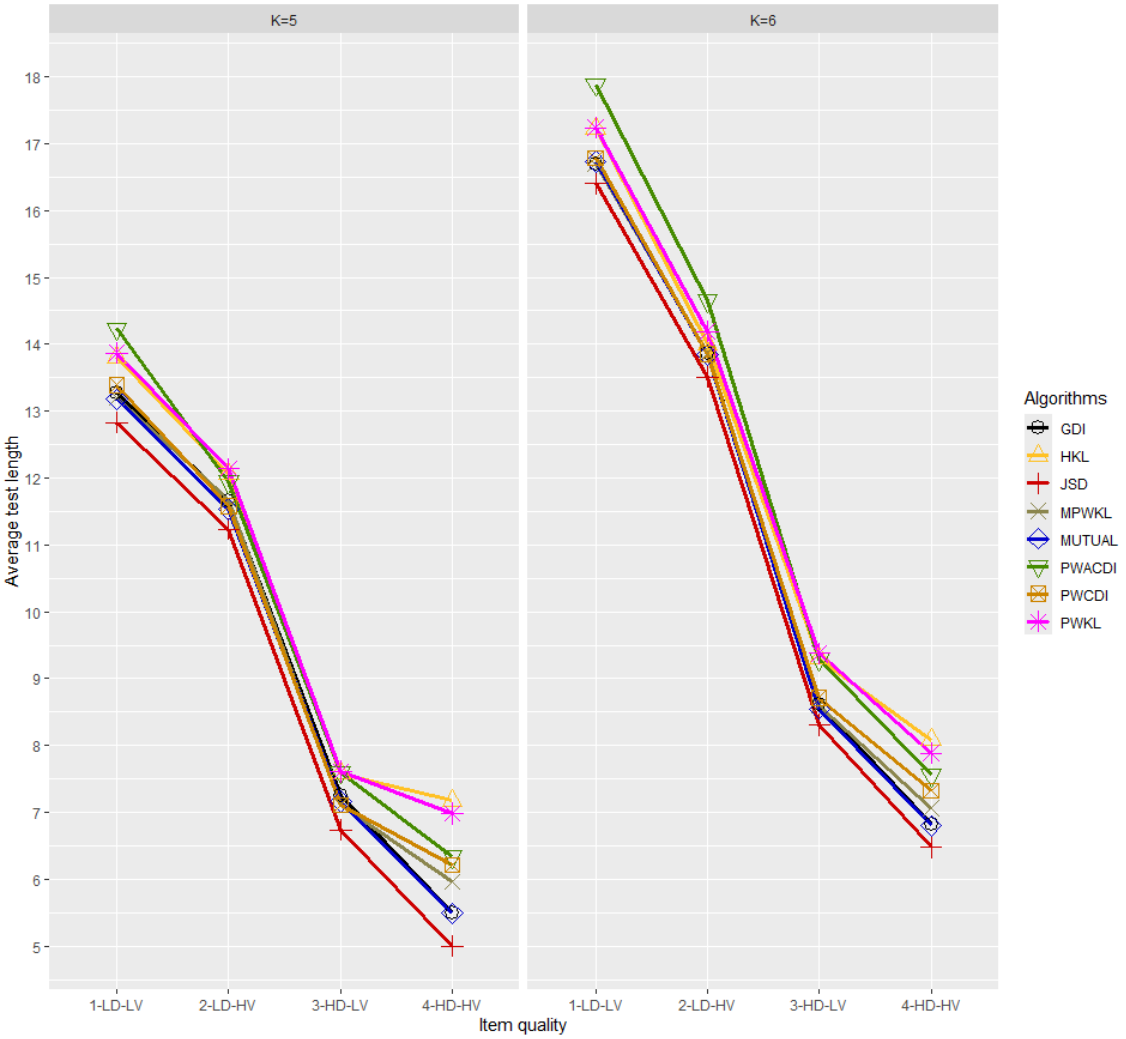


Table 7.

Number of Examinees Who Could not Complete the Test in the Variable-Length CD-CAT (N=3000, p_1=0.80; p_2=0.10, Maximum Test Length=40)

K	Item Quality	GDI	HKL	JSD	MPWKL	MI	PWACDI	PWCDI	PWKL
5	LD-LV	1	2	2	2	2	15	2	7
	LD-HV	0	0	0	0	0	0	0	0
	HD-LV	0	0	0	0	0	0	0	0
	HD-HV	0	0	0	0	0	0	0	0
6	LD-LV	13	12	8	11	7	51	21	13
	LD-HV	2	1	1	0	0	2	1	0
	HD-LV	0	0	0	0	0	0	0	0
	HD-HV	0	0	0	0	0	0	0	0

When analyzing Table 6 and Figure 4, it is seen that the average test lengths of the algorithms varied between 12.83 and 14.23 for the LD-LV item quality level at K=5 and from 16.40 to 17.88 at K=6. For

the LD-HV item quality level at K=5, the average test lengths ranged from 11.24 to 12.06 and from 13.51 to 14.65 at K=6. The average test lengths for HD-LV ranged from 6.74 to 7.61 at K=5 and 8.31 to 9.38 at K=6. For the HD-HV item quality level, the average test lengths varied between 5.00 and 7.18 at K=5 and 6.47 and 8.08 at K=6. Table 6 shows that the average test lengths of the algorithms increased with the number of attributes. Moreover, as the variance in item quality increased, the average test lengths of all algorithms also increased. However, the increase in item discrimination had a more significant impact on the average test lengths than the increase in variance in item quality. The PWACDI algorithm yielded the maximum average test length for items with low discrimination. When item discrimination increased, the HKL algorithm showed a higher average test length than PWACDI at K=5 and higher than HKL and PWKL at K=6. Particularly at the HD-HV level, the difference in average test lengths between these algorithms and others is more pronounced. The JSD algorithm produced the lowest average test length across all item quality levels. The average test lengths of MPWKL, GDI, and MI were similar for the LD-LV, LD-HV, and HD-LV item quality levels. However, at the HD-HV level, and for both K=5 and K=6, the average test length of MPWKL, GDI, and MI algorithms was longer. At K=6, the average test length of the PWCDI algorithm was close to that of MPWKL, GDI, and MI, except at the HD-HV level, where it was higher. At K=5, the average test length of PWCDI was higher than MPWKL, GDI, and MI at most item quality levels, except for HD-LV, which was very close to the average test lengths of these algorithms.

In Table 7, it is shown that some examinees could not complete the test at K=5 for the LD-LV item quality level, while all examinees completed the test for the other item quality levels according to the termination rule. Specifically, for the LD-LV item quality level, fifteen examinees in the PWACDI could not complete the test. Similarly, seven examinees in PWKL, one examinee in the GDI, and two examinees in other algorithms could not complete the test. At K=6, some examinees could not complete the test in any algorithm for the LD-LV item quality level. For the LD-LV, all examinees completed the test for the MPWKL, MI, and PWKL algorithms, while two examinees in the GDI and PWACDI algorithms and one examinee in other algorithms could not complete the test.

Average Computation Times of Item Selection Algorithms: For the variable-length CD-CAT, the average computation times of the item selection algorithms were calculated for various item quality levels and numbers of attributes, similar to the fixed-length CD-CAT. Additionally, the ratio of the average computation time of each algorithm to that of the GDI algorithm is given in Table 8. The relative computation times are graphically represented in Figure 5 for K=5 and Figure 6 for K=6.

Table 8.
Average Computation Time of Item Selection Algorithms for an Examinee at Variable-Length CD-CAT (10 items, milliseconds)

K	Item Quality	GDI	HKL	JSD	MPWKL	MI	PWACDI	PWCDI	PWKL
5	LD-LV	102	294	675	1931	290	467	396	254
	LD-HV	95	226	584	1806	264	362	339	188
	HD-LV	51	146	360	1221	158	254	234	128
	HD-HV	38	211	335	1080	119	181	177	121
6	LD-LV	177	504	1446	4133	513	1370	1292	438
	LD-HV	133	360	1249	3723	422	1164	1016	323
	HD-LV	93	251	797	2448	280	772	749	222
	HD-HV	63	214	655	1941	199	588	581	184

In Table 8, we can see that the average computation times of the algorithms are similar when using a fixed-length CD-CAT. GDI is the fastest, while MPWKL is the slowest algorithm. The JSD algorithm has the second slowest average computation time, following MPWKL. The average computation time decreases as item quality increases, but it increases significantly with more attributes.

Figure 5.

Rates of Average Computation Times of Item Selection Algorithms to Average Computation Time of GDI for an Examinee with K=5 for Variable-Length CD-CAT

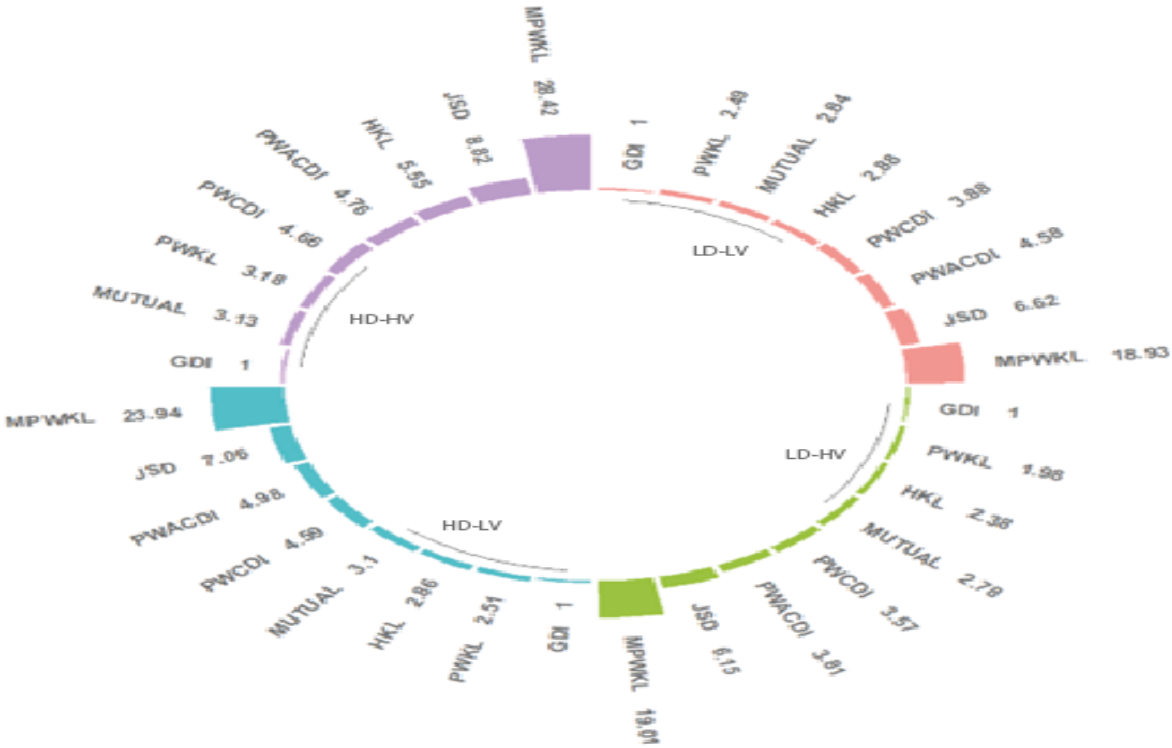
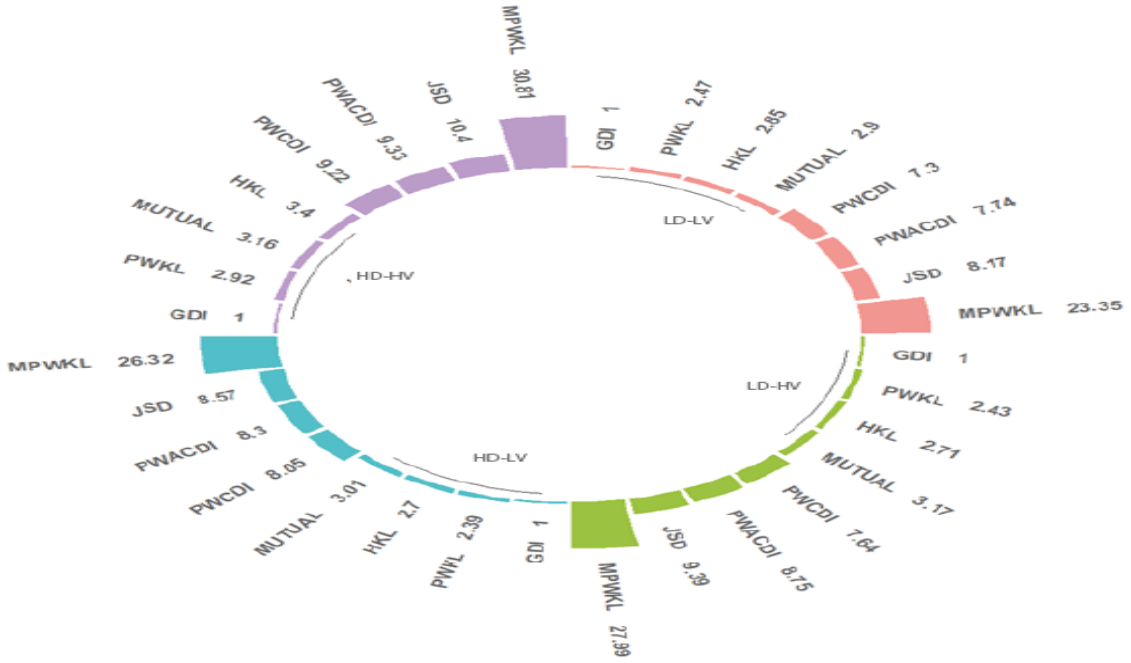


Figure 6.

Rates of Average Computation Times of Item Selection Algorithms to Average Computation Time of GDI for an Examinee with K=6 for Variable-Length CD-CAT



Figures 5 and 6 demonstrate that the GDI algorithm consistently has the lowest average computation time across all conditions. The increase in the relative average computation time of the HKL algorithm at the HD-HV level is more pronounced for 5 attributes. Additionally, the relative average computation

time of the HKL shows minimal variation with an increase in the number of attributes. When the number of attributes is 5, the JSD is approximately 7-9 times slower than the GDI. As the number of attributes increases, this ratio escalates to 8-10 times. The MPWKL is 19-24 times slower than the GDI algorithm at the 5 attribute level and 23-31 times slower at the 6 attribute level. Furthermore, as item discrimination, item variance, and the number of attributes increase, the relative average computation time of the MPWKL increases significantly. The MI computes 2.81-4.98 times slower than the GDI at 5 attributes, with only a slight increase in these ratios when the number of attributes rises to 6. The PWACDI algorithm is 4.42-4.70 times slower than the GDI algorithm at the 5 attribute level, with this ratio increasing to 7.74-9.33 when the number of attributes is 6. Similarly, for the 5 attribute condition, the relative average computation time of the PWCDI algorithm ranges from 3.88 to 4.66, while for the 6 attribute condition, these rates vary between 7.30 and 9.22. The PWKL algorithm, on the other hand, has a relative computation ratio ranging from 2 to 3 times at both the 5 and 6 attribute levels.

Discussion, Conclusion and Recommendations

Nowadays, most psychometric studies focus on tests that measure one-dimensional latent attributes. These tests are often used in outcome-based assessments such as selection and placement. DiBello and Stout (2007) expressed the demand for measurement tools for formative assessments by teachers and education administrators in recent years. Giving fast and accurate feedback plays an important role in process evaluation to increase teaching effectiveness in classroom environments. In order to give effective and accurate feedback to the examinee, the strengths and weaknesses of the examinees must be determined accurately and properly. CDM can be useful in this context. However, giving quick feedback to examinees with a measurement tool developed based on CDM can be difficult due to time limitations in classroom environments. In this respect, the CD-CAT application provides convenience by giving quick feedback to the examinee as soon as possible.

In this study, two different simulation studies were carried out to examine the performance of item selection algorithms under various conditions. In the first simulation study, fixed-length CD-CAT, item selection algorithms through different item quality levels and attributes were evaluated regarding pattern recovery rates and average computation time. In the second simulation study, variable-length CD-CAT, the performance of item selection algorithms through various item quality and number of attribute levels was evaluated according to the average test length and computation times criterion.

In this study, the PRR of item selection algorithms decreased as the number of attributes increased. This is primarily due to the increase in the number of possible cognitive patterns as the number of attributes increases. For instance, with 5 attributes, there are 32 possible cognitive patterns, whereas this number increases to 64 with 6 attributes. Additionally, as item quality and test length increased, the PRR of the algorithms converged for both 5 and 6 attribute conditions. These findings are consistent with those reported by Wang (2013), Lin and Chang (2018), and Huang (2018).

The fixed-length CD-CAT study concluded that random selection is unsuitable for use since the pattern recovery rates of random selection are the lowest in all conditions. This finding is consistent with those reported in previous studies by Cheng (2009), Kaplan et al. (2015), Xu et al. (2003), Wang (2013), and Yigit et al. (2019). The primary reason for the consistently low PRR of the random selection across all conditions is that it does not consider the items' characteristics or the examinee's previous responses during item selection. Besides, it was also found that the attribute and pattern recovery rates of the KL are lower than those of other algorithms. Xu et al. (2003), Cheng (2009), and Zheng and Chang (2016) reported that the PRR of the KL is lower than that of other algorithms, except for the random selection. This study corroborates these results, confirming that the KL algorithm has the lowest PRR following random selection. The main reason is that the KL algorithm treats the probability of each cognitive pattern being the actual cognitive pattern as equal during the estimation process. In contrast, other algorithms adjust the weights of each cognitive pattern based on posterior probabilities after each item is administered, thereby providing more accurate estimations of the true cognitive pattern.

This study found that the PRR of the MI was higher than those of the SHE algorithm for tests with 5 attributes and shorter lengths. However, when the test length increased, and the number of attributes was 6, the PRR of these algorithms converged. These findings are consistent with those reported by Wang (2013). While the HKL and PWKL yielded similar PRR, the HKL generally exhibited a slightly higher PRR than the PWKL. Additionally, the PRR of the SHE and PWKL are very close to each other. These findings align with those reported by Cheng (2009).

In short tests, when the discrimination and variance values of the items were low, the PRR of the JSD and MPWKL were higher than the other algorithms. Kaplan et al. (2015) compared the correct classification rates of the MPWKL, GDI, and PWKL item selection algorithms at 10, 20, and 40 test lengths and across varying item quality levels. The results of that study indicated that the MPWKL and GDI achieved similar classification rates, whereas the PWKL demonstrated a lower classification rate than the MPWKL and GDI. Zheng and Chang (2016) analyzed the PRR of the MI, MPWKL, PWKL, KL, CDI, ACDI, PWCDI, and PWACDI algorithms for test lengths of 5 and 10. They found that the MPWKL, PWCDI, and PWACDI algorithms had the highest PRR, followed by the MI, CDI, ACDI, and KL algorithms. Yigit et al. (2019) compared the classification accuracy of JSD, GDI, and random selection algorithms under the MC-DINA model at test lengths of 5, 10, and 20, and under conditions of low and high discrimination-variance. They reported that the correct classification rates of the JSD were higher than those of the GDI and random selection in most conditions. The findings of this study are consistent with those reported by Cheng (2009), Kaplan et al. (2015), Wang (2013), Yigit et al. (2019) and Zheng and Chang (2016). However, in terms of measurement accuracy, it was observed that JSD and MPWKL could not measure with sufficient accuracy except for 5 test lengths and HD-HV level. In this respect, while using JSD and MPWKL algorithms with 5 test lengths and HD-HV levels can be recommended, longer tests are recommended for these algorithms in different item quality conditions.

In the fixed-length and variable-length CD-CAT studies, GDI had the lowest average computation time, while MPWKL had the highest. This is because, unlike other item selection algorithms, the computational complexity of GDI does not increase exponentially with the number of attributes. Zheng and Chang (2016) found that MPWKL and PWCDI had the longest computation times. The current study's average computation times for MPWKL, JSD, and PWCDI were higher than those for other algorithms. However, the findings related to the amount of time differ from those of Kaplan et al. (2015) and Zheng and Chang (2016). One possible reason could be that Kaplan et al. (2015) worked with a limited number of cognitive patterns, whereas this study used all possible cognitive patterns. Another reason could be that the cognitive pattern estimation method was used. Zheng and Chang (2016) used the MLE estimation method, while this study used the MAP method, which adds values for each cognitive pattern by multiplying the likelihood value with the prior probability value after each item is administered. Additionally, EAP estimation was performed within the CD-CAT process, and items administered and estimated cognitive patterns were recorded in a matrix after each item was administered, potentially affecting computation time. In this study, R 3.6.1 was used for statistical calculations. It is believed that software differences may influence the average computation time of the item selection algorithms.

However, considering measurement accuracy and the average computation times of the JSD and MPWKL, the JSD can be preferred primarily because it performs faster computation. Since item selection algorithms give more accurate results on 10 tests or more, it can be said that 10 test lengths are sufficient for classroom assessments for item banks consisting of items with high discrimination in practical studies. As item quality and test length increase, the classification accuracies of item selection algorithms are close to each other and approach 1. In this respect, when the measurement accuracy and computation time of the item selection algorithms are evaluated together, although the measurement accuracy of the GDI algorithm is slightly smaller than the JSD and MPWKL algorithms, it is recommended to be used in long tests and for item banks with high item discrimination, since the average computation time is faster. MI, SHE, PWKL, HKL, and PWCDI can also be used in long tests (20), and banks consist of items with high discrimination. Due to the decrease in measurement accuracy as the number of attributes increases, in practical applications, it is recommended to avoid very long attribute

numbers or to use longer tests and items with high discrimination in cases where the number of attributes is high.

In a comprehensive review of relevant literature, Kaplan et al. (2015) found that the average test lengths for the MPWKL and GDI were similar across all item quality levels in the variable-length CD-CAT study. However, the PWKL exhibited longer average test lengths than these two algorithms. In another study, Kaplan (2016) reported that the GDI algorithm had a lower average test length than the PWKL, with this difference becoming more pronounced as the number of attributes increased. Additionally, Zheng and Chang (2016) determined that under low item quality conditions, the PWCDI had the shortest average test length, followed by the PWACDI, MI, and PWKL, with MI and PWKL showing similar average test lengths. The shortest average test lengths were observed for the PWCDI and MI in high-item quality conditions, followed by the PWACDI and PWKL. Finally, Yiğit et al. (2019) reported that the JSD had a shorter average test length than the GDI under all conditions. In this study, the JSD consistently had the shortest average test length across all conditions. The average test lengths for the MPWKL, GDI, MI, and PWCDI were similar and slightly longer than those for the JSD. The PWACDI, HKL, and PWKL had longer average test lengths than the other algorithms. These findings are consistent with those reported in other studies within the related literature (Kaplan et al., 2015; Kaplan, 2016; Yiğit et al., 2019; Zheng & Chang, 2016).

In the variable-length CD-CAT study, it was concluded that an increase in item discrimination and variance in item quality results in a decrease in the average test length. Conversely, increasing the number of attributes leads to longer average test lengths. Due to the increase in average test length with a higher number of attributes, it is recommended to avoid an excessive number of attributes or to limit the maximum number of attributes measured by each item. At the HD-HV item quality level, the average test lengths of the algorithms range from 5 to 7 for K=5 and from 6 to 8 for K=6. Consequently, it is posited that classroom assessments with high-quality item banks will facilitate the effective utilization of CD-CAT. Although the JSD algorithm demonstrates the shortest average test length under all conditions, its average computation time exceeds that of other algorithms, except for MPWKL for low item quality levels. Therefore, it is recommended to utilize the JSD algorithm when item quality is high for short tests. When item quality is low, considering computation time, it is advisable to use the GDI and MI algorithms in addition to the JSD algorithm.

In this study, two criterion rules (Hsu et al., 2013) were used in variable-length CD-CAT. In this rule, the highest posterior probability value of the cognitive pattern was 0.80, and the second highest posterior probability value was 0.10. Hsu et al. (2013) suggested that these values should be considered as 0.90 and 0.05, respectively, in high-stake tests. Similar work can be performed using different posterior probability values. Moreover, the maximum test length limitation (40) was determined, as well as the posterior probability value. The performance of item selection algorithms can be examined by changing this value.

In this study, the DINA model was only utilized among the various cognitive diagnostic models. Similar studies can be performed again for different CDMs. In addition, since only the DINA model was used in the study, the Q matrix was developed only under this model. In practice, however, some datasets may fit different CDMs. For this reason, similar studies can be carried out for Q matrices consisting of mixed models.

The results of this study hold significant practical implications. The proposed algorithms are expected to guide future research and practical applications by facilitating the use of shorter tests and reducing the overall testing duration.

Declarations

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical Approval: This study did not necessitate ethical approval as it utilized simulated data

References

Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74(4)*, 619–632. https://dx.doi.org/10.1007/S11336-009-9123-2

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

de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement, 33(3),* 163–183. https://doi.org/10.1177/0146621608320523

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199. https://doi.org/10.1007/S11336-011-9207-7

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. https://doi.org/10.1177/0146621610377081

DiBello, L., Roussos, L. A., & Stout, W. F. (2007). Handbook of Statistics. C. R. Rao ve S. Sinharay (Ed). *Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models.* 26, 979-1030. https://doi.org/10.1016/S0169-7161(06)26031-0

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Corwin Press, https://doi.org/10.4135/9781452219493

Hsu, C. L., Wang, W. H., & Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37(7),* 563-582. https://doi.org/10.1177/0146621613488642

Huang, H. (2018). Effects of Item Calibration Errors on Computerized Adaptive Testing under Cognitive Diagnosis Models. *Journal of Classification*, 35:437-465. https://doi.org/10.1007/s00357-018-9265-y

Izrailev, S. (2020). *tictoc: Functions for Timing R Scripts, as well as Implementations of "Stack" and "StackList" Structures_*. R package
version 1.2.1, <https://CRAN.R-project.org/package=tictoc>

Kaplan, M. (2016). Nitelik Sayısının Madde Seçme Algoritmalarının Performansı Üzerindeki Etkisi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 7(2), 285-295. https://doi.org/10.21031/epod.268486

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. https://doi.org/10.1177/0146621614554650

Lin, C.-J., & Chang, H.-H. (2019). Item Selection Criteria with Practical Constraints in Cognitive Diagnostic Computerized Adaptive Testing. *Educational and Psychological Measurement*, 79(2), 335–357. https://doi.org/10.1177/0013164418790634.

Liu, H., You, X., Wang, W., Ding, S., & Chang, H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*, 30, 152-172. https://doi.org/10.1007/s00357-013-9128-5

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Cham, Switzerland: Springer International Publishing.

McGlohen, M.K., & Chang, H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavioral Research Methods, 40,* 808–821. https://doi.org/10.3758/BRM.40.3.808

Minchen, N. D., & de la Torre, J. (2016, July). *The continuous G-DINA model and the Jensen-Shannon divergence*. Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "Two Disciplines" Problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24(1),* 307–353. https://doi.org/10.3102/0091732X024001307

R Core Team (2021*). R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6(4), 219–262. https://doi.org/10.1080/15366360802490866

Stiggins, R. J. (2002). Assessment Crisis: The Absence of Assessment for Learning. *Phi Delta Kappan*, *83*(10), 758–765. https://doi.org/10.1177/003172170208301010

Stocking, M.L. (1994). *Three practical issues for modern adaptive testing item pools*. Ets research report series, 34. https://doi.org/10.1002/j.2333-8504.1994.tb01578.x

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of Royal Statistics, 65*, 143–157.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society, 51*, 337–350.

Thissen, D., & Mislevy, R. J. (2000). Computerized Adaptive Testing: A primer. H. Wainer, (Ed). *Testing algorithms*, Mahwah, NH: Lawrence Erlbaum Associates, Inc, p. 101-133.

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

van der Linden, W.J., & Glas, G.A.W. (2002). *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report 05-16). Princeton, NJ: Educational Testing Service.

von Davier, M., & Cheng, Y. (2014). Multistage testing using diagnostic models. İn D. L. Yan, A. A. von Davier & C. Lewis (eds.), *Computerized multistage testing: Theory and applications* (p. 219-227). New York, NY: CRC Press.

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing with Short Test Length. *Educational and Psychological Measurement*, *73*(6), 1017–1035.

Wiliam, D. (2011). What Is Assessment for Learning? *Studies in Educational Evaluation*, 37, 3-14. https://doi.org/10.1016/j.stueduc.2011.03.001

Wickham, H., (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York,.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized Adaptive Testing for Cognitively Based Multiple-Choice Data. *Applied Psychological Measurement*, *43*(5), 388–401.

Zheng, C. (2015). *Some practical item selection algorithms in cognitive diagnostic computerized adaptive testing—Smart diagnosis for smart learning.* Unpublished Doctoral Dissertation. University of Illinois at Urbana–Champaign.

Zheng, C., & Chang, H.-H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 40*, 608-6