Journal of Teacher Education and Lifelong Learning
(TELL)

Volume:4 Number:2 Year: 2022 Research Article ISSN: 2687–5713

The Effect of Strong and Weak Unidimensional Item Pools on Computerized Adaptive Classification Testing

Ceylan Gündeğer[1] Sümeyra Soysal[2]

[1] Aksaray University, Department of Educational Sciences, Aksaray, Türkiye
cgundeger@gmail.com
[2] Necmettin Erbakan University, Department of Educational Sciences, Konya, Türkiye
sumeyrasoysal@hotmail.com

Article Info

Article History
Received: 12.09.2022
Accepted: 02.12.2022
Published: 31.12.2022

Keywords:
Computerized adaptive classification testing, unidimesionality, strong unidimensionality, weak unidimensionality, item pool

ABSTRACT

Computerized Adaptive Classification Tests (CACT) aim to classify individuals effectively with high classification accuracy and few items over large item pools. The characteristic features of the item pool include the number of items, item factor loadings, the distribution of the Test Information Function, and dimensionality. In this study, we present the results of a comprehensive simulation study that examined how item selection methods (MFI-KLI), ability estimation methods (EAP-WLE) and classification methods (SPRT-CI) were affected by strong and weak unidimensional item pools. Findings of the study indicate that CI had always produced results with classification accuracy similar to SPRT but with a test length of almost half. Additionally, KLI and MFI item selection methods were not affected by the item pool characteristic as weak or strong unidimensionality. From the findings of this study, it can be recommended to use CI with EAP in CACT studies, whether the item pool is weak or strong unidimensional, but WLE only under strong unidimensional item pools. Additionally, the EAP and SPRT methods are recommended to prefer in the weak unidimensional item pool.

Citation: Gündeğer, C. & Soysal, S. (2022). The Effect of Strong and Weak Unidimensional Item Pools on Computerized Adaptive Classification Testing, *Journal of Teacher Education and Lifelong Learning*, *4*(2), 310-321.

INTRODUCTION

There is a trend toward computer-based tests from paper-pencil tests in measuring individuals' ability. While all individuals answered the same items within a certain time in the paper-pencil tests; items are adjusted in accordance with the individual's ability level in Computerized Adaptive Tests (CAT). Thanks to Item Response Theory 's (IRT) bringing the items and individual (thetas) parameters on the same scale and the invariance of these parameters, each individual can complete the test in a shorter time by answering the questions appropriate to their ability level. Therefore, the ability levels of individuals can be estimated more quickly and with higher reliability, even if individuals' response different items. In educational and psychological measurement applications, decisions such as passed – failed or sick – healthy are made about individuals via Computerized Adaptive Classification Testing (CACT). In CACT, it is aimed to classify individuals into categories with few items and high classification accuracy and reach critical decisions about individuals.

Weiss and Kinsbury (1984) mentioned six main components of CAT that are (i) response model; (ii) item pool; (iii) starting rule; (iv) item selection method; (v) ability estimation method; and (vi) Termination rule. In CACT, these first five components remain constant, and the termination rule is provided by classification criteria. When comparing unidimensional IRT models as 1PL, 2PL, or 3PL, it was seen that the response model impacts the test ending (Jiao & Lau, 2003; Lau, 1996; Reckase, 1983). In the literature, item selection methods, ability estimation methods and classification criteria also affect the test length, classification accuracy and errors like bias, RMSE, or absolute error (Eggen, 1999; Eggen & Straetmans; 2000; Gündeğer, 2017; Lau & Wang, 1999; Lin & Spray, 2000; Nydick et al., 2012; Spray & Reckase, 1994; Spray & Reckase, 1996; Thompson, 2007a; Thompson & Ro, 2007; Thompson, 2009; Thompson, 2011).

Wang and Wang (2001) state that the ability estimation methods are an important CACT component that affects both the selection of the items proper for the estimated ability level and the termination of the test. In the literature, it is seen that Maximum Likelihood Estimation, Weighted Likelihood Estimation (WLE), Expected a Posteriori (EAP), Maximum a Posteriori methods are frequently examined among the ability estimation methods (Breslow & Holubkov; 1997; Cheng & Liou, 2000; Diao & Reckase, 2009; Eggen & Straetmans, 2000; Gökçe, 2012; Kalender, 2011; Kezer, 2013; Penfield & Bergeron, 2005; Tao, Shi & Chang, 2012; Wang, 1997; Wang et al., 1999; Wang & Vispoel, 1998; Wang & Wang, 2001; Warm, 1989; Wouda & Eggen, 2009; Yi, Wang & Ban, 2000). Note that all of these estimation methods make biased estimations to some extent (Warm, 1989). In CACT applications, it is aimed to estimate theta as accurately as possible and select the items appropriate for the estimated theta. Based on this, it can be interpreted that the performances of item selection methods, ability estimation methods and classification criteria depend on each other.

When the literature is examined, it is seen that Maximum Fisher Information (MFI) and Kullback-Leibler Information (KLI) are frequently used among the item selection methods in CACT (Ayan, 2018; Cheng & Liou, 2000; Diao & Reckase, 2009; Eggen, 1999; Eggen & Straetmans, 2000; Gündeğer, 2017; Lau & Wang, 1999; Lin & Spray, 2000; Spray & Reckase, 1994; Thompson, 2007a; Thompson, 2009; Thompson & Ro, 2007). MFI is based on the selection of the item that gives the highest information at the estimated theta level, whereas KLI is based on the selection of the item that gives the highest information at and around the estimated theta level (Eggen, 1999; Reckase, 1983; Spray & Reckase, 1994). These two methods consider the estimated ability level as well as be cut-point based (Thompson, 2007b). In CACT, among the classification criteria, Sequential Probability Ratio Test (SPRT), Generalized Likelihood Ratio (GLR) and Confidence Interval (CI) have also often studied in the literature (Ayan, 2018; Eggen & Straetmans, 2000; Gündeğer, 2017; Thompson & Ro, 2007; Thompson, 2009; Thompson, 2011; Nydick et al., 2012). MFI-KLI item selection methods and SPRT-GLR classification methods have the assume of unidimensionality (Eggen, 1999; Lin & Spray, 2000;

Nydick, 2013; Seitz & Frey, 2013; Spray, Abdel-fatah et al., 1997; Spray & Reckase, 1994). When item selection and classification components work together effectively, the classification of individuals is completed in a shorter time with fewer items, as expected from CACT (Spray & Reckase, 1994).

In CAT and CACT applications, maybe the most important part is the item pool which is needed to be quiet large and have high quality. Hsiehi (2015) and Thompson (2009) state that the quality of the items and number of items reduced the test length significantly (Hsiehi, 2015; Thompson; 2009). Generally, the CACT studies explain the number of items in the item pool as a pool characteristic that is reasonable but inadequate. Since the distrubition of the Test Information Function (TIF) and/or item parameters are important for CAT and CACT, the information about these features should be explained too (Gündeğer & Doğan, 2018; Kezer, 2021; Thompson, 2009; Thompson, 2011). Additionally, item banks' unidimensionality (as weak or strong unidimensionality) is a specific part of these characteristics, which is seldom studied in the literature compared to other characteristics but affects the results significantly since the item selection and classification criteria have the assume of it. Therefore, it is important and essential to examine the unidimensionality of the item pool and what kind of unidimensionality the item pool has. The aim of CACT is to make a high accuracy classification with the least number of items and ensuring this depends on the methods to be employed and naturally on the assumptions of these methods.

Unidimensionality means that there is only one latent trait that the items measure and that underlies the individuals' response performance. In other words, unidimensionality is the explanation of the variance between item responses by a single latent trait. Unidimensionality means that the items depend on a dominant dimension (Hambleton & Swaminathan, 1985). Unidimensional IRT models assume that a single latent trait underlies the responses given to the items. Therefore, unidimensional CACTs require a cut-off point (θ_0) separated range on this latent dimension. The true class decision for individual i depends on the student's estimated ability level (θ_i) relative to θ_0. If $\theta_i > \theta_0$, the student will be classified as pass-successful, and any other decision will be made in Type II error. In contrast, if θi < θ0, the student will be classified as fail-unsuccessful and any other decision will cause a Type I error (Finkelman, 2008).

Considering the characteristics of the item pool and the assumption of the item selection methods and the classification criteria, the concepts of weak and strong unidimensionality come to the fore along with unidimensionality. If the inter-item correlations and the factor loads of the items on one dimension are low, the item pool shows a weak unidimensional factor structure, which is close to the properties of multidimensional structures. However, if the inter-item correlations and the factor loads are high, the item pool indicates a strong unidimensional factor structure. So, how do these methods, which have a unidimensionality assumption, perform when the item pool represents weak or strong unidimensionality? Obviously, in practice, it is hard to set the item pool that has a strong unidimensionality but how do two types of unidimensionality affect the test length, estimations and accuracy? That is the main question expected to be answered by this research.

When CAT and CACT studies are examined, it is seen that Monte Carlo (MC) and Post Hoc (PH) simulations are often carried out in the R environment (Ayan, 2018; Demir, 2019; Erdem Kara, 2019; Gündeğer, 2017; Özdemir, 2015). Some studies present the dimensionality of the item pools (e.g., Ayan, 2018; Aybek, 2016; Demir, 2019; Erdem Kara, 2019; Gündeğer, 2017; Özdemir, 2015; Şahin, 2017); some present the information of unidimensionality with the item loads (e.g., Ayan, 2018; Doğruöz, 2018; Gündeğer, 2017; Şenel, 2017); and some present only the number of items as item pool characteristic (Kaçar, 2016). In fact, when MC data are generated by the software based on unidimensionality, it is important to show some evidences about how the items represent the latent trait. Flaugher (2000) states that the better the quality of the item pool, the more successful the individualized test algorithm will perform (Flaugher, 2000). For this purpose, testing unidimensionality in generated data, examining the item factor loads and specifying which type of unidimensionality data is derived is important both in terms of revealing the item pool characteristic and in terms of the performance of item selection methods and classification criteria. The purpose and importance of this study is to determine how CACT conditions perform on the weak and strong unidimensional factor structures produced in a controlled way. Considering all these discussions, in this research, it was mainly aimed to how the test length, classification accuracy, correlation between the real theta and estimated theta, Root Mean Square Error (RMSE) and absolute error changes when the item pools represent weak or strong unidimensionalty. However, it was also examined how item selection methods (MFI-KLI), ability estimation methods (EAP-WLE) and classification methods (SPRT-CI) are affected by weak and strong unidimensionality which is an assumption for item selection and classification methods, specially. For this purpose, answers to the following sub-problems were sought:

1) How is the test length, classification accuracy, correlation between real and estimated ability levels, RMSE and absolute error in the conditions where MFI and KLI item selection methods, EAP and WLE ability estimation methods, SPRT and CI classification methods are crossed in the item pool that represents strong unidimensionality?

2) How is the test length, classification accuracy, correlation between real and estimated ability levels, RMSE and absolute error in the conditions where MFI and KLI item selection methods, EAP and WLE ability estimation methods, SPRT and CI classification methods are crossed in the item pool, which represents weak unidimensionality?

METHOD

This research was based on a Monte Carlo simulation study. Simulation allows researchers to assume the inherent complexity of organizational systems as a given. If other methods answer the questions “What happened, and how, and why?,” simulation helps answer the question “What if?.” Simulation enables studies of more complex systems because it creates observations by “moving forward” into the future, whereas other research methods attempt to look backward across history to determine what happened, and how (Dooley, 2002). In this section of the study, simulation design and data analysis are presented.

Simulation Design

In line with the aim of the study, ability parameters (thetas), item parameters and item response patterns were generated in SimuMIRT (Yao, 2003). Theta were derived from normal distrubition as N(0,1) for 3000 individuals. In strong and weak unidimensional item pools, items were simulated from normal distribution as N [0,1] for b parameters and from beta distrubition as B (6,16) for c parameters. Parameters were generated from Lognormal distribution as Log [2.5; 0.3] for strong unidimensional item bank and as Log [1.2; 0.3] for weak unidimensional item bank. Test Information Function (TIF) graphics of the weak and strong unidimensional item banks are given below, respectively.

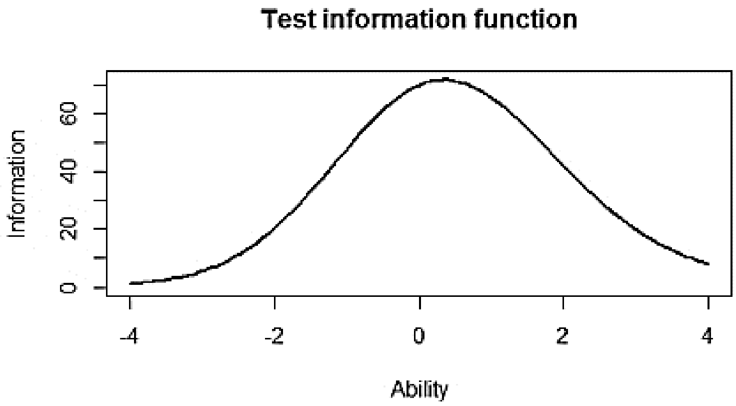

Figure 1. *TIF of the weak unidimensional item banks*

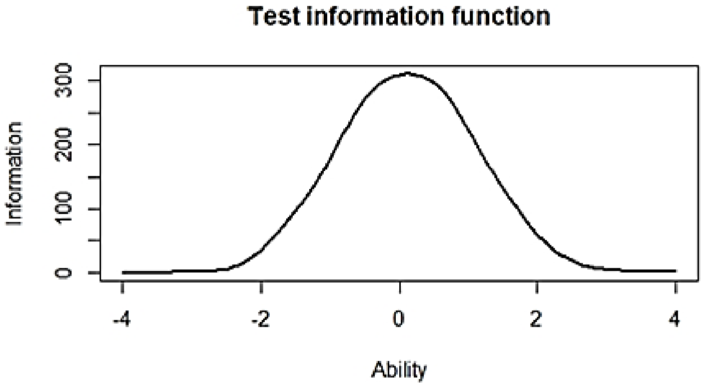

Figure 2. *TIF of the strong unidimensional item banks*

To compare the item pools that could represent the weak and strong unidimensionality similar to Doğan et al. (2017), item factor loads were also fixed in a range. In the strong unidimensional item bank, the item factor loads were set to be in the range of 0.60–1.00 whereas in the weak unidimensional item bank the item factor loads were fixed in the range of 0.30–0.50. After generating ability and item parameters, 25 different item response patterns were simulated in SimuMIRT (Yao, 2003) and item factor loads were checked and confirmed for all patterns using confirmatory factor analysis with lavaan package (Rosseel, 2012). For classification accuracy, the ability points corresponding to the maximum point of the test information function of each item pool were used as the cut-off point. This was determined as 0.5 in both the item pools. Lastly, a Monte Carlo simulation study was performed in the R environment (R Core Team, 2019) in accordance with the study' conditions presented in Table 1.

Table 1. *Simulation conditions of the research*

Conditions	Levels
Item pool/bank unidimensionality	Strong, Weak
Item selection method	MFI, KLI
Ability estimation method	EAP, WLE
Classification method	SPRT, CI

Data Analysis

As seen in Table 1, the independent variables of this study are item bank unidimensionality (weak and strong unidimensional), item selection method (MFI and KLI), ability estimation method (EAP and WLE) and classification method (SPRT and CI). The dependent variables are test length, classification accuracy, correlation between the real theta and estimated theta, Root Mean Square Error (RMSE) and absolute error (AE). Because the simulation study was based on 25 replications, the results were summarized over the average of the replications. RMSE and AE are calculated using the following formula:

$$\text{RMSE} = \sqrt{\frac{\sum_1^N(\theta_t - \theta_e)^2}{N}}$$

$$\text{AE} = \frac{\sum_1^N|\theta_t - \theta_e|}{N}$$

where θ_e is the estimated ability parameter, θ_t is the true ability parameter, N is the number of individuals.

Ethic

Since this study was a simulation study, ethics committee approval was not required.

RESULTS

In line with the purpose of the study, the average values of the test length, classification accuracy, Pearson correlation between the real thetas and estimated thetas, Root Mean Square Error and absolute error are calculated and presented in Table 2.

Table 2. *The average values of the study's conditions*

Item Bank/Pool	Item Selection Method	Ability Estimation Method	Classification Method	TL	CA	r	RMSE	AE
Weak Unidimensional	MFI	EAP	SPRT	49.09	0.92	0.95	0.34	0.27
			CI	20.89	0.91	0.82	0.52	0.43
		WLE	SPRT	49.25	0.92	0.95	0.39	0.31
			CI	19.67	0.90	0.80	0.80	0.61
	KLI	EAP	SPRT	49.08	0.92	0.95	0.34	0.27
			CI	20.77	0.91	0.82	0.52	0.43
		WLE	SPRT	49.25	0.92	0.95	0.39	0.32
			CI	19.68	0.90	0.80	0.81	0.61
Strong Unidimensional	MFI	EAP	SPRT	27.75	0.94	0.94	0.70	0.64
			CI	8.74	0.93	0.87	0.56	0.48
		WLE	SPRT	28.12	0.93	0.94	0.75	0.68
			CI	8.37	0.92	0.87	0.61	0.52
	KLI	EAP	SPRT	27.80	0.94	0.94	0.70	0.64
			CI	8.73	0.93	0.87	0.56	0.48
		WLE	SPRT	28.19	0.93	0.94	0.75	0.68
			CI	8.30	0.92	0.87	0.61	0.52

TL = Test Length; CA = Classification Accuracy; r = Correlation between the real theta and estimated theta, RMSE = Root Mean Square Error; AE = Absolute Error; TL, CA, r, RMSE and AE were calculated by taking the average values of 25 replications.

As shown in Table 2, SPRT has the highest value in all conditions in terms of TL. Based on these findings, it can be said that SPRT needs more items than CI to classify the individuals. In other words, SPRT performed worse than CI in terms of test efficiency. However, a noteworthy point in these findings is that the TL values of SPRT differ significantly between the strong and weak unidimensional item pools. In Figure 3, it is seen that in all conditions, the TL values decreased in the strong unidimensional item pool and SPRT has the most significant decline. While SPRT requires approximately 49 items in the weak unidimensional item pool; it can classify individuals with 28 items in the strong unidimensional item pool. Based on these findings, it can be said that SPRT performs better in the item pools, which represent a strong unidimensionality. The literature showed us that SPRT needs more items to end the CACT and performs worse in terms of the TL (Ayan, 2018; Eggen & Straetmans, 2000; Gündeğer, 2017; Nydick et al., 2012; Thompson, 2011). This may be due to the generated item pools' characteristics. Based on this finding, it can be commented that in order for SPRT to be able to classify individuals with few items, one-dimensionality, which is an assumption of SPRT, must be provided strongly. In other words, if the inter-item correlations and the factor loads are high as in the range of 0.60–1.00, SPRT can classify with less number of items, as expected from CACT applications.

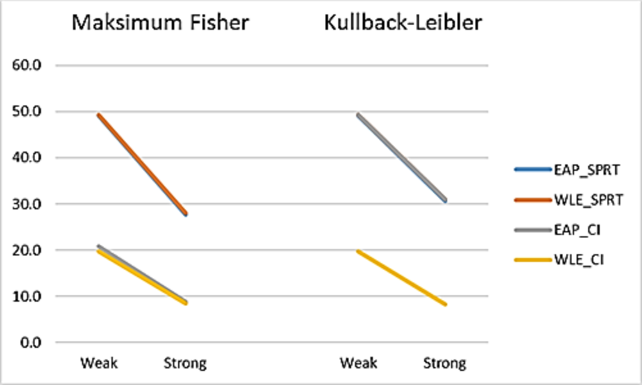

Figure 3. *Findings on test length*

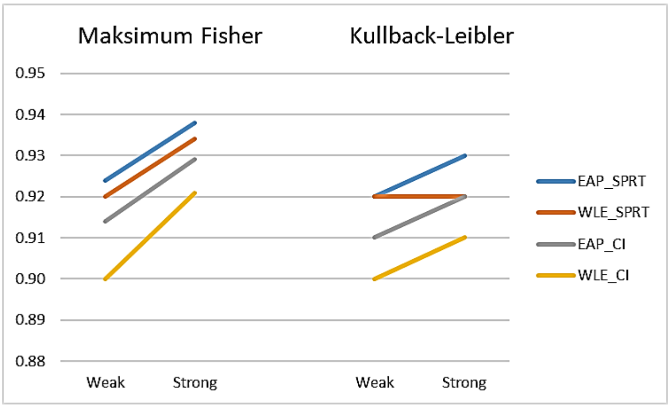

Figure 4. *Findings on classification accuracy*

According to Table 2, in all conditions, the classification accuracy (CA) has a high rate above 90%, which overlaps with the literature (Gündeğer, 2017; Thompson, 2011; Thompson & Ro, 2007; Nydick et al., 2012). In Figure 4, it is seen that almost all the CA values obtained from the strong unidimensional item pool increased compared with the weak unidimensional item pool. Accordingly, it can be said that the item pool characteristics affect the classification of individuals into the pass-fail categories, too. In all conditions, the correlation between the true theta and the estimated thetas has above 0.80 that indicates a positive and high correlation, as expected from CAT and CACT applications. Accordingly, it can be said that in all conditions the methods perform well in terms of the relationships between individuals' true ability levels and estimated ability levels. In Figure 5, it has been concluded that the r values obtained from the CI conditions vary, especially in terms of the item pool characteristic, whereas the r values of the SPRT conditions don't differ much. When the CI method is used in a strong unidimensional item pool, the correlation between the true and estimated values increases. In other words, in the strong unidimensional item pool, the estimated abilities of individuals are quite close to their true ability levels. So, it can be said that CI performs better in terms of r values, when the item pool shows a strong unidimensional characteristic. In the r values of SPRT conditions are higher than the r values of CI values overlap with the literature, too (Gündeğer, 2017).

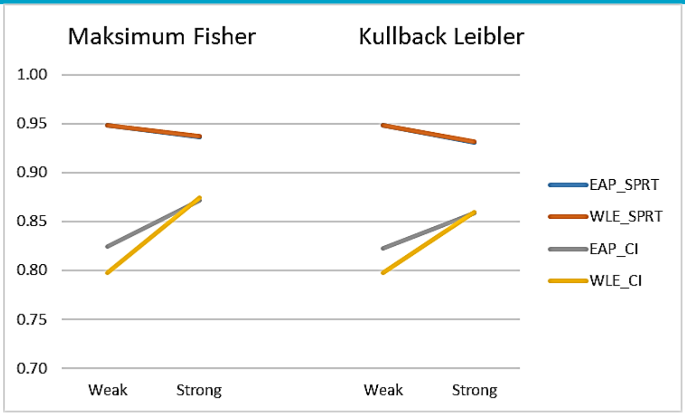

Figure 5. *Findings on correlation*

In Table 2, the RMSE and AE values indicate the error in the final ability estimations. Although there is no absolute threshold for errors, it is possible to make a relative comparison. It can be said that the lower the error, the stronger the prediction. In Figures 6 and 7, it is seen that almost all the errors vary between the strong and weak unidimensional item pools. When the EAP and CI methods are used together, the errors do not show any change in terms of the item pool. Regardless of the item selection methods and ability estimation methods, it was concluded that, in the strong unidimensional item pool, the error values of SPRT increased. Another noteworthy finding of the study is that the errors show a decrease in the strong unidimensional item pool when the WLE and CI methods are used together. Accordingly, the most appropriate methods, in terms of errors, are EAP and CI together in the strong unidimensional item pool, whereas are EAP and SPRT together in the weak unidimensional item pool.

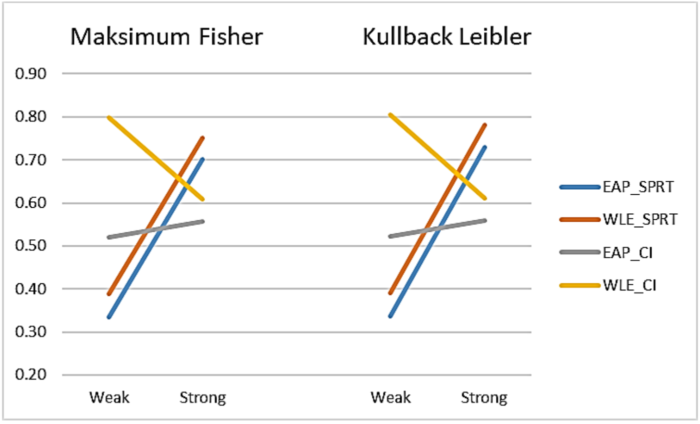

Figure 6. *Findings on RMSE*

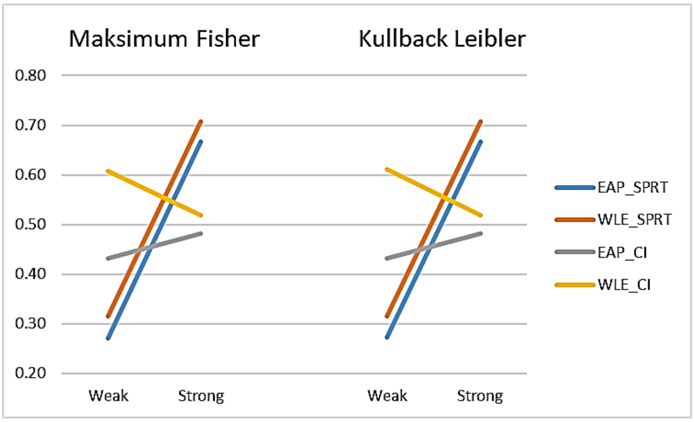

Figure 7. *Findings on absolute error*

DISCUSSION, CONCLUSION, RECOMMENDATIONS

In this research, we investigated how the weak and strong unidimensional item pools affect the CACT ending in terms of test length, classification accuracy, the correlation between the true theta and the estimated theta, RMSE and absolute error, which are the dependent variables of the study. In the line with the aim of the study, 16 conditions were composed of different item pools, item selection, ability estimation and classification methods. The conditions were compared to each other in terms of the dependent variable values by taking the mean over 25 replications. It is impossible to propose a direct method based on the research results and all the dependent variables. Therefore, the dependent variables are handled separately below.

A result of this study is that the classification accuracy (> 0.90) and the correlations between the true theta and estimated theta (> 0.80) were calculated at a very high level in all conditions. In terms of test length, it was concluded that the CI classification criterion was generally more useful than SPRT. These three results of this research coincide with those from the relevant literature (Ayan, 2018; Eggen & Straetmans, 2000; Gündeğer, 2017; Nydick et al., 2012; Thompson, 2011). From this viewpoint, if SPRT is to be used as a classification criterion, it may be recommended to test whether the item pool shows strong unidimensionality and to use SPRT if the item pool has this characteristic. If the item pool shows weak unidimensionality rather than strong, CI should be preferred over SPRT to ensure test effectiveness.

The focus of this research is the unidimensionality that is an assumption for item selection (MFI and KLI) and classification methods (SPRT and CI) and a characteristic for the item pool. A striking finding of the study, when the criteria are test length and accucary classification, is that SPRT was performed better in the strong unidimensional item pool than in the weak unidimensional item pool. In the strong one, SPRT reduced the number of items by half, which is an acceptable number (28) to end a test session. In other words, in a weak unidimensional item pool, SPRT requires approximately 49 items, but in strong one it needs only 28 items to end the CACT application, which shows us that the SPRT is useful when the item pool has strong dimensionality. However, it should not be ignored that as the strength of the unidimensionality increases, the bias of the true ability parameter estimates in SPRT increases noteworthily under both ability estimation methods. When the strength of dimensionality increases in the CI method, the difference between actual and estimated ability barely increases, and the test length becomes significantly shorter with high classification accuracy. Additionally, KLI and MFI item selection methods are not affected by the item pool characteristic as week or strong unidimensionality. Both item selection methods have the same pattern and the same result under all output criteria throughout the type of unidimensionality in themselves.

Many researchers do not provide information about the unidimensionality of the item pool they derived in MC or PH simulations, especially in CAT and CACT studies. However, to draw attention to this situation in this research, the item loads are fixed in a range so that they show weak and strong correlations and unidimensionality. When the literature is examined, it is seen that unfortunately, there is no detail about data generation. Besides, this type of information shows us the item pool characteristic, and these research results prove that the characteristic of the item pool has a significant impact on CACT, it is highly recommended that with the information on how many items the item pool consists of, the item factor loads, inter-item correlations and TIF should be presented in papers and interpreted the results taking these information into account. At this point, researchers and practitioners may be advised to further examine the item pool characteristic with factor analytic methods and/or to report item factor loads. Based on these study results, then it may be recommended to prefer the CI classification method regardless of the unidimensionality of the item pool because CI has always produced results with classification accuracy similar to SPRT but with a test length of almost half.

Another result of the study is that the ability estimation methods and the classification criteria produced errors at different levels in item pools shows strong and weak unidimensionality. It is another remarkable finding that the EAP estimation method outperforms in terms of both classification accuracy and ability parameter recovery under all conditions. It can be recommended to use CI with EAP in CACT studies, whether the item pool is weak or strong unidimensional, but WLE only under strong unidimensional item pools. Additionally, the EAP and SPRT methods are recommended to prefer in the weak unidimensional item pool. With the increase in the strength of the unidimensionality of the item pool in the study, the test length decreased by almost half, but the increase in the bias in ability parameter recovery is also remarkable. We didn't consider whether the cut-off point had an effect on this result. This was because the cut-off point was set as the mode of the test information function in this study. It is considered that the effect of the unidimensionality level of the item pool on Computerized Adaptive Classification Testing needs further investigation with different absolute cut-off points. The consistency between the results of this research and future studies on this subject can be examined.

REFERENCES

Ayan, C. (2018). *Comparing the psychometric features of traditional and computerized adaptive classification test in the cognitive diagnostic model* [Unpublished Doctoral Dissertation]. Ankara University.

Aybek, E. C. (2016). *An investigation of applicability of the self assessment inventory as a computerized adaptive test (CAT)* [Unpublished Doctoral Dissertation]. Ankara University.

Breslow, N. E., & Holubkov, R. (1997). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine, 16*, 103-116.

Cheng, P. E., & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement, 24*(3), 257–265.

Demir, S. (2019). *Investigation of classification accuracy at computerized adaptive classification tests* [Unpublished Doctoral Dissertation]. Hacettepe University.

Diao, Q., & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing* (pp. 1-13). UMI Research Press.

Dogan, N., Soysal, S., & Karaman, H. (2017). Aynı örnekleme açımlayıcı ve doğrulayıcı faktör analizi uygulanabilir mi?[Can exploratory and confirmatory factor analysis be conducted to the same sample?]. In Ö. Demirel & S. Dinçer (Eds.), *Küreselleşen dünyada eğitim* [Education in a globalizing World] (pp. 373- 400). Pegema Publishing.

Doğruöz, E. (2018). *Investigation of adaptive multistage test based on test assembly methods* [Unpublished Doctoral Dissertation]. Hacetteepe University.

Dooley, K. (2002), Simulation research methods. In J. Baum (Ed.), *Companion to organizations,* (pp. 829-848). Blackwell.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249-260.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713-734.

Erdem Kara, B. (2019). *The effect of item ratio indicating differential item functioning on computer adaptive and multi stage tests* [Unpublished Doctoral Dissertation]. Hacettepe University.

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*(4), 442-463.

Flaugher, R. (2000). Item Pools. In Wainer, H. (Ed.), *Computerized adaptive testing: A Primer* (pp. 37-59). Erlbaum.

Gökçe, S. (2012). *Comparison of linear and adaptive versions of the turkish pupil monitoring system (pms) mathematics assessment* [Unpublished Doctoral Dissertation]. Middle East Technical University.

Gündeğer, C. (2017). *A comparison of computerized adaptive classification test criteria in terms of classification accuracy and test length* [Unpublished Doctoral Dissertation]. Hacettepe University.

Gündeğer, C., & Doğan, N. (2018). The effects of item pool characteristics on test length and classification accuracy in computerized adaptive classification testings. *Hacettepe University Journal of Education, 33*(4), 888-896.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Kluwer Nijhoff Publishing.

Hsiehi, M. (2015). Examination of sequential probabılıty ratio tests in the setting of computerized classification tests: A simulation study. *International Journal of Innovative Management, Information & Production, 6*(2), 38-46.

Jiao, H. & Lau, A. C. (2003, April 2-24). *The effects of model misfit in computerized classification test* [Paper presentation]. The annual meeting of the National Council of Educational Measurement. Chicago, IL,USA

Kaçar, M. (2016). *Investigation of maximum fisher item selection method on computerized adaptive testing* [Unpublished Master Thesis]. Necmettin Erbakan, University.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability* [Unpublished Doctoral Dissertation]. Middle East Technical University.

Kezer, F. (2013). *Comparison of the computerized adaptive testing strategies* [Unpublished Doctoral Dissertation]. Ankara University.

Kezer, F. (2021). The effect of item pools of different strengths on the test results of computerized adaptive testing. *International Journal of Assessment Tools in Education, 8*(1), 145–155.

Lau, C. A. (1996). *Robustness of a unidimensional computerized testing mastery procedure with multidimensional testing data* [Unpublished Doctoral Dissertation]. University of Iowa.

Lau, C. A., & Wang, T. (1999, April 19-23). *Computerized classification testing under practical constraints with a polytomous model* [Paper presentation]. AERA Annual Meeting. Montreal, Canada.

Lin, C. J., & Spray, J. A. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test* (Research Report No. 2000-8). ACT.

Nydick, S. W. (2013). *Multidimensional mastery testing with CAT* [Unpublished Doctoral Dissertation]. University of Minnesota.

Nydick, S. W., Nozawa, Y., & Zhu, R. (2012, April 12-16). *Accuracy and efficiency in classifying examinees using computerized adaptive tests: an application to a large scale test* [Paper presentation]. The Annual Meeting of the National Council on Measurement in Education. Vancouver, Canada.

Özdemir, B. (2015). *Examining the effects of ıtem level dimensionality models on multidımensional computerized adaptive testing method* [Unpublished Doctoral Dissertation]. Hacettepe University.

Penfield, R. D., & Bergeron, J. M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Applied Psychological Measurement, 29*(3), 218–233.

R Core Team (2019). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). Academic Press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Seitz, N. N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling, 55*, 105-123.

Spray, J. A., Abdel-fatah, A. A., Huang, C.-Y., & Lau, C. A. (1997). *Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional* (Research Report No.97-5). ACT.

Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computer adaptive test* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, New Orleans, USA.

Spray, J. A., & Reckase, M. D. (1996, April 5-7). Comparison of SPRT and sequential bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*(4), 405-414.

Şahin, M. D. (2017). *Examining the results of multidimensional computerized adaptive testing applications in real and generated data sets* [Unpublished Doctoral Dissertation]. Hacettepe University.

Şenel, S. (2017). *Investigation of the compatibility of computerized adaptive testing on students with visually impaired* [Unpublished Doctoral Dissertation]. Ankara University.

Tao, J., Shi, N. Z., & Chang, H. H. (2012). Item-weighted likelihood method for ability estimation in tests composed of both dichotomous and polytomous items. *Journal of Educational and Behavioral Statistics, 37*(2), 298-315.

Thompson, N. A., & Ro, S. (2007). Computerized classification testing with composite hypotheses. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing* (pp. 1-13). UMI Research Press.

Thompson, N. A. (2007a). *A comparison of two methods of polytomous computerized classification testing for multiple cutscores* [Unpublished Doctoral Dissertation]. University of Minnesota.

Thompson, N. A. (2007b). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12*(1), 1-13.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778-793.

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical assessment. Research & Evaluation, 16*(4), 1-7.

Wang, T. (1997, March 24-28). *Essentially unbiased EAP estimates in computerized adaptive testing* [Paper Presentation]. American Educational Research Association Conference. Chicago, USA.

Wang, T., Hanson, B. A., & Lau, C. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23*(3), 263-278.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(2), 109-135.

Wang, S., & Wang, T. (2001). Precision of warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 317–331.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing* (pp. 1-13). UMI Research Press.

Yao, L. (2003). *SimuMIRT* [Computersoftware]. https://www.psychsoft.soe.vt.edu/report3.php?recordID=SimuMIRT

Yi, Q., Wang, T., & Ban, J. (2000). *Effects of Scale Transformation and Test Termination Rule on the Precision of Ability Estimates in CAT*. ACT Research Report Series, 2000-2.