International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 580–593
https://doi.org/10.21449/ijate.1262199
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Application of the professional maturity scale as a computerized adaptive testing

Süleyman Demir[1,*], **Derya Çobanoğlu Aktan**[2], **Neşe Güler**[3]

[1]Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Türkiye
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[3]İzmir Democracy University, Faculty of Education, Department of Educational Sciences, İzmir, Türkiye

ARTICLE HISTORY

Received: Mar. 08, 2023
Revised: July 29, 2023
Accepted: Sep. 09, 2023

Keywords:
Computerized Adaptive Test (CAT),
CAT application,
Item selection method,
Stopping rule,
Vocational maturity.

Abstract: This study has two main purposes. Firstly, to compare the different item selection methods and stopping rules used in Computerized Adaptive Testing (CAT) applications with simulative data generated based on the item parameters of the Vocational Maturity Scale. Secondly, to test the validity of CAT application scores. For the first purpose, simulative data produced based on Vocational Maturity Scale item parameters were analyzed under different item selection methods (Maximum Fisher Information [MFI],Maximum Likelihood Weighted Information [MLWI] Maximum Posterior Weighted Information [MPWI] Maximum Expected Information [MEI] Minimum Expected Posterior Variance [MEPV] Maximum Expected Posterior Weighted Information [MEPWI]) and stopping rules (Standard Error [SE]<0.30, SE<0.50, SE <0.70, Number of Item [NI]=10, NI=20) by calculating the average number of items, standard error averages, correlation coefficients, bias, and RMSE statistics. For all the conditions of the item selection methods, standard error averages, correlation coefficients, bias, and RMSE statistics showed similar results. When the average number of items is considered, MFI and SE<0.30 were found as most appropriate methods to be used in CAT application. For the second purpose of the study, the paper-pencil form of the Vocational Maturity scale and CAT version were administered to 33 students. A moderate, positive, and statistically significant relationship was found between the CAT application scores and the paper-pencil form scores on the vocational maturity scale. As a result, it can be said that the vocational maturity scale can be applied as a computerized adaptive test and can be used in career guidance processes.

1. INTRODUCTION

The measurement results, which are the foundation of decisions to be made in education and psychology, must be reliable and valid. Decisions made with unreliable and invalid measurement results lead to erroneous evaluations of individuals, teaching methods, and programs. Validity is defined as the process of gathering evidence to support the decisions to be made based on the measurement results. Reliability, on the other hand, is expressed as the degree to which the results obtained from the measurement tool are free from random errors

*CONTACT: Süleyman Demir ✉ suleymand@sakarya.edu.tr Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Türkiye

e-ISSN: 2148-7456 /

(American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], (2014).

The high validity of the measurement results shows that, according to the definition of validity made by Thorndike and Hagen (1961), the measurement results are only related to the variable that is intended to be measured, and that no other feature interferes with the measurement results except for this variable (significantly) (Thorndike & Hagen, 1961; Turgut & Baykul, 2013). These features may be related to the variable to be measured, as well as to include constant, systematic and random errors. The fact that random error does not interfere with the measurement results positively affects both validity and reliability. Therefore, reliable results are required to ensure the results of the validity of the measurement tool. Because unreliable measurement results cannot be valid, the measurement tool should be as reliable as possible with as little random error as possible. According to classical test theory, increasing the number of items in the measurement tool and controlling the sources of random errors as much as possible increases the reliability and thus the validity of the measurement results. Although the number of items in the measurement tool increases the reliability, this increase may cause individuals to lose motivation and fatigue. As a result of this situation, individual-related random errors in measurement occur.

In education and psychology, measurement methods with fewer items have been developed to reduce random errors caused by individuals. Computerized Adaptive Test (CAT) applications are one of these measurement methods. CAT applications can estimate ability levels with fewer items than traditional paper-pencil tests (Gardner et al., 2004; Gibbons et al., 2016; Hol et al., 2008; Kaskatı, 2011; Penfield, 2006; Stochl et al., 2016; Petersen et al., 2016). In traditional paper-pencil tests, individuals answer all items, while in CAT applications they only answer the items relevant to their ability level. The instantaneous ability level is calculated in CAT applications after each item that the individual answers while taking the test. The final ability level calculated for the individual as a result of the CAT application is expected to be similar to the actual ability level. In CAT applications, while the individual answers items based on his or her ability level, he or she is not required to answer items that do not provide information about himself or herself, in other words, items that are higher or lower than his or her ability level (Linacre, 2000; Reckase, 1989; van der Linden, 1998). CAT applications are composed of five basic components: an item pool, a test initiation method, an item selection method, an ability estimation method, and a stopping rule (Dodd et al., 1995; Reckase, 1989; Thompson & Weiss, 2011; Wise & Kingsbury, 2000).

For the reliability and effectiveness of CAT applications to be high, the appropriate components must be used. Monte Carlo simulation studies have been conducted on simulative item parameters and post hoc simulation studies conducted with true item parameters in the literature, methods that allow obtaining measurement results with a high level of validity have been tried to be specified. Furthermore, it is seen that the item selection method is the focus of the vast majority of these studies (Choi & Swartz, 2009; Penfield, 2006; van der Linden, 1998; Veldkamp, 2003).

The item selection method component was defined by Choi and Swartz (2009) as the core of CAT applications, and it was stated that administering items appropriate for the individual's ability level will increase the effectiveness of CAT applications. Item selection methods are examined in two categories: traditional methods and Bayesian methods. Bayesian methods perform item selection methods based on the final distribution, while traditional methods perform item selection based on the item information function.

The Maximum Fisher Information (MFI) method is one of the traditional methods for the item selection in CAT applications. In the MFI method, the item that provides the most information for the instantaneous ability level estimated based on the individual's responses is administered.

In cases where the instantaneous ability level and the true ability level differ, the standard error amount increases because the item used will not be suitable for the true ability level (Hambleton et al., 1991; Penfield, 2006; Thissen & Mislevy, 2000; van der Linden & Pashley, 2000). The MFI method is defined by Lord and Novick (1968) as the Attenuation Paradox, the condition that the reliability and therefore the validity of the measurement results are low despite applying items with maximum information for the instantaneous ability level of individuals.

Figure 1. *Representation of the Attenuation Paradox.*

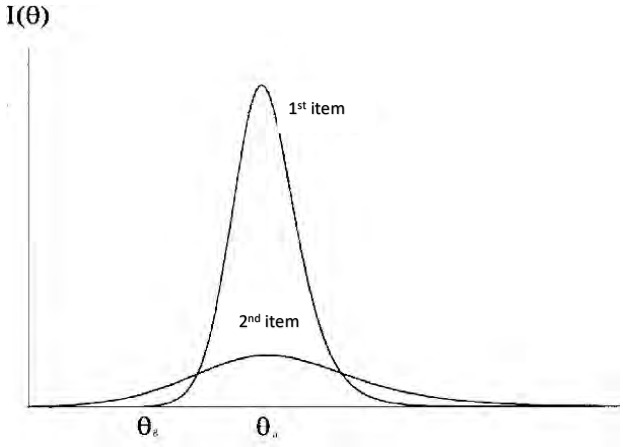

Figure 1 shows the information functions of two different items. θ_g shows the individual's true ability level, while θ_a shows the instantaneous ability level. The first item provides less information on the individual's true ability level (θ_g) than the second item, while the second item provides more information on the individual's instantaneous ability level (θ_a). Therefore, despite providing less information about the actual ability level, the MFI method favors the first item. In this case, the measurement results will be inaccurate due to the application of the item, which provides little or no information on the individual's true ability level.

Another disadvantage of the MFI method is that it results in excessive use of these items due to constant use. Excessive use of some items in maximum information-based methods causes all items not be used and the measurement precision is very high (Davis, 2002; Davis & Dodd, 2008). To avoid the Attenuation paradox and excessive item use, Bayesian statistical approaches to item selection have been developed rather than methods based solely on instead of on information function. In the study conducted by Boztunç-Öztürk and Doğan (2015), it was found that whether item exposure is controlled in maximum information-based and Bayesian item selection methods does not make a significant difference in terms of measurement precision but when item exposure is not controlled in maximum information-based methods, all of the items in the item pool are not used. Not using all of the items is related to item pool size as much as item exposure control (Leroux et al., 2019). In the study conducted by Leroux et al. (2019), it was found that when the item pool is small, all of the items in the item pool are used when even the item exposure is not controlled. Considering this information available in the literature, it can be said that using item selection methods based on maximum information when the item pool is small will not result in high measurement precision or not using all items.

Most CAT research with polytomous models, comparing Bayesian item selection methods and traditional methods appears to be post hoc simulation studies (Choi & Swartz, 2009; Passos et al., 2007; Penfield, 2006; van Rijn et al., 2002; Veldkamp, 2003). While it is expected that measurement results in CAT applications that use the ability estimation method and stopping rules, particularly the item selection method studied in a simulative environment, will have a low level of error, studies on polytomous CAT applications have been limited by method

comparisons in a simulative environment, and the number of application-oriented CAT studies has been limited (Aybek & Demirtaşlı, 2017). Given that measurement tools in the fields of education and psychology are mostly used with paper and pencil in Türkiye, it is believed that increasing the use of CAT applications will be more beneficial to both researchers and participants.

This study compares different item selection methods with the actual application of the "Occupational Maturity Scale" a commonly used educational measurement tool with 40 items and one dimension (Akdaş & Ekinci, 2016; Aktuğ & Birol, 2011; Kutlu, 2012; Orhan & Ültanır, 2011; Sahranç, 2000; Sürücü, 2005; Ulaş & Yıldırım, 2015; Ürün, 2010). The concept of occupational maturity is defined by Super (1957) as meeting the requirements of each professional development step, being ready for the next development step, and having basic abilities that can overcome the difficulties that may be encountered (Kuzgun & Bacanlı, 1995). The feasibility of the Occupational Maturity Scale as a computerized adaptive test is tested, and the amount of error, bias, and correlation coefficients are calculated using the CAT application's simulation under various item selection methods and stopping rules. The relationship between the scores obtained from the CAT application and the paper-pencil test was investigated using minimum error methods with these coefficients.

While data obtained with the scale are more objective, valid, reliable, and useful than data obtained through non-test techniques (such as observation, interview, etc.), there may be random errors in the measurement results due to factors such as low motivation in individuals in answering the scale items, social desirability, psychological fatigue, and the length of the scale. As a result of this situation, researchers are focusing on alternative data collection methods rather than traditional paper-pencil methods. One such method is CAT applications, which can estimate the ability associated with the actual ability level at a high level with a much smaller number of items. This estimate is based on the application of items appropriate to the individual's own ability level. Therefore, selecting the appropriate item for the individual is critical to the effectiveness of CAT applications. However, there is limited research on the comparison between traditional item selection methods and Bayesian item selection methods in real CAT applications and post-hoc simulation studies (Aybek & Demirtaşlı, 2017; Choi & Swartz, 2009; Passos et al., 2007; Penfield, 2006; van der Linden, 1998; van Rijn et al., 2002; Veldkamp, 2003). Thus, this study's results are expected to contribute to both the occupational guidance process and the usability of CAT in scientific studies.

1.1. Research Problems

This study aims to address the following research questions using CAT applications:

1. Does the correlation coefficient between the simulatively estimated occupational maturity level and the actual occupational maturity level differ depending on the item selection method and stopping rules used, mean number of items administered, standard error means, bias and RMSE values?
2. Is there a relationship between the CAT application and the occupational maturity levels obtained from the paper-pencil test application using the determined item selection method and stopping rule?

2. METHOD

2.1. Participants

Before starting to collect data, permission was obtained from Hacettepe University Ethics Commission with the decision dated 24.10.2017 and numbered 433-3695. This study's data were gathered from two distinct groups of participants. The first group's data were employed to determine the item parameters of the Occupational Maturity Scale and to test the IRT

assumptions. The second group's data was used to test the validity of the CAT application of the Occupational Maturity Scale. Table 1 shows the demographic information of the first and second groups.

Table 1. *Demographic information of the participants.*

			Frequency	Percentage
The first group	Gender	Female	366	54.06%
		Male	311	45.94%
	Class	11th Grade	510	75.33%
		12th Grade	167	24.67%
The second group	Gender	Female	16	48.48%
		Male	17	52.52%

The first group consisted of 677 students in the 11th and 12th grades from Adapazarı, Erenler, Hendek and Serdivan districts of Sakarya. For this data group, firstly, permission was obtained from Sakarya National Education Directorate. 12 high schools were determined by cluster sampling from high schools located in Serdivan, Erenler, Hendek and Adapazarı. Data were collected from 677 students on a voluntary basis from 11th and 12th grade students studying in these high schools. Of these 677 students, 366 were female and 311 were male; 510 of them are in the 11th grade and 167 are in the 12th grade.

The second group consisted of 33 students in the 11th and 12th grade from Private School of Sakarya University Foundation. 33 students on a voluntary basis CAT application of the Occupational Maturity Scale and the paper-pencil test application were carried out. When the literature is examined, it is recommended to be 1-2 weeks between the two applications (Bardhoshi & Erford 2017; Cattell, 1986; Cattell et al., 1970; Deyo et al., 1991; Nunnally & Bernstein, 1994), and the application was made by leaving 10 days between the CAT and paper-pencil test applications.

2.2. Data Collection Tools and Methods

The occupational Maturity Scale used in this study was developed by Kuzgun and Bacanlı (2005). The scale consists of 40 items with one dimension and was reported to have an internal consistency coefficient of 0.89, and a test-retest reliability coefficient of 0.82. The scale was administered to the first group of participants to obtain the item parameters of the Occupational Maturity Scale and to test the IRT assumptions.

Based on these data, simulative data were produced according to different item selection methods and stopping rules with the FIRESTAR program using the item parameters of the Occupational Maturity Scale, and the CAT application was prepared with the CONCERTO platform.

2.3. Data Collection

To create an item pool, there must be at least 24-30 items that can provide information at all ability levels. However, having a certain number of items does not necessarily mean that the CAT application will be sufficient. The item information and test information functions are also critical for CAT applications (Dodd et al., 1995).

The item parameters of the Occupational Maturity Scale were determined using the IRTPRO package program. To determine the item parameters, a one-dimensional Item Response Theory (IRT) analysis was performed under a graded response model. The analysis revealed that, the step parameters of the four items (2nd, 4th, 6th and 20th) were outside the ranges (-4.00, +4.00), which are the lower and upper limits for the IRT models. Four items were removed from the item pool because the a_i parameter was less than 0.60, the information functions were weak,

and data production did not occur when these items were in the item pool while generating simulative data.

With the use of the simulated data generated by the FIRESTAR software, various item selection strategies and stopping rules were explored to assure the greatest effect from the CAT application of the Occupational Maturity Scale, whose item parameters were established. Using the simulative data, the average number of items, the correlation between the CAT application and the occupational maturity levels determined by the paper-pencil exam, the bias and the RMSE statistics were calculated. Table 2 lists the methods for the item selection and the stopping rules.

Table 2. *The methods of selection of items used in the simulation and the rules of termination.*

Manipulated Variable	Methods	Number of Conditions
Item Selection Method	Maximum Fisher Information (MFI) Maximum Likelihood Weighted Information (MLWI) Maximum Posterior Weighted Information (MPWI) Maximum Expected Information (MEI) Minimum Expected Posterior Variance (MEPV)	6
Stopping Rule	Standard Error < 0.30 Standard Error < 0.50 Standard Error < 0.70	3
	NI=10 NI=20	2

Table 2, shows that a total of 30 different conditions have occurred under different item selection methods and different stopping rules. A total of 30,000 people's simulated data were produced, with 1,000 people in each situation. In the study, while the item selection method and the stopping rule were manipulated, other variables held constant in the study are listed below.

- Selection of the first item: 0 ability level (θ=0.00)
- Sample mean and standard deviation: *sd*=1.00
- Frequency of item use control: Not used (coded as 1).
- Ability estimation method: Expected Final Estimation Method
- Minimum and maximum ability levels: -4.00, +4.00
- IRT model: Graded Response Model
- Scaling: 1.7
- Ability increase value: 0.10
- Standard error calculation method: Final distribution
- A priori distribution: $\bar{X} = 0.00$, sd=1.00

2.4. Computerized Adaptive Testing Application

During the spring semester of the 2017-2018 academic year, the paper-pencil form and the CAT version of the occupational maturity scale were administered to 33 (11^{th} and 12^{th} grade) students. The paper-pencil test was administered in the students' own classrooms at the beginning of their course by the researcher. Ten days after the paper-pencil form was administered to the students, the CAT version was administered to the participants. The CAT was carried out by a researcher in the computer laboratory located in the same building as the classrooms at the school. Students' transportation from their classrooms to the laboratory was provided by guidance counsellors and course teachers. Since the test was online, all computers and the internet connection were checked and the relevant web page was opened and made

available for students' use. Students logged in by entering their student numbers and gender to give feedback to the CAT application and to match it with the paper-pencil test. After entering the necessary information, the students clicked on the "Continue" button to begin. The student's answers to the items in the CAT application were recorded in the database. If a student wanted to pass an item without answering, a warning message "Please do not pass without answering the Item!" appeared. At the end of the CAT application, which continued until the specified condition was met, an information screen about the Occupational Maturity ability level was displayed.

2.5. Analysis of the Data

To answer the first research problem, correlation coefficient, standard error mean, bias and RMSE statistics were calculated using IRTPRO, Excel and SPSS 17.0. Then these values were examined to determine whether they differed.

• The IRTPRO package program was used to determine the item parameters of the Occupational Maturity Scale. To use the IRTPRO program, a 15-day trial version was rented from Scientific Software International by e-mail at 2018. As a result of the analysis performed under the Progressive Response Model, a_i and β_{ij} were calculated.

• To analyze the simulated data, the correlation coefficient (Pearson Product Moments Correlation Coefficient), and the average number of items applied, the standard error mean (SE), bias, and RMSE statistics were calculated between the simulatively estimated and actual occupational maturity levels for each condition by using Excel and SPSS 17.0. High correlation coefficient, low standard error, bias, and low RMSE statistics (close to 0) indicate that there is no difference (deviation) between individuals' true ability level and estimated ability level. The methodology for calculating bias, RMSE, and standard error averages—three statistics used to compare various stopping rules—is described here:

• The standard error can be calculated in two different ways according to the IRT, depending on the information function and depending on the final distribution. During the production of simulative data, standard error calculation was performed depending on the final distribution.

$$SE_{post} = \sqrt{\text{var}(g(\theta)|U)}$$

• The bias statistic is equal to the average of the difference between the actual value of a parameter and the estimated value.

$$BIAS = \frac{\sum_{i=1}^{n}(\theta_{ig} - \theta_{ik})}{n}$$

• The RMSE statistic is the average of the squares of the difference between the true value and the predicted value of a parameter.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\theta_{ig} - \theta_{ik})^2}{n}}$$

To answer the second research problem, the correlation coefficient was calculated. The correlation coefficient between the CAT application and the paper-pencil application was calculated using the SPSS 17.0 program.

3. FINDINGS

3.1. Findings Related to the First Research Problem

Table 3 presents the findings for the first research problem based on different stopping rules.

Table 3. *The average item obtained under different item selection methods for different stopping rules, mean of SE, correlation coefficient, bias and RMSE statistics.*

Stopping Rule	Item Selection Method	Mean of Item	Mean of SE	r	Bias	RMSE
SE<0.30	MFI	5.44	0.175	0.952	0.147	0.101
	MLWI	5.43	0.174	0.929	0.165	0.157
	MPWI	5.16	0.186	0.944	0.169	0.120
	MEI	5.64	0.171	0.949	0.152	0.111
	MEPV	5.36	0.174	0.951	0.148	0.106
	MEPWI	5.31	0.179	0.945	0.164	0.119
SE<0.50	MFI	4.07	0.259	0.936	0.224	0.136
	MLWI	4.00	0.264	0.936	0.228	0.136
	MPWI	4.01	0.255	0.936	0.223	0.135
	MEI	4.01	0.259	0.937	0.223	0.133
	MEPV	4.01	0.257	0.939	0.217	0.130
	MEPWI	4.01	0.255	0.936	0.223	0.135
SE<0.70	MFI	4.00	0.267	0.933	0.233	0.141
	MLWI	4.00	0.264	0.936	0.228	0.136
	MPWI	4.00	0.257	0.936	0.225	0.135
	MEI	4.00	0.260	0.937	0.225	0.134
	MEPV	4.00	0.257	0.939	0.217	0.129
	MEPWI	4.00	0.255	0.936	0.223	0.135
NI=10	MFI	10.00	0.136	0.965	0.124	0.076
	MLWI	10.00	0.137	0.962	0.130	0.082
	MPWI	10.00	0.139	0.962	0.131	0.082
	MEI	10.00	0.134	0.963	0.128	0.080
	MEPV	10.00	0.125	0.969	0.108	0.066
	MEPWI	10.00	0.139	0.962	0.131	0.082
NI=20	MFI	20.00	0.080	0.986	0.067	0.080
	MLWI	20.00	0.085	0.983	0.065	0.033
	MPWI	20.00	0.076	0.985	0.062	0.033
	MEI	20.00	0.080	0.987	0.057	0.029
	MEPV	20.00	0.077	0.988	0.054	0.027
	MEPWI	20.00	0.079	0.985	0.061	0.033

The results show that the correlation coefficient between the estimated and true occupational maturity level of individuals produced simulatively under different item selection methods and stopping rules, the average number of items applied, standard error averages, bias, and RMSE statistics provide similar outcomes. In addition, when using variable-length stopping rules (SE<0.30, SE<0.50, SE<0.70), the test is completed using an average of between 4 to 5.64 items. In this case, the number of items decreases by 84% to 90% compared to the original scale.

Under different item selection methods and stopping rules, the lowest mean standard error was determined as 0.076 (for the condition MPWI; NI=20), while the highest mean standard error was determined as 0.267 (for the condition MFI; SE<0.70). The highest correlation coefficient

between the predicted occupational maturity level and the actual occupational maturity level was determined as 0.988 (for the condition MEPV; NI=20), and the lowest correlation coefficient was determined as 0.929 (for the condition MLWI; SE<0.30). The lowest bias statistic was calculated as 0.054 (for MEPV; NI=20 condition) and the highest bias statistic was calculated as 0.233 (for MFI; SE<0.70 condition). The lowest RMSE statistic was obtained as 0.027 (for MEPV; NI= 20 condition) and the highest RMSE statistic was obtained as 0.141 (for MFI; SE<0.70 condition).

Overall, it is seen that the most appropriate stopping rule is NI=20, and the most appropriate item selection method is MEPV. When the stopping rule is set as NI=20, the CAT application is expected to end with 45% fewer items than the original scale, while when the stopping rule is SE<0.30, the CAT application is expected to end with 85% fewer items than the original scale. Therefore, it is suggested that the SE<0.30 stopping rule should be used for real CAT applications, considering the low level of differences between correlation coefficients, bias, and RMSE statistics, and the significant decrease in the number of items.

When the stopping rule was determined as SE<0.30, the highest correlation coefficient (0.952), the lowest bias (0.147) and RMSE statistics (0.101) were obtained based on the MFI method. Thus, it was predicted that it would be more appropriate to determine the item selection method as MFI and the stopping rule as SE<0.30 in the actual CAT application.

3.2. Findings Related to the Second Research Problem

Table 4 presents descriptive statistics of the occupational maturity levels of individuals obtained from the CAT application and the paper-pencil test administration.

Table 4. *Descriptive statistics of occupational maturity levels obtained from CAT and paper-pencil test applications.*

	N	Minimum	Maximum	Mean	*sd*
CAT	33	-0.19	2.46	1.283	0.626
Paper-Pencil	33	-0.06	2.26	0.812	0.600

The results show that a minimum of -0.19 and a maximum of 2.46 occupational maturity level were estimated from the CAT application. The average of the occupational maturity levels obtained from the CAT application was 1.28, while the standard deviation was 0.63. On the other hand, the minimum and maximum occupational maturity levels estimated from the paper-pencil test application were -0.06 and 2.26, respectively. The average of the occupational maturity levels obtained from the CAT application was 0.81, while the standard deviation was 0.60. Furthermore, there was a moderate (r=0.535) positive and statistically significant relationship between the CAT application and the occupational maturity levels obtained from the paper-pencil test application (p<0.05).

4. DISCUSSION and CONCLUSION

The results of the study indicate that the correlation coefficient between the estimated occupational maturity level and the true occupational maturity level of individuals produced simulatively under different item selection methods and stopping rules, the average number of items applied, standard error averages, bias and RMSE statistics provide similar results. This finding is consistent with previous studies that compared different methods of item selection and stopping rules (Aybek & Demirtaşlı, 2017; Choi & Swartz, 2009; Ho, 2010; Veldkamp, 2003). However, the results of this study differ from Penfield's (2006) study, which found that the MPWI and MEI methods had a higher level of measurement precision than the MFI method

(in case the information functions of the items in the item pool were flat), while the MPWI and MEI methods were not different from each other.

In a study by van der Linden's (1998) that compared the item selection methods and stopping rules under the 2-parameter model, the MFI and MPWI methods had the highest bias statistics (NI=5 and 10), while the other three item selection methods (MEI, MEPV and MEPWI) had lower bias.

The study also found that when variable length stopping rules are used (SE<0.30; SE<0.50; SE<0.70), ability estimation can be made with 84%-90% fewer items than the original scale. This finding is in parallel with the advantage of the following ratios in the number of items: the rate of 73.3% was reached as a result of the study of Gardner et al. (2004); the rates of 36%-65% reached as a result of the study of Smits et al. (2011); the 50.86% rate reached as a result of Aybek and Demirtaşlı's (2017) study; the rate of 75% obtained in the study of Gibbons et al. (2016); the rate of 67% obtained in the study by Stochl et al. (2016); the 50%-85% rates obtained in the study by Petersen et al. (2016); the 30%-71% rates obtained in the study by Choi and McClenen (2020); the rate of 75% obtained in the study by Harrison et al. (2020); the 50%-63% rates obtained in the study by Yasuda et al. (2021); the 62%-96% rates obtained in the study by Liu et al. (2022); the rate of 78% obtained in the study by Giordano et al. (2023). In the studies conducted by Smits et al. (2011) and Aybek and Demirtaşlı (2018), it can be said that the low ratio in test lengths is because there is a more limited pool of items compared to other studies.

Furthermore, there was a moderate (r=0.535) positive and statistically significant relationship between the CAT application and the occupational maturity levels obtained from the paper-pencil test application (p<0.05). The correlation to be obtained from CAT and paper-pencil application is expected to be high as in simulation studies. Compared to the correlation coefficient obtained in the simulation CAT study (r=0.952; MFI, SE<0.30; see Table 3), the correlation coefficient obtained in the real CAT study (r=0.535) is lower. It can be said that there are several reasons for this. Firstly, the correlation coefficient obtained from the real CAT application is relatively lower than the correlation coefficient obtained from the simulation data, due to the sample size. Because the sample size is effective in calculating the correlation coefficient (Green, 1991; Harris, 1985; Tabachnick & Fidell, 1996; Wilson & Morgan, 2007). In addition, the application of CAT to students in the computer laboratory instead of the classroom may have caused random errors to be mixed in the measurement results. In this case, the difference in the measurement results obtained from the CAT and paper-pencil application caused the correlation coefficient to be low. In addition, in studies with dichotomous and polytomous measurement tools, it is seen that the correlation coefficient obtained from the real CAT study is lower than the correlation coefficient obtained from the simulation CAT study (Aybek & Demirtaşlı, 2018; Şahin & Gelbal, 2020).

Acknowledgments

This study was produced from the doctoral thesis titled “Investigation of different item selection methods in terms of stopping rules in polytomous computerized adaptive testing”.

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**: Hacettepe University, 24.10.2017, 433-3695.

Authorship Contribution Statement

Süleyman Demir: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Derya Çobanoğlu Aktan**: Methodology, Supervision, and Validation. **Neşe Güler**: Methodology, Supervision, and Validation.

Orcid

Süleyman Demir https://orcid.org/0000-0003-3136-0423
Derya Çobanoğlu Aktan https://orcid.org/0000-0002-8292-3815
Neşe Güler https://orcid.org/0000-0002-2836-3132

REFERENCES

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Akdaş, G., & Ekinci, M. (2016). Sağlık meslek lisesi öğrencilerinin mesleki olgunluk düzeylerinin ve algıladıkları aile desteğinin incelenmesi [Analysis of vocational school of health students' professional maturity and family support perception levels]. *Uluslararası Hakemli Psikiyatri ve Psikoloji Araştırmaları Dergisi 7, 83-100.* https://doi.org/10.17360/UHPPD.2016723147

Akıntuğ, Y., & Birol, C. (2011). Lise öğrencilerinin mesleki olgunluk ve karar verme stratejilerine yönelik karşılaştırmalı analiz [Comparative analysis of vocational maturity and decision making strategies of high school students]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41,* 1-12. http://efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/675-published.pdf

Aybek, E.C., & Demirtaşlı, R.N. (2017). Computerized adaptive test (CAT) applications and item response theory models for polytomous items. *International Journal of Research in Education and Science, 3*(2), 475-487. https://doi.org/10.21890/ijres.327907

Aybek, E.C., & Çıkrıkçı, R.N. (2018). Kendini Değerlendirme Envanteri'nin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliği [Applicability of the Self-Assessment Inventory as a computerized adaptive test]. *Turkish Psychological Counseling and Guidance Journal, 8*(50), 117-141. https://dergipark.org.tr/tr/pub/tpdrd/issue/40299/481364

Bardhoshi G., & Erford B.T. (2017). Processes and procedures for estimating score reliability and precision. *Measurement and Evaluation in Counseling and Development, 50*(4), 256-263. https://doi.org/10.1080/07481756.2017.1388680

Cattell R.B. (1986). The psychometric properties of tests: Consistency, validity, and efficiency. In Cattell R.B., Johnson R.C. (Eds.), *Functional psychological testing* (pp. 54-78). Brunner/Mazel.

Cattell R.B., Eber H.W., & Tatsuoka M.M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Institute for Personality and Ability Testing.

Choi, S.W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33(1),* 644-645. https://doi.org/10.1177/0146621608329892

Choi, S.W., & Swartz, R.J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33(1),* 419-440. https://doi.org/10.1177/0146621608327801

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences, 10*(22), 81-96. https://doi.org/10.3390/app10228196

Davis, L.L. (2002). *Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items* [Unpublished doctoral dissertation]. The University of Texas.

Davis, L.L., & Dodd, B.G. (2008). Strategies for controlling item exposure in computerized adaptive testing with the partial credit model. *Journal of Applied Measurement, 9*(1), 1-17. https://pubmed.ncbi.nlm.nih.gov/18180546/

Deyo, R.A., Diehr, P., & Patrick, D.L. (1991). Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Controlled Clinical Trials, 12*(4), 142-158. https://doi.org/10.1016/S0197-2456(05)80019-4

Dodd, B.G., De Ayala, R.J., & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*(1), 5-22. https://doi.org/10.1177/014662169501900103

Gardner, W., Shear, K., Kelleher, K.J., Pajer, K.A., Mammen, O., Buysse, D., & Frank, E., (2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry, 4*(13).

Gibbons C., Bower P., Lovell K., Valderas J., & Skevington S. (2016). Electronic quality of life assessment using computer-adaptive testing. *Journal of Medical Internet Research,* 18. https://doi.org/10.2196/jmir.6053

Giordano, A., Testa, S., Bassi, M. et al. (2023). Applying multidimensional computerized adaptive testing to the MSQOL-54: a simulation study. *Health Qual Life Outcomes* 21, 61 https://doi.org/10.1186/s12955-023-02152-8

Green, S.B. (1991). How many subjects does it take to do a regression analysis?. *Multivariate Behavioral Research*, *26*, 499-510.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Sage.

Harris, R.J. (1985). *A primer of multivariate statistics.* Academic Press.

Harrison, C., Loe, B.S., Lis, P., & Sidey-Gibbons, C. (2020). Maximizing the potential of patient-reported assessments by using the open-source concerto platform with computerized adaptive testing and machine learning. *Journal of Medical Internet Research, 22*(10), 1–8. https://doi.org/10.2196/20950

Ho, T. (2010). *A Comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on the Generalized Partial Credit Model,* [Unpublished doctoral dissertation]. University of Texas.

Kutlu, M. (2012). Anadolu ve genel lise öğrencilerinin çeşitli değişkenlere göre mesleki olgunluk düzeylerinin incelenmesi [An analysis of vocational maturity levels of anatolian and general high school students in terms of some variables]. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 13*(1), 23-41. https://dergipark.org.tr/tr/download/article-file/92233

Kuzgun, Y., & Bacanlı, F. (2005). Mesleki Olgunluk Ölçeği el kitabı [Professional Maturity Scale handbook]. MEB Basımevi.

Linacre, J.M. (2000). Computer-adaptive testing: A methodology whose time has come. In Chae, S., Kang, U., Jeon E., & Linacre J.M. (Eds.), *Development of computerized middle school achievement test (in Korean)*. Komesa Press.

Liu, K., Zhang, L., Tu, D., & Cai, Y. (2022). Developing an Item bank of computerized adaptive testing for eating disorders in Chinese University students. *SAGE Open*, *12*(4). https://doi.org/10.1177/21582440221141273

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Nunnally, J., & Bernstein, I.H. (1994). *Psychometric theory*. McGraw-Hill.

Orhan, A.A., & Ültanır, E. (2014). Lise öğrencilerinin mesleki olgunluk düzeyleri ile karar verme düzeyleri [Vocational maturity level and decision making strategies of high school

students]. *Ufuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 3*(5), 43-55. https://dergipark.org.tr/tr/download/article-file/1358749

Passos, V., Berger, M.P.F., & Tan, F.E. (2007). Test design optimization in CAT early stage with the Nominal Response Model. *Applied Psychological Measurement, 31*(3), 213–232. https://doi.org/10.1177/01466216062915

Penfield, R.D. (2006). Applied Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*, 1-20. https://doi.org/10.1207/s15324818ame1901_1

Petersen, M.A., Gamper, E.M., Costantini, A., Giesinger, J.M., Holzner, B., Johnson, C., Sztankay, M., Young, T., Groenvold, M. (2016). An emotional functioning item bank of 24 items for computerized adaptive testing (CAT) was established. *Journal of Clinical Epidemiology, 70*, 90–100. https://doi.org/10.1016/j.jclinepi.2015.09.002

Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement Issues and Practice, 8,* 11-15. https://doi.org/10.1111/j.1745-3992.1989.tb00326.x

Sahranç, Ü. (2000). *Lise öğrencilerinin mesleki olgunluk düzeylerinin denetim odaklarına göre bazı değişkenler açısından incelenmesi [A Study on some variables affecting career maturity levels of high school students depending on their locus of control],* [Unpublished master dissertation]. Hacettepe University.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research, 188*(1), 147-155. https://doi.org/10.1016/j.psychres.2010.12.001

Stochl, J., Böhnke, J.R., Pickett, K.E., & Croudace, T.J. (2016). An evaluation of computerized adaptive testing for general psychological distress: Combining GHQ-12 and Affectometer-2 in an item bank for public mental health research. *BMC Medical Research Methodology, 16*(1), 1-15. https://doi.org/10.1186/s12874-016-0158-7

Super, D.E. (1957). *The psychology of careers*. Harper.

Sürücü, M. (2005). *Lise öğrencilerinin mesleki olgunluk ve algıladıkları sosyal destek düzeylerinin incelenmesi [High school students' vocational maturity and perceived social support level],* [Unpublished master dissertation]. Gazi University.

Şahin, M.D., & Gelbal, S. (2020). Development of a multidimensional computerized adaptive test based on the bifactor model. *International Journal of Assessment Tools in Education, 7*(3), 323-342. https://doi.org/10.21449/ijate.707199

Tabachnick, B.G., & Fidell, L.S. (1996). *Using multivariate statistics*. HarperCollins.

Thissen, D., & Mislevy, R.J. (2000). Testing algorithms. In H. Wainer (Ed.). *Computerized adaptive testing*, (101-135). Lawrence Erlbaum Assc.

Thompson, N.A., & Weiss, D.J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research and Evaluation, 16*(1), 1-9. https://doi.org/10.7275/wqzt-9427

Thorndike R.L., & Hagen, E. (1961). *Measurement and evaluation in psychology and education.* John Wiley and sons.

Turgut, M.F., & Baykul, Y. (2013). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education].* PegemA Yayıncılık.

Ulaş, Ö., & Yıldırım, İ. (2015). Lise öğrencilerinde mesleki olgunluğun yordayıcıları [Predictors of career maturity among high school students]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 30*(2), 151-165. http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/14-published.pdf

Ürün, A.E. (2010). *Lise öğrencilerinin kendine saygı düzeyleri ile mesleki olgunlukları arasındaki ilişki [The relationship between the self-esteem level and the vocational maturity of high school students]* [Unpublished master dissertation]. Balıkesir University.

van der Linden, W.J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika, 62*, 201–216. https://link.springer.com/article/10.1007/BF02294775

van der Linden, W.J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Kluwer.

Van Rijn, P., Eggen, T.J., Hemker, B.T., & Sanders, P.F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement 26*, 393- 411. https://doi.org/10.1177/014662102237796

Veldkamp, B.P. (2003). Item selection in Polytomous CAT. In H., Yanai, A., Okada, K., Shigemasu, Y., Kano & J.J. Meulman (Eds.), *New Developments in Psychometrics* (pp. 207-214). Springer Verlag.

Wilson, C.R., & Morgan, B.L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology, 3*(2), 43-50. https://doi.org/10.20982/tqmp.03.2.p043

Wise S.L., & Kingsbury G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*, 135-155. https://www.uv.es/revispsi/articulos1y2.00/wise.pdf

Yasuda, J., Mae, N., Hull, M.M., & Taniguchi, M., (2021). Optimizing the length of computerized adaptive testing for the force concept inventory. *Physical Review Physics Education Research, 17*(1), 1-15. https://doi.org/10.1103/PhysRevPhysEducRes.17.010115