ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 47-61

Developing A Computerized Adaptive Test Form of the Occupational Field Interest Inventory

Volkan ALKAN* Kaan Zülfikar DENİZ**

Abstract

In this research, the aim was to apply the Occupational Field Interest Inventory (OFII), which was developed in paper-pencil format, as a Computerized Adaptive Test (CAT). For this purpose, the paper and pencil form of the OFII was applied to 1425 high school students and post-hoc simulations were carried out with the obtained data. According to results obtained from the simulations, it was decided that the most ideal criteria for the CAT application were GPCM as the IRT model, .40 standard error value as the test termination rule, and MFI as the item selection method. The OFII ended with an average of 59 items, and the correlations between scores obtained from the paper-pencil form and thetas (θ) estimated by simulation ranged between .91-.97. According to post-hoc simulation results, the CAT application was applied to 150 students. It was observed that the correlations between the scores of students from the online application of the paper-pencil form and θ levels estimated by the CAT form varied between .73 and .91.

Keywords: Computerized Adaptive Test, Item Response Theory, Occupational Field Interest Inventory, Occupational Interest

Introduction

Having an occupation is an important factor for people to maintain their lives to a certain standard by obtaining the necessary income to do so, which can also play an important role in determining an individual's social prestige as well as their achievement of happiness (Altın, 2020). While choosing an occupation suitable for oneself, individuals often make their choice based on comparing their personal knowledge (i.e., lifestyle, interest, skills, values, etc.) with the available occupations as well as the conditions of those occupations (Akar, 2012). According to Yoo (2016), the factors which affect an individual's career choice can be listed as occupational interest, talent, personality, value, socioeconomic status, and gender. Among these factors, one of the variables that most affects an individual's career choice is occupational interest.

Occupational interest is initially determined by an individual's liking for people who do a specific job. For example, occupational interests are determined by assuming that someone who loves teachers will in effect have an interest in the teaching occupation. Later, this method was abandoned, and occupational interests were then determined according to the individuals' enjoyment of behaviors belonging to various occupations (Deniz, 2009). Today, occupational interest is mostly determined by asking individuals about their level of interest in a range of work activities through inventories.

When the measurement tools used to measure occupational interest were examined, it could be seen that most of them were developed based on Classical Test Theory (CTT). CTT has been widely used in measurement applications such as test development, application, and evaluation since the early 1900s (Hambleton et al., 1991). In CTT, the sum of scores that an individual obtains from items of the measurement tool are defined as the degree of possessing the feature to be measured. Although there are some exceptions, a low score that the individual obtains from the measurement tool generally indicates

* PhD., Ankara University, Faculty of Education, Ankara-Türkiye, volkanalkan114@gmail.com, ORCID ID: 0000-000-8264-8190
** Prof. Dr., Ankara University, Faculty of Education, Ankara-Türkiye, kzdeniz@ankara.edu.tr, ORCID ID: 0000-0003-0920-538X

To cite this article:
Alkan, V., & Deniz, K. Z. (2023). Developing a computerized adaptive test form of the Occupational Field Interest Inventory. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 47-61. https://doi.org/10.21031/epod.1153713

Received: 3.08.2022
Accepted: 15.02.2023

that the level of possessing the desired feature being measured is low, while a high score indicates that the level of having the desired feature being measured is high. CTT applications are mostly concerned with test-level information such as reliability. In addition, it should be noted that although CTT allows for obtaining item-level information such as item discrimination index and average of item scores, CTT does have important limitations. The limitations of CTT are that item statistics are dependent on the group, test scores obtained by individuals are dependent on the test items, the inability to distinguish individuals being different from the average in terms of ability level, and measurement error is considered the same for all individuals while measurement error is actually different for each individual (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Meyer, 2010). As a result, over time the limitations of CTT have been discussed and many have sought a new model to eliminate these limitations. Thus, the model developed by taking these limitations into account is the Item Response Theory (IRT).

IRT is an item-based theory based on the psychological measurement studies of Binet, Simon and Terman in 1916. The first studies regarding IRT were made by Thorndike, Thurstone, Horst and Symonds in the 1920s, and in the subsequent years, Lord, Novic and Lawley continued studies regarding IRT as well as significantly contributed to the theory's development (Ostini & Nering, 2010). IRT, especially as a result of developments in computer technology, has frequently been used in the measurement of various characteristics within the field of psychology and education, and has also been developed as an alternative for addressing limitations arising from the structure of CTT (Harvey & Hammer, 1999).

IRT has many application areas, and in particular, one of these application areas is the Computerized Adaptive Test (CAT), which was developed using IRT. CAT is a computer-based application in which each individual test-taker does not answer the same items, but instead only the items appropriate to their skill-feature levels as measured within the test (Kezer & Koç, 2014).

The development of CAT models suitable for both two-category and multi-category items have brought to the forefront the idea of developing CAT-forms for the measurement of items normally applied in a paper-pencil format. Thus, valid and reliable measurement tools, which can be difficult to implement in terms of application time, are applied in a shorter amount of time through the use of CAT applications due to fewer items being needed than in the paper-pencil form (Özbaşı & Demirtaşlı, 2015). In addition, with the help of CAT, it is possible to perform more reliable measurements in a shorter amount of time by not querying individuals using items that are well above or well below their ability level (Şahin & Özbaşı, 2017). Furthermore, some measurement tools may need updating according to the technological and social developments experienced, which may ultimately reduce the usefulness of the scales by causing an increase in the number of items used within the measurement tools. In this respect, the adaptation of valid and reliable measurement tools originally applied in a paper-pencil format to a CAT format also facilitates updating studies to be carried out regarding these scales.

The Occupational Field Interest Inventory (OFII), which is the subject of this research, is one of the inventories for which the CAT application had yet to be developed. The OFII is an interest inventory which includes 14 subscales, consisting of 156 items, and the paper-pencil application of this inventory takes approximately 15-20 minutes. When the CAT application for this inventory was developed, it was clear that the usefulness of the inventory would increase by shortening the application time as well as new sub-scales belonging to different occupations could be added. The primary purpose of the current research was to further develop the CAT form by determining the most appropriate IRT model, test termination rule, and item selection method for the OFII, which was developed to assist students in their career choices as part of student occupational guidance services. GPCM Generalized Partial Credit Model (GPCM) and Graded Response Model (GRM) were preferred as the IRT model, .30, .40, .50 as the standard error value and Maximum Fisher Information (MFI), Maximum Expected Information (MEI), Minimum Expected Posterior Variance (MEPV) and Maximum Expected Posterior Weighted Information (MEPWI) as the item selection method. These are preferred because the platform used in the research allows working with these options, and these options are generally preferred in scales developed in accordance with multi-category models (Aybek & Çıkrıkçı, 2018; Boyd et al., 2010;

Özbaşı, 2014; Şimşek 2017; Van der Linden, 1998). In line with the determined general purpose, answers were sought to the following questions.

1. When different IRT models (GPCM and GRM) and three different standard error values (.30, .40, and .50) and different item selection methods (MFI, MEI, MEPV, and MEPWI) are used as test termination rules:
   a) How many items are used on average in the OFII-CAT (OFII-C) simulations?
   b) How do the standard error values to be obtained from the OFII-C simulations change?
   c) How does the direction and level of the relationship change between the θ levels obtained from the OFII-C simulation and the paper-pencil application of the OFII (OFII-PP)?
2. When the OFII-C application is created using the test termination rule, item selection method, and IRT model determined according to the findings obtained from the OFII-C simulation results:
   a) How does the frequency of item use change in OFII-C and OFII-C simulation applications?
   b) How do the average number of items and test times change in the OFII-C and OFII paper-pencil form online application (OFII-PPOA)?
   c) Is there a significant relationship between the θ levels obtained by the students from the OFII-C application and the scores they received from the OFII PPOA?

Method

Information regarding the research model, research group, data collection tool, data collection process, and data analysis are presented in this section.

Research Groups

There were two different research groups which took part in this study. The first research group was the one with whom the OFII-PP application was carried out, and the data based on the post-hoc simulation application was obtained. In this group, there were 1425 students from the 10th, 11th or 12th grade studying at different types of high schools located in Mersin, Turkey during the 2018-2019 academic year. The second research group is the group in which both OFFI-PPOA and OFFI-C applications were carried out. In this group, there are 150 students studying in the 10th, 11th and 12th grades of different types of high schools in Mersin. Data were collected from this research group in April 2020.

Data Collection Tools

The OFII-PP developed by Deniz (2009) was used in the first data collection phase of this research. In the second phase, OFII-PPOA (which is the computerized version of OFII) and OFII-C, which was developed as a result of post-hoc simulations, were used. The OFII is an inventory aimed at assisting individuals in selecting an occupation. To develop OFII, first, by making use of the university's educational programs and literature, 14 occupational fields (i.e., computer, law, health, psychology, mathematics, literature, visual arts, foreign language, political sciences, science, communication, education, agriculture, and engineering) were reviewed, and 25 items each focused on measuring the interest of students were written. The items were rated on a 5-point Likert scale: (1) I find little interest, and (5) I find it very interesting (Deniz, 2009). To analyze the validity and reliability of the inventory, it was applied to 1373 students studying at 10 high schools in Ankara, Turkey. Thus, to determine the validity of the inventory, exploratory factor analysis was applied on 1373 students and confirmatory factor analysis was applied to a data group of 216 randomly selected students from the original group of 1373. To determine the reliability of the inventory, Cronbach alpha internal consistency coefficient was calculated with the data obtained from two separate groups of 673 and 700 people, which were determined from the original group of 1373. In addition, the inventory was reapplied to 109 students

selected from among the group of 1373 for whom the inventory was applied, and the test-retest reliability coefficient was calculated (Deniz, 2009).

To determine the content validity of OFII, expert opinion on the content validity of the sub-scales of OFII was obtained from 88 academics who work in various occupational fields at a variety of universities and have earned at least a doctorate degree within their field of expertise. Following the content validity, exploratory factor analysis was applied to determine the construct validity of OFII. As a result of confirmatory factor analysis performed using 14 subscales obtained from the exploratory factor analysis, 11 or 12 items with high factor load values belonging to the subscale were determined and a final version of the inventory consisting of 156 items was obtained. Thus, as a result of the confirmatory factor analysis, it was determined that the goodness of fit indexes of the subscales (CFI, GFI, NNFI, and AGFI) were above .90, except for the AGFI value (.87) of the science subscale. Similarly, in terms of RMSEA values, RMSEA values below .08 were obtained for all factors, except for the science subscale (.087). In addition, correlation coefficients were calculated to determine the relationships between subscales and values ranging from -.43 to .50 were obtained (median: -.07). The fact that the majority of the correlation coefficients obtained had negative values indicated that the subscales sufficiently diverged from each other. Another application carried out to determine the construct validity of the inventory was to calculate the correlation between the scores that students directly gave to their occupation names, varying between 1 and 9, and the scores they got from the relevant sub-scale of the inventory. As a result of these calculations, it was seen that there were significant positive correlations between .49 - .80 (Deniz, 2009).

To estimate the reliability of the inventory, Cronbach's alpha reliability coefficient was calculated for each subscale. Cronbach's alpha internal consistency coefficients for the subscales of OFII were found to range from between .79 and .95 (Median: .88). In addition, the test-retest reliability of the inventory was determined by re-administering the inventory to the same participants eight weeks later. After these applications, it was determined that the test-retest reliability coefficients of the subscales of OFII varied between .75 and .95 (Deniz, 2009).

Data Collection Process

In the scope of this research, an individualized form of the OFII was developed within a computer environment. The data obtained from the first research group were collected from seven high schools accessible to the researcher. Before the data were collected, permission was obtained from the Turkish Ministry of National Education (MoNE) and data were collected with the help of psychological counselors working within the schools. The data collection process was carried out in the classroom environment and each student was provided a booklet containing 156 items belonging to the OFII along with an answer sheet, and they were asked to only answer using the answer sheet. Participation in the research was strictly voluntary and students were informed that if they participated in the research, their individual results would be shared with them at a later date. After the obtained data were analyzed, the occupational field interest profiles of each student were sent to the guidance counselors within their schools and the results shared with the individual students. As a result, it was observed that the students carefully examined their occupational field interest profiles as well as shared and discussed the results with their friends.

As part of the first application, the item parameters of the OFII according to GRM and GPCM were obtained using data obtained from the OFII-PP form. Then, based on the item parameters obtained from the application, a post-hoc simulation was carried out for the CAT form via the Firestar (Choi & Swartz, 2009) software. Thus, the CAT simulation was carried out separately for each subscale, and correlation coefficients were calculated to determine the relationship between the average number of items in each subscale, the mean standard error values, θ levels estimated for all items, and θ levels estimated as a result of the simulation. According to the results, the most suitable IRT model to be used in the OFII-C application was determined as the item selection method and test termination rule. In the second phase of the research, the OFII-C application was developed on the Concerto platform and applied to 150 students online at www.meslekialanilgienvanteri.com. Similarly, it was applied to the same students

using the Google Survey Application for the OFII-PPOA. To prevent rank effect, a 15-day period wait period was carried out between the applications and 75 students who had taken the OFII-PPOA in the first application received the OFII-C within the second application.

Analysis of Data

In the analysis of the current research data, IBM SPSS Statistics 20.00, LISREL 8.51, R, Firestar and PARSCALE software were used.

Examining Assumptions

First, the one-dimensionality assumption, which is a basic assumption of the IRT, was examined through confirmatory factor analysis due to the factor structure of OFII being predetermined. Confirmatory factor analysis was performed with LISREL 8.51 software separately for each dimension and the assumptions of the analysis were checked prior to the factor analysis. Goodness-of-fit indices of the confirmatory factor analysis which were performed separately for each subscale are presented in Table 1.

Table 1

Confirmatory Factor Analysis Fit Indices Applied for the One-Dimensional Assumption

Subscale	χ^2	sd	p	χ^2/sd	RMSEA	SRMR	AGFI	NFI
Computer	58.35	44	.00	1.32	.043	.039	.93	.97
Law	74.62	44	.00	1.70	.056	.051	.91	.95
Health	62.42	44	.00	1.42	.044	.042	.92	.96
Psychology	71.88	44	.00	1.63	.053	.044	.90	.95
Math	73.39	44	.00	1.67	.055	.048	.91	.93
Literature	95.46	44	.00	2.17	.072	.065	.88	.92
Visual arts	97.63	44	.00	2.22	.073	.067	.88	.92
Foreign language	89.52	54	.00	1.66	.057	.054	.90	.94
Political science	106.48	54	.00	1.97	.065	.061	.89	.92
Science	111.04	44	.00	2.52	.082	.079	.86	.90
Communication	63.17	44	.00	1.44	.045	.040	.91	.96
Education	59.82	44	.00	1.36	.044	.040	.91	.94
Agriculture	84.25	44	.00	1.92	.067	.063	.89	.93
Engineering	74.33	44	.00	1.69	.056	.052	.92	.95

In Table 1, the goodness of fit indices obtained for each subscale of the OFII were evaluated according to the criteria determined by Schermelleh-Engel et al. (2003). As a result of the confirmatory factor analysis performed for each subscale of the OFII, χ^2/sd values was found below 3 for all subscales, indicating that the data fit perfectly with the model. In addition, it was seen that the RMSEA value for only the science subscale did not show a good fit. Although the RMSEA value obtained for the science subscale was above 0.80, the 0.90 confidence interval of the RMSEA value for the science subscale indicated that the RMSEA value of 0.082 was acceptable. Thus, according to the results obtained, it can be stated that each subscale of the OFII provided an assumption of unidimensionality.

After testing the unidimensionality assumption, the invariance of the item parameters were tested. For this purpose, two different data groups consisting of 500 people were created randomly from the data set of 1425 people. For each data group created, first of all, the item parameters were calculated using the PARSCALE software, then the relationship between the item parameters calculated for both groups were determined by calculating the Spearman Rank Differences Correlation Coefficient due to the scarcity of items in the subscales. In the next step, the items in each subscale were divided into two groups, and the relationship between the students' θ values estimated according to the items in both groups were determined using the Pearson Product Moments Correlation Coefficient.

The findings regarding the invariance of the θ estimations and item parameters for each subscale of the OFII are presented in Table 2 for both GPCM and GRM. Since the θ estimations and a parameter gave

close values for the GRM and GPCM models, only the position parameter was calculated according to the different IRT models.

Table 2

Findings on the Invariance of θ Estimates and Item Parameters

	r_a	r_L		r_θ
Subscales		GPCM	GRM	
Computer	.69	.82	.82	.43
Law	.82	.94	.94	.56
Health	.83	.94	.93	.57
Psychology	.91	.96	.95	.49
Math	.81	.91	.92	.46
Literature	.79	.92	.92	.61
Visual arts	.84	.94	.95	.67
Foreign language	.73	.88	.89	.52
Political science	.84	.95	.94	.57
Science	.59	.82	.81	.46
Communication	.76	.89	.89	.41
Education	.68	.86	.86	.65
Agriculture	.84	.92	.93	.55
Engineering	.72	.87	.89	.48

All correlation coefficients provided in Table 2 were found to be significant ($p < .05$). Accordingly, it can be stated that the item parameters and θ estimations showed the invariance feature.

Data Analysis for Post-hoc Simulation

After the item parameters were determined, the appropriate syntax was created for the simulation to be carried out in the R software using Firestar (Choi, 2009). While performing the simulations, GRM and GPCM were used as the IRT model. In the first item selection, θ = 0.00 was determined and MEPV, MEI, MEPWI, MFI were used as the item selection method. Standard error values of 0.30, 0.40, and 0.50 were preferred, provided that at least three items were used as the test termination rule. While the range [-3,3] was determined as the θ interval, the θ increment was determined as 0.10. The BS (EAP) was preferred as the θ estimation model. The mean distribution was determined as 0.00 and the standard deviation was 1.00 as the a priori distribution as well as the posterior distribution was preferred as the standard calculation method. While item use control was not carried out, the scaling value was determined as D = 1.7. As in this research, while deciding on the specified simulation conditions, studies were used in which the CAT form of an affective measurement tool was developed and successful results were obtained (Aybek & Çıkrıkçı, 2018; Şimşek, 2017). In addition, the limitations of the Concerto application, in which the OFII-C application is carried out, were considered.

According to the simulation result, for each subscale, the average number of items the application ended with was determined as well as the average standard error and correlation coefficients between the θ values obtained as a result of the simulation and the θ values obtained from the whole test were obtained. In addition, according to the results obtained, it was decided which test termination rule, item selection method, and IRT model were most suitable for the OFII-C.

Data Analysis for OFII-C Application

The Pearson Product-Moment Correlation Coefficient was used to calculate the correlation between the θ levels estimated from the OFII-C application and the scores obtained from the OFII-PPOA application. Furthermore, frequency of item use was determined, and frequency analysis was performed for the items used. After calculating the correlation coefficients and the frequency of item use, the OFII-PPOA scores and the OFII-C estimations were provided in the same graph as a way of comparing the occupational field interest profiles obtained from the OFII-PPOA and OFII-C applications. For this purpose, the raw scores obtained by the students from the OFII-PPOA form were converted into standard z scores.

Thus, to determine whether the OFII-PPOA and OFII-C profiles matched, the relationship between the θ levels obtained by the students in the second research group of 150 people from the OFII-C application and the scores they obtained from the OFII-PPOA application were determined by calculating the Pearson correlation coefficient.

Results

The first sub-objective of this research was to determine the mean number of items used, the mean standard error values obtained, and the mean standard error values obtained in the post-hoc simulations performed using different IRT models (GPCM and GRM) along with different test termination rules (0.30-0.40-0.50 standard error). The correlations between the θ estimates are presented in Table 3.

Table 3

GRM and GPCM Values for OFII .30, .40 and .50 Standard Error Test Termination Rules

		k		SEM		r	
	Subscales	GPCM	GRM	GPCM	GRM	GPCM	GRM
SEM= 0.30	Computer	10.12	11.00	.32	.41	.99	1.00
	Law	9.52	11.00	.31	.32	.91	1.00
	Health	10.00	11.00	.31	.31	.98	1.00
	Psychology	11.00	11.00	.33	.39	1.00	1.00
	Math	8.43	11.00	.31	.31	.81	1.00
	Literature	9.41	11.00	.31	.47	.88	1.00
	Visual arts	11.00	11.00	.33	.35	1.00	1.00
	Foreign Language	10.34	12.00	.32	.36	.99	1.00
	Political science	7.86	11.45	.30	.38	.66	.99
	Science	10.05	11.00	.32	.39	.98	1.00
	Communication	6.28	10.57	.30	.40	.56	.99
	Education	6.71	10.52	.30	.45	.59	.99
	Agriculture	10.20	11.00	.32	.33	.99	1.00
	Engineering	8.27	11.00	.30	.35	.79	1.00
SEM= 0.40	Computer	4.80	9.15	.38	.43	.94	.98
	Law	4.12	8.27	.36	.40	.93	.97
	Health	4.20	8.37	.37	.41	.93	.97
	Psychology	5.25	8.12	.38	.41	.96	.99
	Math	3.53	7.48	.36	.41	.91	.96
	Literature	3.81	8.06	.36	.40	.92	.96
	Visual arts	5.00	9.93	.38	.41	.96	.98
	Foreign Language	5.53	9.74	.38	.41	.95	.98
	Political science	3.29	6.87	.35	.39	.91	.95
	Science	4.30	8.50	.37	.39	.94	.98
	Communication	3.18	6.31	.35	.38	.90	.95
	Education	3.22	6.51	.35	.39	.91	.94
	Agriculture	5.11	9.29	.38	.42	.94	.98
	Engineering	3.41	7.11	.36	.40	.91	.96
SEM= 0.50	Computer	4.54	8.70	.41	.46	.90	.95
	Law	3.93	7.16	.39	.44	.89	.93
	Health	3.97	7.75	.40	.45	.89	.93
	Psychology	5.17	9.95	.43	.48	.92	.96
	Math	3.17	6.76	.39	.44	.89	.93
	Literature	3.29	6.96	.39	.44	.89	.93
	Visual arts	4.81	9.81	.43	.48	.91	.95
	Foreign Language	4.32	9.56	.43	.48	.91	.95
	Political science	3.15	6.28	.40	.46	.88	.93
	Science	4.00	8.20	.40	.46	.90	.94
	Communication	3.01	6.05	.37	.43	.88	.92
	Education	3.00	6.16	.38	.43	.88	.92
	Agriculture	4.57	9.09	.42	.46	.90	.94
	Engineering	3.15	6.58	.38	.43	.88	.93

Thus, according to the post-hoc simulation results, when a standard error value of .30 was preferred as the test termination rule, it was seen that the mean standard error was above this value in all subscales. Whereas when a standard error value of .40 was used as the test termination rule, it was determined that an average standard error of over .40 was obtained for the subscales of GRM except in the subscales of political sciences, science, communication, and education. However, it was determined that GPCM had a standard error of less than .40 in each subscale as well as this occurred by using approximately 4.2 items. When the standard error value of .50 was preferred as the test termination rule, an average standard error value of less than .50 was obtained for all subscales in both the GRM and GPCM.

As a result of the calculation made using the data obtained from the OFII-PP application, it was determined that the Cronbach's alpha internal consistency coefficients of the subscales ranged between .81 and .94. When the .40 standard error was selected as the test termination rule, the average of the reliability coefficients of the reliability scales became .84.

According to the results of the post-hoc simulation research, a standard error of .40 was found to be more appropriate as a test termination rule than other rules, and as a result, the decision was made to use the .40 standard error as test termination rule in the OFII-C application. In addition, according to the simulation studies carried out, it was seen that GPCM achieved similar results with fewer items than GRM. Also, it was determined that GPCM used approximately 62% fewer items than the 156 items in the original form as well as provided feature estimation with an error of less than .40. Due to all of these reasons, the decision was made to use GPCM as the IRT model in the OFII-C application.

Thus, with a .40 standard error value as the test termination rule, the GPCM and MFI, MEI, MEPV, and MEPWI item selection methods as the IRT model, the correlations between all test-simulation θ estimations obtained as well as the standard error values and average number of items used are presented in Table 4.

Table 4

Findings According to Item Selection Methods According to 0.40 Standard Error Value as Test Termination Rule

Subscale	MFI			MEI			MEPV			MEPWI		
	k	SEM	r	k	SEM	r	k	SEM	r	k	SEM	r
Computer	4.80	.38	.94	4.82	.38	.94	4.83	.38	.94	4.84	.38	.94
Law	4.12	.36	.93	4.12	.36	.93	4.13	.36	.93	4.13	.36	.93
Health	4.20	.37	.93	4.20	.37	.93	4.21	.37	.94	4.22	.37	.94
Psychology	5.25	.38	.96	5.27	.37	.94	5.26	.37	.94	5.27	.37	.94
Math	3.53	.36	.91	3.54	.36	.92	3.53	.36	.92	3.54	.36	.92
Literature	3.81	.36	.92	3.81	.36	.92	3.80	.36	.92	3.80	.36	.92
Visual arts	5.00	.38	.96	5.02	.38	.96	5.02	.38	.97	5.02	.38	.97
Foreign Language	5.53	.38	.95	5.53	.38	.94	5.52	.38	.94	5.53	.39	.95
Political science	3.29	.35	.91	3.28	.35	.93	3.27	.35	.94	3.28	.35	.94
Science	4.30	.37	.94	4.31	.37	.94	4.32	.37	.94	4.34	.37	.94
Communication	3.18	.35	.90	3.18	.35	.90	3.19	.35	.90	3.19	.35	.90
Education	3.22	.35	.91	3.22	.35	.92	3.21	.35	.92	3.21	.35	.92
Agriculture	5.11	.38	.94	5.12	.36	.94	5.11	.36	.94	5.13	.37	.94
Engineering	3.41	.36	.91	3.43	.36	.92	3.42	.37	.92	3.43	.37	.92

When Table 4 is examined, it can be seen that different item selection methods did not cause a significant change in the correlation coefficients between the θ levels (all-θ) estimated using the entirety of the items and the θ levels (sim-θ) estimated by simulation as well as the standard error values or average number of items applied. Therefore, in the OFII-C application, the MFI method was preferred as the appropriate item selection method.

In addition, when GPCM was preferred as the IRT model, a .40 standard error as the test termination rule, and MFI preferred as the item selection method, it was seen that the OFII-C simulation ended with an average of 59 items. Considering that the OFII-PP consisted of 156 items, it can be stated that

approximately 62% less items were used with the OFII-C simulation. Furthermore, the correlation coefficients between the sim-θ and all-θ were determined to be at values between .91 and .97.

For the second sub-purpose of this research, the students' data for the OFII-C application were taken from a database of the website created by the researcher for the OFII-C application. Using the data obtained from that database, the frequency of use of each item was determined. Similarly, the frequency of use of the items in the post-hoc simulation application were also determined, and the frequency of use of the items according to both applications are compared in Table 5.

Table 5

Subscales of OFII Item Use Frequencies for OFII-C and OFII-C Simulation Applications

			Comp.	Law	Heal.	Psy.	Math	Lit.	Vis.	Fore.	Pol.	Sci.	Com.	Edu.	Agri.	Eng.
Item 1	Live	%	58.5	48.1	52	62.5	35.1	69.5	100	0	0	100	14	14.4	39.5	11.7
	Sim	%	62.6	44.4	43.6	53.2	25.6	60	81.6	5.7	2.5	82.5	11.3	11.2	32.7	9.2
Item 2	Live	%	12.3	56.2	74.5	54.7	43.4	55.9	12.4	69.5	0	31.7	17.1	58.3	14.3	16.2
	Sim	%	25.2	60	60.2	61.5	40.7	55.1	9.5	60.2	8.5	26.4	14.3	47.9	11.1	15.78
Item 3	Live	%	52.3	20.3	14.3	39.2	25.6	17.2	15.6	75.6	0	25.0	25.1	60.1	33.7	35.6
	Sim	%	47.8	30.3	22.5	35.5	22.2	20.5	11.4	71.2	6.1	22.0	18.5	55.2	27.5	30.1
Item 4	Live	%	24.6	100	21.3	100	100	0	33.8	0	100	17.5	100	100	42.4	100
	Sim	%	17.8	86.5	20.3	84.7	89.5	4.6	31.2	4.2	76.5	19.3	92.3	85.2	38.9	89.1
Item 5	Live	%	15.6	15.1	19.0	28.3	0	42.0	46.7	17.3	48.7	23.4	0	25.6	0	25.2
	Sim	%	20.4	20.2	14.0	22.6	7.33	50.6	40.1	15.3	39.5	27.6	6.2	30.8	8.5	28.9
Item 6	Live	%	100	91.2	5.29	14.5	0	100	0	43.6	57.8	52.2	13.4	12.5	100	14.4
	Sim	%	87.2	85.1	7.25	18.9	4.5	92.9	6.6	37.4	65.1	45.8	17.7	17.5	89.5	19.5
Item 7	Live	%	45.7	18.4	12.2	12.3	26.7	0	39.0	26.7	0	24.5	0	24.7	9.25	0
	Sim	%	51.6	15.0	15.39	15.1	28.5	3.4	34.6	23.4	11.4	32.7	10.1	33.0	14.3	5.9
Item 8	Live	%	72.0	32.7	23.6	14.6	13.5	12.4	7.6	0	0	20.2	60.5	11.5	39.5	49.1
	Sim	%	65.2	29.5	20.9	20.0	16.1	15.3	13.9	6.9	5.9	25.5	54.6	9.2	32.2	53.0
Item 9	Live	%	0	0	11.2	30.3	0	0	14.7	0	61.2	13.3	32.5	0	44.1	39.8
	Sim	%	5.08	7.58	8.83	27.6	5.5	7.8	12.3	2.5	63.5	10.4	25.1	2.5	35.4	34.1
Item 10	Live	%	33.6	27.5	65.5	0	49.5	74.2	51.6	43.7	0	17.1	45.1	45.8	66.1	43.8
	Sim	%	43.5	31.0	46.1	6.54	45.7	70.0	57.3	35.6	8.2	20.5	33.4	38.2	70.3	38.5
Item 11	Live	%	33.5	22.4	100	51.3	61.0	65.9	67.9	0	60.2	52.5	54.5	41.4	12.5	20.0
	Sim	%	37.0	19.3	92.2	45.0	67.3	54.1	59.4	6.2	65.8	42.4	47.2	43.8	9.3	23.3
Item 12	Live	%								31.5	54.3					
	Sim	%								25.8	50.2					

Live: OFII-C Sim: OFII-C Simulation

When Table 5 is examined, it can be seen that all the items were used in the post-hoc simulation, but that some items were not used in the OFII-C application. As a result, items not used in the OFII-C application were the ninth item for the computer factor; ninth item for the law factor; tenth item for the psychology factor; fifth, sixth, and ninth items for the mathematics factor; fourth, seventh, and ninth items for the literature factor; sixth item for the visual arts factor; first, fourth, eighth, ninth, and eleventh items for the foreign language factor; first, second, third, seventh, eighth, and tenth items for the political sciences factor; fifth and seventh items for the communication factor; ninth item for the education factor; fifth item for the agriculture factor; and seventh item for the engineering factor. Thus, it was seen that the frequency of use of a majority of the items which were never used in the OFII-C application was below 10% within the post-hoc simulation. In addition, it can be seen that the α parameters of the items which were never used, generally belonged to items with the lowest coefficient in each factor. When the

data of the OFII-C application was examined, it could be seen that one item in each factor was directed to all the participants. Since the $\theta = 0$ was chosen as the initial value of the test, the starting material was the same for all participants and the frequency of use of these items was determined to be 100%. In other words, in the OFII-C application, it was determined that one item in each factor was directed to all the participants. At the same time, the frequency of use of these items in the OFII-C simulation was over 80%.

Thus, to compare the average number of items used in the OFII-C application and the OFII-PPOA, the average of the number of items answered by each participant for each factor was determined through the OFII-C application. The average number of items used in the OFII-C application and the OFII-PPOA is presented in Table 6.

Table 6

Number of Items Used in OFII-C and OFII-PPOA and Test Durations

	OFII-C				OFII-PPOA	
Subscale	Minimum Number of Items	Maximum Number of Items	Average Number of Items	Average Test Time (minutes)	Average Number of Items	Average Test Time (minutes)
Computer	3	5	3.22	.29	11	.92
Law	3	5	3.35	.31	11	.95
Health	3	5	3.14	.28	11	1.02
Psychology	3	7	3.43	.31	11	.91
Math	3	6	3.74	.33	11	.94
Literature	3	5	3.42	.30	11	.88
Visual arts	3	6	3.53	.34	11	.90
Foreign Language	3	7	4.14	.36	12	1.10
Political science	3	7	4.19	.37	12	1.15
Science	3	7	3.52	.31	11	.99
Communication	3	5	3.34	.30	11	.85
Education	3	5	3.72	.33	11	.96
Agriculture	3	7	3.45	.31	11	1.00
Engineering	3	6	3.63	.32	11	.97
Total			53.49	4.46	156	13.51

When Table 6 is examined, it can be seen that the least used item subscale in the OFII-C application was the health subscale, and that the most used item subscale was the political sciences subscale. In the OFII-C application, each participant responded to an average of 53.49 items. Since the OFII-PPOA consisted of 156 items, it was stated that 65.71% less items were used with the OFII-C. While the participants answered the OFII-C within an average of 4.46 minutes, the average response time for the OFII-PPC was 13.51 minutes. In this respect, it could be stated that the application time of the OFII decreased by 66.99% with the CAT application.

The Pearson correlation coefficients were calculated to determine whether there was a significant relationship between the θ levels estimated by the students within the OFII-C application and the scores they had obtained from the OFII-PPOA. Thus, the results of the correlation coefficients are presented in Table 7.

Table 7

The Correlations Between Levels of θ Estimated from OFII-C Form and Scores Obtained from OFII-PPOA

Subscale	r	p
Computer	.88	.00
Law	.83	.00
Health	.92	.00
Psychology	.82	.00
Math	.79	.00
Literature	.91	.00
Visual arts	.85	.00
Foreign Language	.78	.00
Political science	.84	.00
Science	.74	.00
Communication	.79	.00
Education	.81	.00
Agriculture	.73	.00
Engineering	.76	.00

When Table 7 is examined, it can be seen that the correlation coefficients calculated for the relationship between the OFII-C and OFII-PPOA were significant for all the subscales and that the highest correlation coefficient was 0.92 for the health subscale, while the lowest correlation coefficient was 0.73 for the agriculture subscale. The median value of the obtained correlation coefficients was found to be 0.82. As a result, it can be stated that there were highly significant relationships between the OFII-C and OFII-PPOA for all subscales.

Discussion and Conclusion

The most important difference of the CAT applications from paper-pencil applications is that the number of items directed to each person differs. This is due to the test termination rule, which is one of the basic components of CAT applications. For example, according to the test termination rule, following each item answered by a test taker, it is determined whether the test should be terminated or continue (Hambleton et al., 1991). Although there are many different options which make up the test termination rule in CAT applications, the most preferred rule is the use of the standard error value. In the current research, the standard error criterion was used as a rule for terminating the test. There is a negative relationship between the standard error and measurement precision. As the standard error increases, the measurement precision decreases. Thus, by making use of the relationship between measurement accuracy and standard error, the standard error criterion is determined to obtain the desired measurement precision (Özkan, 2014). As a result, in cases where the standard error criterion is applied as the test termination rule, when a participant answers an item, it is determined whether the calculated standard error value is less than the determined critical value. If the standard error value calculated with this method is less than the standard error value determined as the critical threshold, the CAT application is terminated. It is revealed in several past studies, (Babcock & Weiss, 2012; Eroğlu & Kelecioğlu, 2015; Gnambs & Batinic, 2011; Stochl et al., 2016), that the test termination rule is one of the most important CAT components which directly affects test length. When the literature was examined, it was recognized that the standard error criteria of 0.30, 0.40, and 0.50 were most widely used in studies (Aybek & Çıkrıkçı, 2018; Özbaşı & Demirtaşlı, 2015; Şimşek 2017) which investigated the adaptability of affective measurement tools for CAT. Therefore, a standard error criteria of 0.30, 0.40, and 0.50 were also used in this research. Thus, according to the results of the post-hoc simulation research, the .40 standard error was found to be more appropriate as a test termination rule than other rules, and as a result, the decision was made to use the .40 standard error as the test termination rule for the OFII-C application. In addition, according to the simulation studies carried out, it was seen that GPCM achieved similar results with fewer items than GRM. Furthermore, it was ultimately determined that GPCM used approximately 62% fewer items than the 156 items from the original form as well as gave feature

estimation with an error of less than .40. For the reasons just discussed, the determination was made to use GPCM as the IRT model for the OFII-C application.

In CAT applications, a variety of methods are used to determine which items are directed to the individual test taker following the selection of the starting material. For example, Van der Linden (1998) states that if one of the MEI, MEPV or MEPWI methods is selected in CAT applications, the θ estimation will be more reliable than other methods. It is also recommended by Boyd et al. (2010) that MFI, MEI, MEPV, and MEPWI methods be preferred as the item selection methods in CAT applications. Therefore, the MFI, MEI, MEPV, and MEPWI methods were the preferred item selection methods for this research. It has been recognized that different item selection methodologies do not cause a significant change in correlation coefficients, standard error values or the average number of items applied between θ levels estimated using all items (all-θ) and θ levels estimated through simulation (sim-θ). These findings were in line with the finding from Choi and Swartz (2009) using the CTM model, which the estimated θ level and number of items used in cases where the item pool is small do not differ according to the item selection method. Also, Veldkamp (2003) states that although different item selection methods are used, the same items are found at a rate between 85% to 100%. While Aybek & Çıkrıkçı (2018) used MEPWI, MEI, MFI, and MEPV item selection methods, and find that item selection methods do not have a significant effect on the estimated θ level, standard error values, and number of items used. In the current research, in accordance with the literature, it was determined that different item selection methods under both the GPCM and GRM models did not cause a significant change in the estimated θ level, standard error values, and the number of items used. Therefore, in the OFII-C application, the MFI method was the preferred item selection method.

When GPCM was preferred as the IRT model, .40 standard error as the test termination rule, and MFI as the preferred item selection method, it was seen that the OFII-C simulation ended with an average of 59 items. Considering that the OFII-PP consists of 156 items, it can be stated that approximately 62% less items were used with the OFII-C simulation. In addition, it was determined that the correlation coefficients between sim-θ and all-θ gained values between .91-.97. In Scullard (2007), an investigation of the adaptability of the Strong Interest Inventory, a measurement tool similar to OFII for individuals in the computer environment, reached the same findings obtained from our research. The correlation coefficient obtained in Scullard (2007) ranged from .90 to .98 between the sim-θ and all-θ estimations and the test length decreased by approximately 60%. The fact that many researchers (Betz & Turner, 2011; Chien et al., 2011; Gibbons et al., 2012; Hol et al., 2007; Smits et al., 2011) obtained similar findings to those obtained in the current research highlights the reliability of the findings.

The OFII-C was developed through the Concerto program, using GPCM as the IRT model, .40 standard error value as the test termination rule, and MFI as the item selection method, which are the most suitable criteria for OFII-C. The OFII-C and OFII-PPOA were applied to 150 high school students. While all items were used in post-hoc simulations, it was determined that some items were not used in the OFII-C application. It was also observed that the majority of the items which were never used in the OFII-C application were the items with a very low frequency of use in the post-hoc simulation. In addition, it was determined that the α parameters of the items which were never used, generally belonged to the items with the lowest coefficient in each factor.

It was determined that approximately 66% less items were used in the OFII-C application compared to the OFII-PPOA, and that the OFII-C was 67% more advantageous in terms of time. When the literature was examined, it was found that there were findings similar to the current research regarding the OFII-C application ending with fewer items as well as a shorter time period compared to the OFII-PPOA application. For example, Hol et al. (2007) adapted the measurement tool from a paper-pencil format to a CAT format and had a 62.50% decrease in the number of items. While Gibbons et al. (2012) showed a 95% decrease in the number of items for their 626-item measurement tool adapted to the CAT format. Also, when Kocalevent et al. (2009), adapted a 104-item measurement tool to the CAT format, the decrease in the number of items was 85%. Whereas Aybek and Çıkrıkçı (2018) adapted their measurement tool to the CAT format and had a 52% decrease in the number of items. In addition, Şimşek (2017) adapted the measurement tool in the CAT format and had a 50% decrease in the number of items. Thus, in a variety of other studies, it can be seen that the CAT application saves between 50-70% in the

number of items used as well as an overall decrease in test time (Betz & Turner, 2011; Bulut & Kan, 2012; Choi et al., 2010; Cömert, 2008; İşeri, 2002; Jodoin et al., 2006; Kalender, 2012; Kezer & Koç, 2014; McDonald, 2002; Öztuna, 2008; Rezaie & Golshan, 2015; Scullard, 2007; Smits et al., 2011; Weiss, 2011).

It was ultimately determined that there were highly significant relationships between OFII-C and OFII-PPOA for all subscales, and although correlation coefficients varying in the range of 0.91-0.97 were obtained between θ levels obtained from the OFII-PP and OFII-C simulation applications; the correlation coefficients varying in a range of 0.73-0.91 were obtained between the θ levels obtained from the OFII-C and OFII-PPOA applications. Thus, in this case, it can be stated that although the relationship between the OFII-C and OFII-PPOA was high, it was relatively lower than the correlation coefficients obtained from the OFII-PP and OFII-C simulation studies. In addition, when the literature was examined, it was seen that the correlation coefficients obtained in post-hoc simulation studies were higher than in the CAT studies (Achtyes et al., 2015; Aybek & Çıkrıkçı, 2018; Betz & Turner, 2011; Gibbons et al., 2012; Simms & Clark, 2005; Stochl et al., 2016).

As a result of this study, CAT form of OFI was developed successfully within the limitations of the research. The simulation phase of the research was limited to 0.30, 0.40, 0.50 standard error values, FEYB, BEYB, BEDSV, BEYSAB item selection methods and GKPM and KTM models due to the program used. At the same time, existing subscales of OFI were used while creating the CAT form and no new subscales were added. In line with these limitations, researchers can be recommended to develop CAT form of OFI by using different standard error values, item selection methods, IRT models, and adding new professional interests that have emerged in accordance with the technological developments of our age.

Declarations

Author Contribution: Volkan Alkan: conceptualization, investigation, methodology, data curation, supervision, writing - review & editing. Kaan Zülfikar Deniz: conceptualization, methodology, writing - original draft, formal analysis, visualization.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Ethical Approval: The study was ethically approved by the Ministry of National Education (research number: 81576613/605.01/5603857, dated 18/03/2019). In addition, this study was found ethically appropriate with the decision of Ankara University Rectorate Ethics Committee numbered 56786525-050.04.04/49481. This study has been produced from the dissertation of Volkan Alkan that was conducted under the supervision of Prof. Dr. Kaan Zülfikar Deniz.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Achtyes, E. D., Halstead, S., Smart, L., Moore, T., Frank, E., Kupfer, D. J., & Gibbons, R. D. (2015). Validation of computerized adaptive testing in an outpatient nonacademic setting: The vocations trial. *Psychiatric Services*, 1-6. https://doi.org/10.1176/appi.ps.201400390

Akar, C. (2012). Factors affecting university choice: A study on students of economics and administrative sciences. *Journal of Eskişehir Osmangazi University Faculty of Economics and Administrative Sciences, 7*(1), 97-120.

Altın, M. (2020). Education, status and social mobility in Turkey. *Mecmua, 10*, 180-196. https://doi.org/10.32579/mecmua.789249

Aybek, E. C., & Çıkrıkçı, R. N. (2018). Applicability of self-assessment inventory as an individualized test in computer environment. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, *8*(50), 117-141.

Babcock, B., & Weiss, D. (2012). Termination criteria in computerized adaptive tests: Do variable - length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*(1), 1-18. https://doi.org/10.7333/1212-0101001

Betz, N. E., & Turner, B. M. (2011). Using Item Response Theory and Adaptive Testing in Online Career Assessment. *Journal of Career Assessment*, *19*(3), 274–286. https://doi.org/10.1177/1069072710395534

Boyd, A., Dodd, B., & Choi, S. (2010). Polytomous models in computerized adaptive testing. Nering, M., & Ostini, R. (Ed.). *Handbook of polytomous item response theory models*. (229-255). Routledge.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research*, (49), 61–80.

Chien, T.-W., Lai, W.-P., Lu, C.-W., Wang, W.-C., Chen, S.-C., Wang, H.-Y., & Su, S.-B. (2011). Web-based computer adaptive assessment of individual perceptions of job satisfaction for hospital workplace employees. *BMC Medical Research Methodology*, *11*(1), 1-8. https://doi.org/10.1186/1471-2288-11-47

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program forpolytomous item response theory models. *Applied Psychological Measurement*, *33*(8), 644–645.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement*, *33*(6), 419-440. https://doi.org/10.1177/0146621608327801

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, *19*(1), 125–136. https://doi.org/10.1007/s11136-009-9560-5

Cömert, M. (2008). *Development of computer-aided assessment and evaluation software adapted to the individual*. Unpublished Master's Thesis, Bahçeşehir University Institute of Science and Technology, İstanbul.

Deniz, K. Z. (2009). Occupational Interest Inventory (OFII) development study. Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, *6*(1), 289-310.

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

Eroğlu, M. G., & Kelecioğlu, H. (2015). Comparison of Different Termination Rules in terms of Measurement Accuracy and Test Length in Individualized Computerized Testing Applications. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, *28*(1), 31-52.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. a, Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of General Psychiatry*, *69*(11), 1104-12. https://doi.org/10.1001/archgenpsychiatry.2012.14

Gnambs, T., & Batinic, B. (2011). Polytomous adaptive classification testing: Effects of item pool size, test termination criterion, and number of cutscores. Educational and Psychological Measurement, 71(6), 1006–1022. https://doi.org/10.1177/0013164410393956

Hambleton, R., & Swaminathan, R. (1985). *Fundementals of item response theory.* Sage Publications, Inc.

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

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist*, *27*(3), 353-383.

Hol, M. A., Vorst, H. C. ve Mellenbergh, G. J. (2007). Computerized Adaptive Testing for Polytomous Motivation Items: Administration Mode Effects and a Comparison with Short Forms. *Applied Psychological Measurement*, *31*(5), 412–429. https://doi.org/10.1177/0146621606297314

İşeri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures*. Unpublished Doctoral Dissertation, Middle East Technical University, Ankara.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203–220. https://doi.org/10.1207/s15324818ame1903_3

Kalender, İ. (2012). Computerized adaptive testing for student selection to higher education. *Yükseköğretim Dergisi*, *2*(1), 13-19.

Kezer, F, & Koç, N. (2014). Comparison of Individualized Test Strategies in Computer Environment. *Eğitim Bilimleri Araştırmaları Dergisi*, *4*(1), 145-174. https://doi.org/10.12973/jesr.2014.41.8

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., ... & Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology*, *62*(3), 278-287.

McDonald, P. L. (2002). *Computer adaptive test for measuring personality factors using item response theory*. Unpublished Doctoral Dissertation. The University Western of Ontario, London.

Meyer, J. P. (2010). *Understanding measurement: Reliability*. Oxford University Press.

Ostini, R., & Nering, M. L. (Eds.). (2010). *Polytomous item response theory models.* Taylor and Francis Group.

Özbaşı, D., & Demirtaşlı, N. (2015). Developing the computer literacy test as an individualized test in the computer environment. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(2), 218-237.

Özkan, Y. Ö. (2014). A comparison of estimated achievement scores obtained from student achievement assessment test utilizing classical test theory, unidimensional and multidimensional IRT. *International Journal of Human Sciences*, *11*(1), 20-44.

Öztuna, D. (2008). *Application of computer adaptive testing method in disability assessment of musculoskeletal problems*. Unpublished Doctoral Dissertation. Ankara University Institute of Health Sciences, Ankara.

Rezaie, M., & Golshan, M. (2015). Computer adaptive test (CAT): Advantages and limitations. *International Journal of Educational Investigations*, *2*(5), 128–137.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.

Scullard, M. G. (2007). *Application of item response theory based computerized adaptive testing to the strong interest inventory*. Unpublished Doctoral Dissertation. University of Minnesota, USA.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment*, *17*(1), 28–43. https://doi.org/10.1037/1040-3590.17.1.28

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), 58. https://doi.org/10.1186/s12874-016-0158-7

Şahin, A., & Özbaşı, Ö. (2017). Effects of Content Balancing and Item Selection Method on Ability Estimation in Computerized Adaptive Tests. *Eurasian Journal of Educational Research*, *69*, 21-36.

Şimşek, A. S. (2017). *Adaptation of skills confidence occupational interest inventory and development of computerized individualized testing*. Unpublished Doctoral Dissertation, Ankara University Institute of Educational Sciences, Ankara.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1-23.

Van der Linden, W. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201-216.

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

Yoo, J. H. (2016). The effect of professional development on teacher efficacy and teachers' self-analysis of their efficacy change. *Journal of Teacher Education for Sustainability*, *18*(1), 84–94. https://doi.org/10.1515/jtes-2016-0007