International Journal of Assessment Tools in Education
2022, Vol. 9, No. 1, 118–137

https://doi.org/10.21449/ijate.909616

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Multidimensional Computerized Adaptive Testing Simulations in R

F. Gul Ince Araci[1,*], **Seref Tan**[1]

[1]Gazi University, Gazi Faculty of Education, Department of Measurement and Evaluation in Education, Ankara Turkey

ARTICLE HISTORY

Received: Apr. 04, 2021
Revised: Nov. 25, 2021
Accepted: Dec. 18, 2021

Keywords:
Computerized adaptive testing,
Multidimensional,
R,
MCAT,
Simulation.

Abstract: Computerized Adaptive Testing (CAT) is a beneficial test technique that decreases the number of items that need to be administered by taking items in accordance with individuals' own ability levels. After the CAT applications were constructed based on the unidimensional Item Response Theory (IRT), Multidimensional CAT (MCAT) applications have gained momentum with the improvement of multidimensional IRT (MIRT) models in recent years. Researchers often benefit from simulation studies in order to design the final adaptive testing application and to test the effectiveness of adaptive testing applications they developed with different methods. Recently, R has become one of the most widely used programming languages in Monte Carlo Simulation studies since it is a free and open-source software. The aims of this study are to present the MCAT simulation process step by step in the R environment, to examine the effects of the conditions that researchers can handle during the simulation process according to two different dimensional models, and to examine the effect of treating multidimensional structures as unidimensional structures on simulation results. In this direction, datasets generated in accordance with within-item dimensionality and between-item dimensionality models, MCAT simulation studies were constructed with different customizations, and MCAT simulation results were compared with unidimensional CAT simulation results. All commands required for each simulation example were explained and results were shared for each condition.

1. INTRODUCTION

The integration of computers and the internet into education has gained tremendous momentum through the development of information technology. Although most of the exams are still applied as a paper-and-pencil method starting from the primary education level, internet-based distance education and test applications are rapidly increasing. In paper-and-pencil exam applications, the items that each examinee is expected to answer, the number of items are the same. The connection between abilities of an individual and the properties of the items are not taken into consideration. In other words, item difficulty cannot be matched with the examinee's ability. However, in CAT applications, the individual encounters the properties of items determined according to his/her ability level during the application process. In this way, test applications can be conducted with fewer items tailored to the individual, shorter time, and higher reliability. Moreover, test results and feedback can be presented to the individual as soon as the test ends (Weiner, 1993, Segall, 2005; Weiss

*CONTACT: F. Gul INCE ARACI ✉ gulincegazi@gmail.com Gazi University, Gazi Faculty of Education, Department of Measurement and Evaluation in Education, Ankara, Turkiye.

e-ISSN: 2148-7456 /© **IJATE** 2022

& Gibbons, 2007, Lin 2012). In addition, cheating during the testing time in cognitive tests is significantly reduced in online exams with CAT applications by using different questions to individuals. In addition to its advantages for individuals, CAT has also enabled researchers to form their own tests with different methods. Researchers can perform a wide variety of applications by differentiating the IRT model, starting the test, choosing the item to be administered, estimating the ability parameter, and changing the test stopping rules. For instance, when designing a test in order to decide which IRT model is suitable, simulation applications can be used to decide how many items will be included in the test and what the item exposure rate will be. Simulation applications allow the comparison of CAT applications constructed under different methods and constraints (Thompson & Wise, 2011; Meneghetti & Junior, 2018). Researchers can run CAT simulations based on various datasets: Monte Carlo simulations by generating data, post-hoc simulations on the basis of parameters derived from real-time applications, or hybrid simulation by imputing missing values to the real-time applications (Nydick & Weiss, 2009).

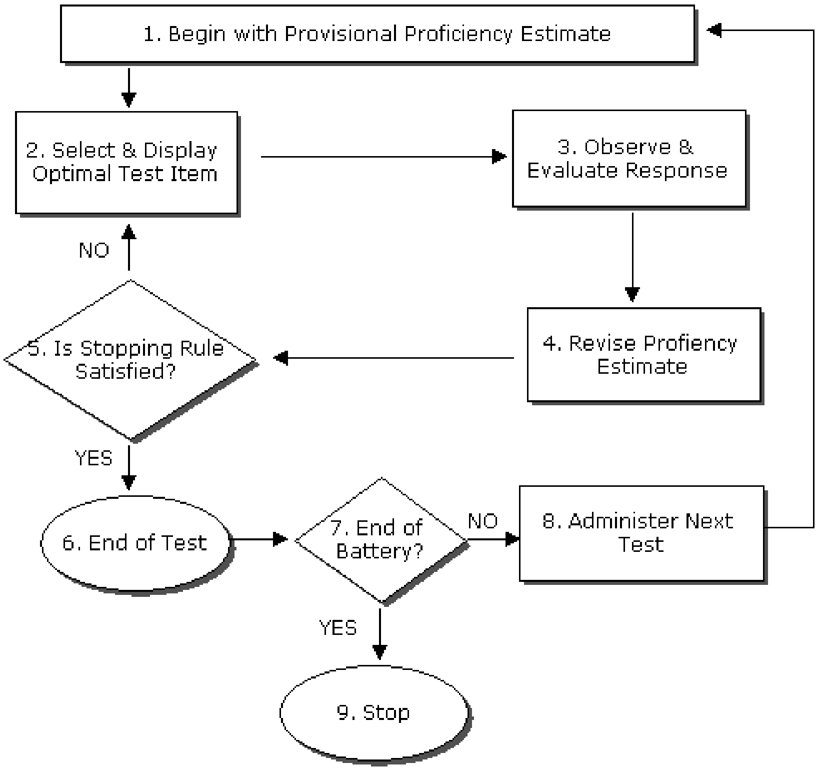
IRT models are frequently used in Monte Carlo Simulation studies in the field of psychometrics (Bulut & Sünbül, 2017). Most of the CAT studies performed in the literature are based on unidimensional IRT. Nevertheless, many psychological structures are multidimensional. Through MCAT applications developed using multidimensional IRT (MIRT) models in multidimensional structures, it is possible to decrease the number of items required to be administered to an individual to increase the precision of measurement and measure multiple traits at the same time (Seo & Weiss, 2015; Chalmers, 2016). In order to take these advantages of MCATs, there is a variety of software developed. One of the most popular software is R, which is a free and open-source platform. Real-time or simulation applications of MCATs, which researchers can customize by writing their own functions, can be performed on R (R Core Team, 2020). R allows researchers to customize their own applications by writing their own functions. The mirtCAT package (Chalmers, 2016) in R, which allows researchers to develop customized MCAT applications by writing their own code, consists of utile tools performing Monte Carlo simulations, and it is the only package that allows MCAT applications for now.

1.1. Computerized Adaptive Testing

Adaptive testing is an advanced test application where examinees encounter items according to their abilities, which is estimated based on the response pattern. Each individual completes a tailored test by preventing them from taking easy and difficult questions for them. Therefore, examinees encounter fewer items and save time (Embretson & Reise; 2000; Van der Linden, 2002).

In adaptive testing applications, the process starts with temporary θ estimation for an examinee. Frequently, the starting θ is considered to be 0. By presenting the first item in accordance with the starting rule to the examinee, the estimation is performed again, and the items are presented according to the item selection rule. This process continues until the stopping criterion is met. If the test does not consist of subscales, the application is stopped when the stopping criterion is met. If it consists of subscales, the other test starts, and the same procedures are repeated. The flowchart (Thissen & Mislevy, 2000), which represents the application process of adaptive testing, is presented in Figure 1.

Figure 1. *A flowchart showing the adaptive testing process**


1. Begin with Provisional Proficiency Estimate
2. Select & Display Optimal Test Item
3. Observe & Evaluate Response
4. Revise Profiency Estimate
5. Is Stopping Rule Satisfied?
NO
YES
6. End of Test
7. End of Battery?
NO
8. Administer Next Test
YES
9. Stop

*: From Thissen, D. and Mislevy, R. J., 2000. Testing Algorithms. In H. Wainer (Eds.). *Computerized Adaptive Testing: A Primer* (p.106). Lawrence Erlbaum Assc. Copyright 2000 by Lawrence Erlbaum Associates, Publishers."

Adaptive tests can be developed as a unidimensional CAT and MCAT according to the dimension of the measured construct. Six components are required to carry out an MCAT application (Weiss & Kingsbury, 1984; Thompson & Weiss, 2011; Chalmers, 2016):

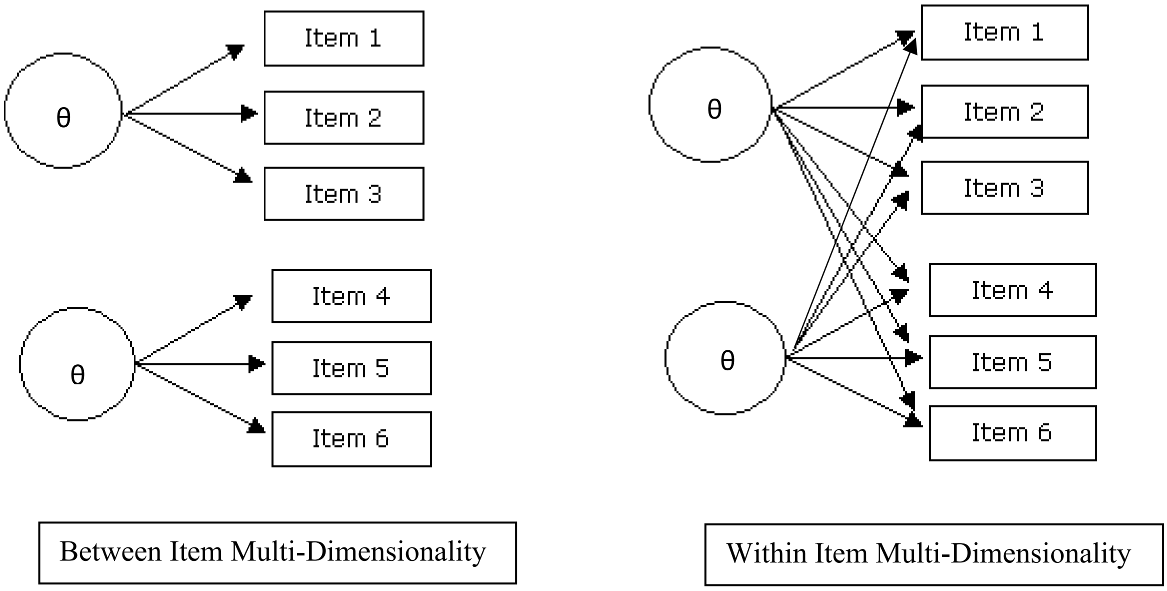
1. Multidimensional Item Response Theory (MIRT) model,
2. Calibrated item pool,
3. Starting rule,
4. Item selection method,
5. Estimation method,
6. Stopping rule.

These components are briefly explained below:

***MIRT model*:** The interaction between examinees and test items may not always be unidimensional because, while answering the test item, the individual may need to use more than one ability or skill field, so MIRT models may be required in complex structures (Bock & Aitkin, 1981, Reckase, 2009, Chen, 2012). MIRT models are divided into two according to item level: within-item dimensionality and between-item dimensionality models. In the within-item dimensionality model, items load to all dimensions, and in the between-item dimensionality model, each item loads to a specific dimension as shown in Figure 2 (Wang & Chen, 2004).

One of the crucial steps in implementing adaptive testing applications is determining the model to be used in item bank calibration, ability estimation, and item selection methods (Magis et al., 2017). The methods to be used vary in accordance with the dimensionality of the model and whether the items have dichotomous or polytomous scored response categories (Weiss & Kingsbury, 1984, Ackerman, 1991; Wang & Chen, 2004, De Ayala, 2009). Therefore, it should be decided on the MIRT model to be used before MCATs are formed.

Figure 2. *Between and Within Item Multi-Dimensionality (Wang & Chen, 2004).*



Calibrated item pool: An item pool consisting of many quality items to be used during MCAT application should be created. There should be a sufficient number of items in the item pool suitable for individuals with different levels of the measured attribute. According to Stocking (1994), the item pool should have at least six times as many items as the test length. And, Reckase (2003) stated that a pool of approximately 200 items is appropriate for examinees sampled from a standard normal distribution. For CAT applications to work efficiently, it is essential to have sufficient quantity and quality of items. On the other hand, a high number of items will not be sufficient by itself. Many researchers have stated that the distribution of item parameters, content weighting, and item exposure rates should also be taken into consideration while developing the pool (He & Reckase, 2013).

Starting Rule: At this stage of MCAT application, it is usually required to define the initial estimation of the latent trait and the hyper-parameter distributions. Hyper-parameters are parameters obtained from the preliminary distribution without the real dataset observed. Suppose the first item administered to the examinee during the application is not determined specifically. In that case, the latent trait's initial estimation is used to determine the first item to be selected. Hyper-parameter distributions are used as a component in the item selection method and they provide prior distribution information while updating the latent trait estimation after the individual responds to each item during the MCAT application. MCATs can also be started with the methods of starting from the first item, starting according to the item selection method, and assigning the average ability level of the population as the initial theta. When the initial value of the examinees' latent trait estimates is unknown, it is common to assume it 0 (Thompson, 2007; Riggelsen, 2008; Chalmers, 2016).

Item selection method: During the MCAT application, after the examinee encounters the first item and estimates the ability parameters, the item selection method should be determined for determining which item will be presented next. Item selection methods are usually based on the idea of maximizing information about an examinee's location on the θ-coordinate or minimizing the error in the location estimation The basis of all item selection methods is maximizing or minimizing some criterion values in the final θ estimation. What makes these methods different from each other is the definition of the criterion (Reckase, 2009). While there are many item selection methods for CAT applications in the literature, there is a limited number of item selection methods available for MCAT applications. Some of these criteria are as follows: A-rule, E-rule, D-rule, T-rule, W-rule, Kullback-Leibler Information Criteria (KL), and Continuous Entropy method. While the basis of A-rule method is minimizing trace of the asymptotic co-variance matrix, the basis of E-rule is minimizing the information matrix and Wrule method based on maximizing the weighted information criteria. D-rule method is based on maximizing the determinant of the information matrix and T-rule based on maximizing the trace of the asymptotic covariance matrix. And other methods KL and Continious Entropy method, maximize posterior expected KL information and minimize the expected continuous entropy, respectively. And other methods KL and CL are based on maximizing the posterior expected KL information and minimizing the expected continuous entropy, respectively. (Veerkamp & Berger, 1997; Segall, 2001; Wang & Chang, 2004; Mulder & van der Linden, 2009; Wang & Chang 2011).

Estimation method: The estimation method for calculating the examinees' latent trait parameters should be selected. The Maximum Likelihood Method (MLE) (Lord and Novick, 1968), EAP and MAP (Segall, 1996) are the most frequently used methods. However, if all the answers are correct or incorrect, EAP and MAP methods are suggested to make estimations with low standard errors (Hambleton and Swaminathon, 1985). In addition to these methods, the weighted MLE (MWLE) method was revealed by Wang (2015) for multidimensional tests, which provides robust estimations

Stopping rule: At this stage of the CAT application, the stopping rule of testing should be determined. In CAT applications, stopping rules may be used in accordance with the fixed test length, standard error, change in the amount of the latent trait estimation, or the fixed application time. When the stopping rule is determined according to the fixed test length, erroneous results can be obtained in CAT applications for the examinees who are at the end of the skill distribution (Finkelman et al., 2009). For this reason, researchers can stop the CATs when a standard error level they previously specified is reached, and they obtain measurements that are more precisely. When specifying the standard error, if the test used is multidimensional, each dimension can be terminated with the same or different standard error values (Chalmers, 2016). If the standard error value is not determined for each dimension by customizing the codes to be used, the application stops in accordance with the standard error value specified for the first dimension and estimates according to the different standard error values for the other dimensions. Therefore, if the standard error-based stopping rule will be used in MCAT applications, it is essential to add the standard error value for each dimension to the code to be run.

1.2. Monte Carlo Simulations

Monte Carlo simulations have a crucial role in studies in the field of psychometrics. Within the scope of their studies, researchers may not be able to access empirical data or may not prefer to test applications for data collection purposes. One reason for the simulation requirement is that collecting empirical data can be time-consuming and costly when the number of items used in studies is long, and the number of examinees to be applied is high. In some studies, there are losses in the empirical data collected, and this loss of data affects the results of the analysis. Another, probably the most important, reason is that the working conditions to be examined cannot be obtained with real-time applications (Davey et al., 1997; Feinberg and Rubright, 2016). But this approach also has several limitations. Firstly, how realistic the conditions modeled in Monte Carlo simulation studies are affects the usefulness of the results. In this respect, the modeled conditions (e.g., assumed distribution of the parameters) should be defensible in terms of reality. Another limitation is that it is difficult to assess the quality of the random number generator in Monte Carlo simulations (Stone, 1993). Post-hoc and hybrid simulations can be done as a solution to the concern that the results obtained from Monte Carlo simulations cannot be generalized to real test applications. In post-hoc simulations, real item-response vectors obtained from paper-and-pencil test or adaptive test are used instead of generated item responses. In hybrid simulations, simulations are performed after missing value imputation based on empiricalreal data set (Thompson & Weiss, 2011). However, it is not easy to obtain real answers due to time and cost. Monte Carlo studies allow modeling realistic data conditions and can be used in competitor statistical comparisons that cannot be made with empirical data (Harvell et al., 1996). When there are a large number of conditions to manipulate, Monte Carlo simulations are preferred. Because, Monte Carlo simulations provide researchers with the opportunity to test a large number of models in a short time, which are hard to test in real life.

Monte Carlo simulation studies can be carried out by researchers in order to examine the applicability in CAT applications and to make an application plan. Post-hoc or hybrid simulation studies are preferred to determine the final application conditions (Thompson and Weiss, 2011). Monte Carlo and post-hoc simulations are frequently used in CAT applications performed in the literature. The two most crucial variables of Monte Carlo simulations are average test length and precision of the test scores. In traditional tests, the number of items in the test is constant, and the precision is variable. The number of items administered to examinees in adaptive tests is usually the variable, but it can be designed to provide equal precision to each examinee. In this regard, simulation studies are essential (Thompson & Weiss, 2011).

1.3. R Statistical Programming Environment

As a result of the development of computer technology, there are some commercial and open-source softwares that can carry out CAT simulations. Some of these softwares are CATSim, SimulCAT, SimuMCAT, Firestar software, and R software environment. CATSim is presented as a commercial product, and other software is presented as open-source access (Aybek, 2016). While unidimensional CAT simulations are possible with CATSim, SimulCAT, and Firestar software, MCATs simulations are possible with the SimuMCAT and the mirtCAT package in the R software environment.

The R programming language, which has been widely used in academic studies in recent years, is a programming language developed with the contributions of researchers from different parts of the world since 1997 (Hornik, 2020). The use of R has increased rapidly due to its open source code. R programming language offers the opportunity to be used in many fields such as statistics, data mining, machine learning and simulation applications. The R statistical programming environment (R Core Team, 2020) enables the opportunity to conduct simulation studies free of charge. Researchers who ask for generating data in R may generate data in accordance with a different probability distribution (normal, log-normal, uniform, etc.). There is a root name setting out each distribution, and usually, four functions are defined for each. Each distribution's commands begin with a letter to indicate functionality:

p: cumulative distribution function,

q: quantile function,

d: density function,

r: randomly generated numbers.

For instance, for log-normal distribution, rlnorm (the multivariate lognormal distribution), plnorm (the log normal cumulative distribution), dlnorm (the log normal probability density) and qlnorm (the log normal quantile) functions can be defined. Random data is generated for the rlnorm function according to the log-normal distribution. The qlnorm function sets the quantile of the log-normal distribution at a given cumulative density. Normal, log-normal, and uniform distributions are frequently used in studies where data generation is performed based upon IRT.

In this study, the steps of MCAT simulations according to within-item and between-item dimensionality models with the mirtCAT (version: 1.10) package in the RStudio (version: 1.3.1073) software environment will be demonstrated in terms of ease of use and prevalence.

After all required components are prepared, the function that starts MCAT simulations with the mirtCAT package is the mirtCAT() function. It is essential to introduce the item and individual parameters, IRT model, inter-dimensional correlations, starting rule of the test, item selection criteria and stopping rule to perform a multidimensional simulation with this function. The functions that are basically required to perform MCAT simulation with mirtCAT package are described in Table1, below (Chalmers, 2016).

Table 1. *Some functions to perform MCAT simulation with mirtCAT package*

mo: It is used in the model definition phase. The model defined in the mirt package is drawn to mirtCAT with this function. This object is required if test items are to be scored.

generate.mirt_object: It is the function used to form a mirt object from known population parameters and transfer it to mirtCAT.

method: It is used to determine the parameter estimation method. "EAP", "MAP", "ML", "WLE", "EAPsum", "fixed" are the methods that can be selected.

criteria: It is the function for determining the method of item selection. "seq", "random", "MI", "MEPV", "MLWI", "MPWI", "MEI", "IKL", "IKLP", "IKLn", "IKLPn", "Drule", "DPrule", "Erule", "EPrule", "Trule", "TPrule", "Arule", "APrule", "Wrule", "WPrule", "KL", "KLn".

start_item: It is the function by which the starting rule of MCAT application will be determined.

A MCAT design can be customized using different MIRT models, different item selection rules, different estimation methods etc. Since the methods to be used in a design will affect the measurement result, it is important to determine the most effective methods according to the application purpose. Besides these, interdimensional correlations and the dimensional structures are important issues for CATs (Su, 2016). Because interdimensional correlation can change the dimensionality of the structure and this in turn can change the MCAT implementation to be carried out.

1.4. Purposes

The purposes of this study are to present how MCAT designs can be generated and executed through Monte Carlo simulations in R environment; to show the effect of simulation conditions, which can be considered according to different dimensionality models, on the simulation results, and to investigate the effect of treating multidimensional structures as unidimensional structures. For the purposes, different Monte Carlo simulation studies were presented and the steps of simulations were demonstrated.

1.5. Research Questions

In line with the research purposes, answers to the following questions were sought:

1. How is an MCAT simulation designed according to the within-item dimensionality model affected by different item selection methods, ability parameter estimation methods and interdimensional correlations?
2. How is an MCAT simulation designed according to the between-item dimensionality model affected by different item selection methods, ability parameter estimation methods and interdimensional correlations?
3. How does treating each dimension of the multidimensional structure as a single dimension affect the simulation results?

2. METHOD

In this study, three different Monte Carlo simulation studies were carried out and all simulation steps were presented. In all three studies, data generation, Monte Carlo simulation steps and findings obtained as a result of analysis are presented. In order to answer the first research question, Simulation Study 1 is carried out. In this example, MCAT simulations were conducted according to the within-item dimensionality model with different conditions. Different conditions that could affect the precision of MCATs were considered: (1) interdimensional correlation, (2) item selection method, (3) parameter estimation method. In order to answer the second question, MCAT simulations conducted according to the between-item dimensionality model are carried out and the same conditions in the first study were examined in Simulation Study 2. Lastly, in order to seek an answer to the third questions, Simulation Study 3 is carried out. In this study, Unidimensional CAT (UCAT) simulations were conducted using the item and ability parameters of the multidimensional structure. MCAT and UCAT simulations performed with the data produced according to the between-item dimensionality model were compared.In all studies, RMSE, bias and r(θi, $\widehat{\theta}$ j) values obtained from all simulations were examined. R was used in order to complete simulation steps. The presented steps are completed in R for Windows 4.0.2. Simulation study examples of MCAT applications were performed on the mirtCAT (Chalmers, 2016) package.

2.1. Simulation Study 1: The Within-item Dimensionality Model

In the first simulation study, the within-item dimensionality model was handled. A simple non-customized MCAT simulation example is presented. Item selection methods, ability parameter estimation methods and the interdimensional correlations were examined as changing simulation conditions. In the first step of the MCAT simulation, packages to be used on the R platform should be downloaded.

```
# Install required packages
install.packages("mirt")
install.packages("mirtCAT")
install.packages("mvtnorm")
install.packages("plyr")
install.packages("SimDesign")
```

After the downloading process is completed, the required packages should be activated. Before the analyses are carried out, the `set.seed()` command ensures that the outputs of the application are reproducible. Any number can be written in parentheses, and the same results are obtained when `set.seed()` is run with the same number.

```
# Load packages into the current session
library(mirt)
library(mirtCAT)
library(mvtnorm)
library(plyr)
library(SimDesign)
# Set the seed for reproducible results
set.seed(1111)
```

After the packages were drawn and activated, item and ability parameters were generated. In accordance with the within-item dimensionality model, 2-dimensional MCAT simulations were carried out for 1000 examinees. For this example, parameters for a multidimensional test consisting of 300 dichotomous items and 2 dimensions were generated. The *a* parameters were drawn from the log normal distribution (*a* ~ *lnN* (.0, .3)), and item intercept parameters were drawn from the uniform distribution (*d~ U* (-2, 2)). After slopes and intercepts were generated, they were combined in a single object (parameters) with the `data.frame` function.

```
#Generate Multidimensional IRT parameters
testlength <- 300 # Bank size
N <- 1000 # Sample size
a <- matrix(rlnorm(testlength*2,.0,.3),testlength) # Generate item
discrimination parameters
d <- runif(n = testlength, -2, 2) # Generate intercepts
parameters <- data.frame(a, d) # Combine parameters in a single dataset
colnames(parameters) <- c('a1', 'a2', 'd') # Name the columns
```

In the next step, the variance-covariance matrix of the ability parameters for two-dimensional structure is demonstrated on a matrix. Ability parameters were drawn from the multivariate normal distribution ((θ~*(0,*$\boldsymbol{\Sigma}$)) depending on the defined correlations. For the two-dimensional structure, the inter-dimensional correlation was determined as 0.3, 0.6 and 0.9, and parameters generated with `rmvnorm()` function.

```
#Set intercorrelations between latent traits
latent_cov <- matrix(c(1, r, r, 1), 2, 2)

#Generate multidimensional theta parameters
thetas <- rmvnorm(N, sigma = latent_cov)
```

Then the `mod` object required for MCAT Simulation was formed. This object is used while generating the response pattern and creating the MCAT design. The `generate_pattern ()` function was used to generate the response pattern.

```
# Create mirt_object
mod <- generate.mirt_object(parameters, itemtype = '2PL', latent_covariance
= latent_cov)

#Generate response data
responsepattern <- generate_pattern(mo = mod, Theta = thetas)
```

In the next step, required components were specified for the MCAT simulation to be conducted. These components were defined by the function `design()` and `mirtCAT()`. In these definitions, SE ≤0.4 was specified as stopping rule. Five item selection methods were examined for two dimensional complex model using two estimation methods. By using Arule, Drule, Trule, Wrule and KLn (Kullback-Leibler item selection method with root-N adjustment) item selection methods, estimations were made according to both EAP and MAP methods. The item starting rule for each condition is the same as for the item selection method. 5x2x3 (30) simulation application including stopping criterion, estimation method and correlation was carried out.

```
# Run the MCAT simulations with mirtCAT function and store results
design <- list(min_SEM = 0.4)
mcat1 <- mirtCAT(mo = mod, local_pattern = responsepattern, method = ' ',
start_item = " ", criteria = " ", design = design)
```

When the application is completed, the results can be reached by running the `mcat1` object where the results are saved. The mirtCAT package presents the avarage number of items administered, the ability parameter and true ability parameter for each dimension, and the standard errors of these parameters in the output of the simulation. If researchers ask for examining the test efficiency or the effect of different MCAT simulation designs on test efficiency; bias and

RMSE values can be computed. In this situation, firstly, an object should be formed in which theta estimates of both dimensions are listed. Theta estimates are collected in a single object with the `laply()` function that can be computed using the `bias()` and `RMSE()` commands of the SimDesign (Chalmers et al., 2020) package, with the following code:

```
#Show average number of items answered, theta estimations, bias and RMSE
itemsanswered <- laply(mcat1, function(x) length(x$items_answered))
mean(itemsanswered)
estimation1 <- laply(mcat1, function(x) x$thetas[1])
estimation2 <- laply(mcat1, function(x) x$thetas[2])
bias(thetas[,1], estimation1) # Compute bias
bias(thetas[,2], estimation2)
RMSE(thetas[,1], estimation1) # Compute root mean square error
RMSE(thetas[,2], estimation2)
cor(thetas[,1], estimation1)
cor(thetas[,2], estimation2)
```

According to the simulation outputs, the average number of items answered, bias, RMSE and the correlation between estimated θ and true $\hat{\theta}$ (r(θ i, $\hat{\theta}$ j)) obtained for both dimensions are presented. The conditions with interdimensional correlations of 0.3, 0.6, and 0.9 are presented in Table 2, Table 3 and Table 4, respectively.

Table 2. *Statistics from MCAT when interdimensional correlation is 0.3*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	r_1(θi, $\hat{\theta}$ j)	r_2(θi, $\hat{\theta}$ j)
Arule	EAP	67.057	-0.004	-0.013	0.374	0.398	0.926	0.917
	MAP	63.848	-0.008	-0.013	0.402	0.373	0.926	0.915
Drule	EAP	72.517	-0.02	0.002	0.373	0.398	0.927	0.917
	MAP	69.56	-0.05	-0.001	0.376	0.403	0.925	0.915
Trule	EAP	94.364	-0.017	0.000	0.355	0.396	0.934	0.917
	MAP	91.238	-0.018	0.001	0.359	0.399	0.932	0.916
Wrule	EAP	100.424	-0.016	0.004	0.357	0.392	0.933	0.919
	MAP	97.532	-0.016	0.003	0.359	0.395	0.932	0.918
KLn	EAP	103.496	-0.017	0.009	0.359	0.393	0.932	0.919
	MAP	100.6	-0.02	0.005	0.366	0.397	0.929	0.917

Note. MTL (Mean test length) represents the average number of items administered.

When the correlation between dimensions is 0.3, the MCAT application resulting in the least average number of items was performed with the MAP estimation method and the ARule stopping rule. However, when the Arule stopping rule was used, the RMSE value obtained for the first dimension was not below 0.40. The simulation application that resulted in the highest number of items was carried out with the KLn method. All MCAT applications ended with fewer items with the MAP estimation method. Correlation between estimated θ and true $\hat{\theta}$ calculations was high and similar in all conditions. All calculated bias values were negligible.

As seen in Table 3, the increase in interdimensional correlation decreased the number of items required to terminate the MCAT application. The simulation that resulted in the least number of items was carried out with the Arule stopping rule and the MAP estimation method. All applications performed with the MAP estimation method ended with fewer items than the applications performed with EAP. The RMSE value for the second dimension was not below 0.4

for Arule, Drule, Trule and Wrule methods. Only, the RMSE value obtained for both dimensions with the KLn method fell below 0.4. It should be noted that the KLn is the only method that can be used in common in unidimensional and multidimensional CAT applications. Correlation between estimated θ and true $\hat{\theta}$ calculations and bias values were similar. The results obtained under the condition that the interdimensional correlation is 0.9 are presented in Table 4.

Table 3. *Statistics from MCAT when interdimensional correlation is 0.6*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \hat{\theta} j)$	$r_2(\theta i, \hat{\theta} j)$
Arule	EAP	45.006	0.025	-0.024	0.375	0.41	0.928	0.918
	MAP	42.485	0.024	-0.023	0.385	0.414	0.924	0.916
Drule	EAP	46.05	0.024	-0.022	0.368	0.412	0.931	0.917
	MAP	43.839	0.017	-0.025	0.374	0.420	0.928	0.914
Trule	EAP	66.181	0.027	0.027	0.354	0.401	0.936	0.921
	MAP	58.440	0.026	-0.011	0.357	0.402	0.935	0.921
Wrule	EAP	64.706	0.023	-0.010	0.354	0.398	0.936	0.922
	MAP	63.014	0.023	-0.012	0.355	0.400	0.936	0.922
KLn	EAP	66.333	0.028	-0.016	0.352	0.393	0.937	0.925
	MAP	64.840	0.024	-0.019	0.352	0.396	0.937	0.924

Table 4. *Statistics from MCAT when interdimensional correlation is 0.9*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \hat{\theta} j)$	$r_2(\theta i, \hat{\theta} j)$
Arule	EAP	12.832	0.019	0.009	0.375	0.386	0.932	0.926
	MAP	11.119	0.028	0.013	0.402	0.406	0.922	0.918
Drule	EAP	11.216	0.015	0.004	0.370	0.386	0.934	0.926
	MAP	9.820	0.022	0.01	0.390	0.401	0.927	0.92
Trule	EAP	10.849	0.022	0.008	0.356	0.380	0.940	0.928
	MAP	9.219	-0.018	0.035	0.383	0.398	0.931	0.922
Wrule	EAP	10.401	0.019	0.010	0.371	0.386	0.934	0.926
	MAP	8.975	0.031	0.022	0.388	0.388	0.929	0.926
KLn	EAP	10.84	0.026	0.018	0.367	0.376	0.935	0.930
	MAP	9.645	0.022	0.014	0.377	0.397	0.933	0.922

When the interdimensional correlation was 0.9, the number of items required to complete the MCAT simulation was greatly reduced. The result of the application that ends with the least number of items was obtained with the Trule stopping criterion and the MAP estimation method. When Arule and Drule stopping methods are used with MAP parameter estimation method, the RMSE value for the second dimension was not below 0.4. In line with the simulation results, it was observed that the Arule and Drule methods gave similar results in all conditions. However, since they finished the application with fewer items, it was observed that although they provided the desired RMSE value in the first dimension, they could not provide in the second dimension. Simulations performed using the MAP estimation method in all conditions resulted in fewer items than EAP. As the interdimensional correlation increased and the structure approached unidimensionality, the methods gave results closer to each other and in all conditions KLn provided the desired stopping rule for both dimensions. In all conditions, $r(\theta i, \hat{\theta} j)$ obtained for both dimensions was high and similar. Bias for all dimensions was negligible.

2.2. Simulation Study 2: The Between-item Dimensionality Model

In simulation study 1, MCAT simulations were performed with the stopping rule not customized. In simulation study 2, MCAT simulations were performed according to the between item dimensionality model by customizing the stopping rule for each dimension. Interdimensional correlation values and stopping methods were used the same as Study 1. Due to the fact that the necessary packages are loaded in the first example, the packages will be activated directly in this example. The packages required for this study are called via commands written to the console. In order to MCAT Simulation results to be reproducible, the `set.seed()` command is used.

```
library(mirt)
library(mirtCAT)
library(mvtnorm)
library(plyr)
library(SimDesign)
set.seed(2222)
```

According to the between-item dimensionality model for the MCAT simulation, parameters for a multidimensional test consisting of 600 polytomous items and 2 dimensions were generated. Item parameters were generated for polytomous items with four categories and Multidimensional Graded Response Model (MGRM) was chosen as the MIRT model. The item parameters are distributed in the same way as in the study of Jiang, Wang, and Weiss (2016). The *a* parameters were drawn from the uniform normal distribution ($a \sim U\,(1.1,\, 2.8)$). First category boundary parameter (d_1) were drawn randomly from the uniform distribution (d_1~ (0.67, 2)), second category boundary parameter from (d_1~ (-0.67, 2-0.67)) and third category boundary parameter from (d_1~ (-0.67, -2)). Thus, all item bounce parameters ranged from [-2,2]. After that, the generated parameters were combined in a single dataset.

```
Generate Multidimensional IRT parameters
testlength <- 600 # Bank size
N <- 1000 # Sample size
# Generate  parameters
itemnames <- paste0("Item.", 1: testlength)
a <- matrix(runif(testlength *2, 1.1, 2.8), testlength)
a[1:300, 2] <- a[301:600, 1] <- 0
d1 <- runif(n = 600, min = 0.67, max = 2) # Generate first category boundary
parameter
d2 <- d1 - runif(n = 600, min = 0.67, max = 1.34)
d3 <- d2 - runif(n = 600, min = 0.67, max = 1.34)
d <- as.matrix(cbind(d1, d2, d3), ncol = 3)
parameters <- data.frame(a, d) # Combine parameters in a single dataset
colnames(parameters) <- c('a1', 'a2', paste0('d', 1:3))
```

In the next step, the variance-covariance matrix of the ability parameters for the two-dimensional structure is demonstrated on a matrix (cov). For the two-dimensional structure, interdimensional correlations were determined as 0.3, 0.6 and 0.9 between all dimensions. Ability parameters were drawn from the multivariate normal distribution ($\theta \sim (0, \boldsymbol{\Sigma})$) depending on the defined correlations. Then the mod object was created and the response pattern was generated using this object.

```
#Set intercorrelations between latent traits
latent_cov <- matrix(c(1, r, r, 1), 2, 2)
#Generate theta parameters for 2 dimensions
thetas <- rmvnorm(N, sigma = latent_cov)
#Create mirt_object
mod <- generate.mirt_object(parameters, itemtype = 'graded', latent_covari-
ance = cov)
#Generate response pattern
responsepattern <- generate_pattern(mo = mod, Theta = thetas)
```

In the next stage, unlike the first example, the minimum SE values were determined for each dimension by customizing the commands. The simulation was stopped on the condition that each dimension had a minimum SE value below 0.4 using `customNextItem()`, `extract.mirtCAT()` and `findNextItem ()` functions. In this regard, each dimension is considered as a block.

As item selection criteria, Arule, Drule, Trule, Wrule and KLn methods were used. EAP and MAP estimation methods were used as in the first example. The stopping rules of the application was determined by the `customNextItem()` function, the item selection method is defined by the `findNextItem()` function. A total of 30 simulations including the stopping rule (5), the estimation method (2) and the correlation value (3) were carried out.

```
customNextItem <- function(design, person, test){
browser()
      }
customNextItem <- function(design, person, test){
block1 <- 1:300
block2 <- 301:600
#Stop when the SE value falls below 0.4.
total <- sum(!is.na(extract.mirtCAT(person, 'items_answered')))
if(total< 300 && extract.mirtCAT(person, 'thetas_SE')[1] >= 0.4){
block <- block1
} else if(total < 600 && extract.mirtCAT(person, 'thetas_SE')[2] >= 0.4){
block <- block2
} else return(NA)
ret <- findNextItem(person=person, design=design, test=test, subset=block,
criteria = '')
ret
}
```

In the last step, simulation design was constructed with `mirtCAT()` function. Average number of items answered, bias, RMSE and the correlation between estimated θ and true $\widehat{\theta}$ values calculated with the commands presented below.

```
mcat2 <- mirtCAT(mo = mod, local_pattern = responsepattern, method = ' ',
start_item = " ", criteria = " ",
design = list(customNextItem=customNextItem))

#Show average number of items answered, bias, RMSE and r(θ i, ,θ .j).
itemsanswered <- laply(mcat2, function(x) length(x$items_answered))
mean(itemsanswered)
estimation1 <- laply(mcat2, function(x) x$thetas[1])
```

```
estimation2 <- laply(mcat2, function(x) x$thetas[2])
bias(thetas[,1], estimation1) # Compute bias
bias(thetas[,2], estimation2)
RMSE(thetas[,1], estimation1) # Compute root mean square error
RMSE(thetas[,2], estimation2)
cor(thetas[,1], estimation1)
cor(thetas[,1], estimation1)
```

The average number of items administered, bias, RMSE and the correlation between estimated θ and true $\hat{\theta}$ (r(θ i, $\hat{\theta}$ j)) obtained for both dimensions are as follows. The values calculated when the correlation is 0.3 are presented in Table 5.

Table 5. *Statistics from MCAT when interdimensional correlation is 0.3*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \hat{\theta} j)$	$r_2(\theta i, \hat{\theta} j)$
Arule	EAP	12.035	-0.02	0.011	0.359	0.356	0.938	0.934
	MAP	9.986	-0.013	0.014	0.388	0.397	0.928	0.918
Drule	EAP	16.27	0.001	0.002	0.386	0.353	0.927	0.935
	MAP	13.887	0.022	0.014	0.403	0.398	0.922	0.918
Trule	EAP	12.035	-0.02	0.011	0.359	0.356	0.938	0.934
	MAP	9.986	-0.013	0.013	0.388	0.397	0.928	0.918
Wrule	EAP	12.035	-0.020	0.011	0.359	0.356	0.938	0.934
	MAP	9.986	-0.013	0.014	0.388	0.397	0.928	0.918
KLn	EAP	12.101	0.006	0.010	0.370	0.354	0.933	0.935
	MAP	10.240	0.006	0.002	0.390	0.398	0.927	0.918

When the interdimensional correlation for the between-item dimensionality model is 0.3, the number of items required to finish the simulation is similar for the conditions. However, when the Drule method was used as stopping rule, average number of items administered were higher compared to other methods. The condition with the highest number of items administered is the condition in which the Drule stopping rule and EAP estimation method are used. Under the condition that the Drule stopping rule and EAP estimation method are used, the RMSE value obtained for the first dimension is more than 0.4. The results obtained from simulations using the Trule and Wrule stopping rules are the same. All simulations ended with fewer items with the MAP estimation method. The calculated bias and r_1(θ i, $\hat{\theta}$ j) values are similar for all conditions. The values calculated for the condition that the interdimensional correlation is 0.6 are presented in Table 6.

Table 6. *Statistics from MCAT when interdimensional correlation is 0.6*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \hat{\theta} j)$	$r_2(\theta i, \hat{\theta} j)$
Arule	EAP	11.52	-0.012	0.002	0.35	0.355	0.94	0.934
	MAP	9.661	-0.007	0.005	0.375	0.396	0.931	0.919
Drule	EAP	15.686	-0.003	-0.009	0.374	0.356	0.931	0.934
	MAP	13.643	0.018	0.009	0.389	0.394	0.926	0.919
Trule	EAP	11.52	-0.012	0.002	0.350	0.355	0.940	0.934
	MAP	9.661	-0.007	0.005	0.375	0.396	0.931	0.919
Wrule	EAP	11.52	-0.012	0.002	0.350	0.355	0.94	0.934
	MAP	9.661	-0.007	0.005	0.375	0.396	0.931	0.919
KLn	EAP	11.648	-0.01	-0.008	0.349	0.367	0.940	0.929
	MAP	10.015	0.003	-0.002	0.373	0.399	0.932	0.917

Under the condition that the interdimensional correlation is 0.6, the average number of items administered is fewer than the correlation is 0.3. The average number of items obtained from the simulation application performed with the Drule stopping rule is higher than other methods. The same results were obtained with Arule, Trule and Wrule methods. Bias and $r_1(\theta\ i, \hat{\theta}\ j)$ values are very close to each other for all conditions.

Finally, for the between-item dimensionality model, the values calculated according to the condition that the interdimensional correlation is 0.9 are presented in the Table 7.

Table 7. *Statistics from MCAT when interdimensional correlation is 0.9*

Item Selection	Estimation	MTL	Bias1	Bias2	RMSE1	RMSE2	$r_1(\theta i, \hat{\theta} j)$	$r_2(\theta i, \hat{\theta} j)$
Arule	EAP	9.302	0.002	-0.004	0.316	0.363	0.950	0.931
	MAP	7.793	0.010	0.006	0.354	0.390	0.938	0.922
Drule	EAP	13.283	-0.002	-0.001	0.342	0.367	0.941	0.930
	MAP	11.585	0.018	0.013	0.361	0.398	0.935	0.918
Trule	EAP	9.302	0.002	-0.004	0.316	0.363	0.950	0.931
	MAP	7.793	0.010	0.006	0.354	0.390	0.938	0.921
Wrule	EAP	9.302	0.002	-0.004	0.316	0.363	0.950	0.931
	MAP	7.793	0.010	0.006	0.354	0.390	0.938	0.921
KLn	EAP	9.346	0.002	-0.003	0.329	0.36	0.946	0.933
	MAP	7.885	0.015	0.013	0.358	0.394	0.937	0.920

When the interdimensional correlation is 0.9, that is, if the structure is similar to unidimensional structure, the average number of items administered is the fewest. As in other conditions, simulations performed with the MAP estimation method resulted fewer items than simulations performed with EAP method. The calculations obtained using the stopping rules Arule, Wrule and Trule are the same. Simulation with Drule method ended with more items and higher RMSE values than others. The bias values calculated for both dimensions are negligible.

2.3. Simulation Study 3: Comparison of MCAT and CAT Results

In the third simulation study presented , we investigate the effect of treating multi-unidimensional structures as unidimensional structures on adaptive testing results. In line with the purpose, using the item and ability parameters used in the second example, a unidimensional CAT simulation was performed and the outputs were compared with the MCAT simulation. Since it is an item selection method that can be used in both CATs and MCATs, the "KLn" method was used. MAP was used as the estimation method. The data generated for the two-dimensional structure is exported in csv format with the Haven package (Wickham & Miller, 2020). After obtaining the item and ability parameters with the commands example 2, the parameters were exported through the following commands, the ".csv" files were divided and saved for each dimension.

```
#Export parameters
library(haven)
df <- data.frame(a1 = a[,1], a2= a[,2], d1 = d[,1], d2 = d[,2], d3 = d[,3])
write.csv(df, "parameters.csv")
write.csv(thetas, "thetas.csv")
```

After the data sets were saved separately for each dimension, simulation studies continued with ".csv" files. In the UCAT simulation phase, SE($\hat{\theta}$) <0.4 stopping criteria is determined as in the MCAT examples.

```
#Design and start simulation
design = list(min_SEM = 0.4, max_items=300)
mcat3 <- mirtCAT(mo=mod, local_pattern=response, start_item = 'KLn', criteria
= 'KLn', design = design)
```

Average number of items administered, bias, RMSE and r (θ i, $\hat{\theta}$ j) values obtained from CAT simulation performed with MCAT parameters are presented in the Table 8.

Table 8. *Statistics from UCAT simulation.*

Dimension	Interdimensional Correlation	MTL	Bias	RMSE	r (θ i, $\hat{\theta}$ j)
1	0.3	5.335	0.003	0.398	0.924
2		5.130	-0.008	0.407	0.914
1	0.6	5.142	-0.004	0.412	0.917
2		5.120	0.004	0.412	0.911
1	0.9	5.326	-0.006	0.393	0.923
2		5.120	-0.005	0.401	0.917

Compared to MCAT and unidimensional CAT in terms of the average number of items administered, MCAT has a lower average number of items in all conditions. As the interdimensional correlation increases, the average number of items decreases. Unidimensional CAT simulation, on the other hand, resulted in a similar number of items in all conditions. In addition, in MCATs, SE ≤ 0.4 criterion was provided for both dimensions, whereas in CATs, this criterion was only provided for the first dimension when the correlation was 0.3 and 0.9. As in MCAT simulations, bias values are negligible in UCAT.

3. DISCUSSION and CONCLUSION

Since CATs are used for selection, classification and diagnosing purposes, it has important functions for society (Chang, 2015). Technological developments have increased the popularity of CAT applications. With CATs, test length and test session duration are reduced compared to the paper-pencil applications of both achievement tests and psychological scales. While this decrease, the increase in measurement precision makes adaptive testing applications more important. Through the widespread use of MIRT models, MCAT applications are becoming widespread. Researchers frequently apply simulations before CAT and MCAT applications to design the appropriate design for their studies. In this study, data were generated using Monte Carlo simulations by using within-item and between-item dimensionality models. With the generated data, MCAT simulation application codes customized according to different conditions were presented. R programming language was used in this study as it is an open-source and free software. The simulation findings obtained under different conditions are shared. The average number of items administered, RMSE, BIAS and r (θ i, $\hat{\theta}$ j) values obtained using different interdimensional correlation values, different item selection criteria and different parameter estimation methods were examined.

3.1. Main Findings

In this study, the steps of MCAT simulations according to within-item and between-item dimensionality models with the mirtCAT (version: 1.10) package in the RStudio (version: 1.3.1073) software environment were demonstrated. In more detail, multidimensional models applied at the item level to MCAT under within-item and between-item dimensional models using three interdimensional correlation levels, five item selection methods and two parameter estimation methods. MCAT and CAT results performed with data generated according to the between-item dimensionality model were compared. Results showed that MCAT simulations performed with data produced according to the multidimensional models, as the interdimensional correlation increased, the average number of items required to terminate the test decreased. In the MCAT simulations performed according to the within-item dimensionality model, the number of items required to complete the test was higher than the between-item dimensionality model. While increasing the correlation in the within-item dimensionality model greatly changes the average number of items, the average number of items is quite similar in the between-item model.

Wang and Chen (2004) concluded in their study that the higher the correlation between the traits, the less number of items required to reach the same test reliability degree and MCATs will be more efficient than CATs. Similarly, in this study, as the correlation between features increased, the number of items required to complete the MCAT according to the standard error rule decreased. In other words, as the correlation value between traits increases, the number of items required to achieve similar accuracy decreases. And, MCAT was more effective than CAT at meeting the required termination criteria.

When comparing UCATs and MCATs with data generated according to MIRT, the average number of items used in UCAT simulation is higher than MCAT. According to MCAT results, SE <0.4 rule was provided for each dimension, but according to UCAT results, this rule was not provided for all dimensions. A similar result was obtained in Paap, Born, and Braeken's (2018) study. They conducted simulations with the standard error-based termination rule for different design cells and concluded that while meeting the MCAT's termination criteria, CAT failed 80% to meet the termination criterion.

According to the findings obtained, ability parameter estimation method, interdimensional correlations and item selection methods did not much affect measurement fidelity. However, as in the Yao's (2013) study, it can be said that MAP performs similar or better than EAP. The bias values obtained for the different conditions indicate that MCATs give unbiased estimates of ability. The size of interdimensional correlation, item selection criterion and parameter estimation method did not have a considerable effect on the calculated BIAS values for three examples. An interesting finding obtained as a result of the simulations was that KLn was the only method that provided the standard error stopping criteria, regardless of the methods used.

3.2. Future Directions

On the basis of results, for MCAT simulations that researchers will design according to the standard error-based stop rule, it is suggested to add MAP as the estimation method to the simulation conditions. If the standard error rule cannot be defined separately for each dimension, it is recommended to add the KLn rule as the item selection criterion to the simulation conditions. It should be noted that if the stopping rule is not defined for each dimension in MCAT applications, the standard error-based stop rule may not be provided. If the stopping rule is defined by customizing for each dimension, the items continues to be applied until the termination rule is met in all dimensions. Therefore, it is required to specify the standard error rule by customizing it at the desired level for each dimension. If the structure is multidimensional, it is recommended to use MCAT instead of applying separate CAT to each dimension.

Lastly, although the number of MCAT studies has increased in the last decade, more research is needed to investigate scenarios beyond the factors included in the study. For example, different stopping rules, content balancing and other MIRT models can also be investigated. It is important for MCAT practitioners to know with which criteria they can perform MCAT applications more efficiently and effectively.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Since the data in this study were generated by Monte Carlo simulations, there is no need for an ethics committee document.

Authorship Contribution Statement

F. Gul Ince Araci: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing the original draft. **Seref Tan:** Methodology, Supervision, and Validation.

Orcid

F. Gul Ince Araci https://orcid.org/0000-0001-5620-6911
Seref Tan https://orcid.org/0000-0002-9892-3369

REFERENCES

Ackerman, T.A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement*, *15*(1), 13-24.

Aybek, E.C. (2016*). Kendini Değerlendirme Envanteri'nin bilgisayar ortamında bireye uyarlanmış test (BOBUT) olarak uygulanabilirliğinin araştırılması [An investigation of applicability of the self assessment inventory as a computerized adaptive test (CAT)]* [Doctoral Dissertation, Ankara University]. https://dspace.ankara.edu.tr/xmlui/bitstream/handle/20.500.12575/37233/eren_can_aybek.pdf?sequence=1&isAllowed=y

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algortihm. *Pschometrika, 46*(4), 443-459.

Bulut, O., & Sünbül, Ö. (2017). Monte carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266-287. https://doi.org/10.21031/epod.305821

Boyd, A.M., Dodd, B.G., & Choi, S.W. (2010). *Polytomous models in computerized adaptive testing.* In M. L. Nering & R. Ostini (Eds.), Handbook of polytomous item response theory models (pp. 229–255). Routledge.

Chalmers, R.P. (2015). mirtCAT: Computerized adaptive testing with multidimensional item response theory. *R package version 0.6*, *1*. https://CRAN.Rproject.org/package=mirtCAT

Chalmers, R.P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 139. https://doi.org/10.18637/jss.v071.i05

Chalmers, P., Sigal, M., Oguzhan, O., & Chalmers, M. P. (2020). SimDesign: Structure fororganizing monte carlo simulation designs. *R package version 2.2.* https://CRAN.R-project.org/package=SimDesign

Chen, J. (2012*). Applying Item Response Theory methods to design a learning progression based science assessment* [Unpublished Doctoral Dissertation]. Michigan State University.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (Vol. 97, No. 4). ACT, Incorporated.

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

Embretson, S.E., & Reise, S.P. (2000*). Item response theory for psychologists*. Erlbaum.

Feinberg, R.A., & Rubright, J.D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36-49. https://doi.org/10.1111/emip.12111

Finkelman, M., Nering, M.L., & Roussos, L.A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement*, *46*(1), 84103. http://doi.org/10.1111/j.1745-3984.2009.01070.x

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

He, W., & Reckase, M.D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement*, *74*(3), 473-494. https://doi.org/10.1177/0013164413509629

Hornik, K., & FAQ, R. (2010). Frequently asked questions on R. *The R project for Statistical*. https://CRAN.R-project.org/doc/FAQ/RFAQ.html

Lin, H. (2012). *Item selection methods in multidimensional computerized adaptive testing adopting polytomously-scored items under multidımensional generalized partial credit model* [Unpublished Doctoral Dissertation, University of Illinois]. https://hdl.handle.net/2142/34534

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of zihinsel test scores*. Oxford.

Magis, D., Yan, D., & von-Davier, A. (Eds.). (2017). *Computerized adaptive and multistage testing with R: Using packages catr and mstr*. Springer.

Meneghetti, D.D.R., & Junior, P.T.A. (2017). *Application and simulation of computerized adaptive tests through the package catsim*. https:// arxiv.org/pdf/1707.03012.pdf

Mulder, J., & van der Linden, W.J. (2009). Muldimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74*(2), 273-296. https://doi.org/10.1007/s11336-008-9097-5

Nydick, S., & Weiss, D.J. (2009). A hybrid simulation procedure for developments of CATs. *In Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* https://www.iacat.org/sites/default/files/biblio/cat09nydick.pdf

Paap, M.C., Born, S., & Braeken, J. (2019). Measurement efficiency for fixed-precision multidimensional computerized adaptive tests: Comparing health measurement and educational testing using example banks. *Applied Psychological Measurement*, *43*(1), 68-83. https://doi.org/10.1177/0146621618765719

R Core Team (2020). R: *A language and environment for statistical computing* [Computer software manual]. *http://www.R-project.org/*

Reckase, M.D. (2009*). Multidimensional item response theory: Statistics for social and behavioral sciences*. Springer.

Riggelsen, C. (2008). Learning Bayesian networks: a MAP criterion for joint selection of model structure and parameter. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 522-529). IEEE.

Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2), 331-354.

Segall, D.O. (2001). General ability measurement: An application of multidimensional itemresponse theory. *Psychometrika*, *66*, 79-97.

Segall, D.O. (2005). *Computerized adaptive testing.* In K. Kempf-Leonard (Ed.),Encyclopedia of Social Measurement. Academic Press.

Seo, D.G., & Weiss, D.J. (2015). Best Design for Multidimensional Adaptive Testing With the Bifactor Model. *Educational and Psychological Measurement, 75*(6), *954-978.*

Su, Y.H. (2016). A comparison of constrained item selection methods in multidimensional computerized adaptive testing. *Applied psychological measurement*, *40*(5), 346-360. https://doi.org/10.1177/0146621616639305

Team, R. (2020). RStudio: Integrated Development for R (1.3.1073) [Computer software]. RStudio. https://rstudio.com/products/rstudio/

Thissen, D., & Mislevy, R.J., 2000. Testing algorithms. In H. Wainer (Eds.). *Computerized Adaptive Testing*. Lawrence Erlbaum Assc.

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

Thompson, N.A., & Weiss, D.J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation. 16*(1*). 1-9.*

Van der Linden, W., & Glas, G. A. W. (2002). *Computerized adaptive testing: theory and practice.* Kluwer Academic Publishers.

Veerkamp, W.J., & Berger, M.P. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*(2), 203-226. https://doi.org/10.3102/10769986022002203

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J. Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: a primer*. Lawrence Erlbaum.

Wang, C., & Chen, H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement,* 28*, 295-316.* https://doi.org/10.1177/0146621604265938

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, *80*(2), 428-449. https://doi.org/10.1007/s11336-013-9399-0

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computer adaptive testing to educational problems. *Journal of Educational Measurement, 21(4), 361–375.*

Weiss, D.J., & Gibbons, R.D. (2007). Computerized adaptive testing with the bifactor model. *Paper presented at the New CAT Models session at the 2007 GMAC Conference on Computerized Adaptive Testing.* https://mail.iacat.org/sites/default/files/biblio/cat07weiss%26gibbons.pdf

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedure with different stopping rules. *Applied Psychological Measurement*, *37*(1), 3-23. https://doi.org/10.1177/0146621612455687