ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(4); 328-344

# Applicability and Efficiency of a Polytomous IRT-Based Computerized Adaptive Test for Measuring Psychological Traits

Ahmet Salih ŞİMŞEK* Ezel TAVŞANCIL**

**Abstract**

Currently, research on computerized adaptive testing (CAT) focuses mainly on dichotomous items and cognitive traits (achievement, aptitude, etc.). However, polytomous IRT-based CAT is a promising research area for measuring psychological traits that has attracted much attention. The main purpose of this study is to test the practicality of the polytomous IRT-based CAT and its equivalence with the paper-pencil version. Data were collected from 1449 high school students (45% female) via the paper-pencil version. The data were used for IRT parameter estimates and CAT simulation studies. For the equivalence study, the research group consisted of 81 students (47% female) who participated in both the paper-pencil and live CAT applications. The paper-pencil version of the vocational interest inventory consists of 17 factors and 164 items. When the EAP estimation method and setting SE < .50 as the termination criterion, better performance was obtained compared with other CAT designs. The Item selection did not help to reduce test duration or increase measurement accuracy. As a result, it was found that an area of interest can be assessed with four items. The results of the live CAT application showed that the estimates of CAT were strongly positively correlated with its paper-pencil version. In addition, the live CAT application increased applicability compared to the fixed-length test version by reducing test length by 50% and time by 77%. This study shows that the polytomous IRT-based CAT is applicable and efficient for measuring psychological traits.

*Keywords:* polytomous item response model, computerized adaptive test, equivalence, efficiency, measurement precision

## Introduction

Likert scales are commonly used measurement tools to measure the psychological characteristics of individuals. Responses are considered valid as long as individuals answer sincerely. However, because the test duration is quite long for some measurement instruments, the person's motivation to respond may decrease, and the validity of the measurements may be negatively affected (Crocker & Algina, 1986; Gardner et al., 2004). This situation, seemingly related only to the usefulness of the measurement instrument, also raises validity issues. Such validity issues can be overcome with the use of technology and the measurement model.

The use of technology has somewhat increased the practicality of fixed-length paper-pencil tests (PPTs). However, non-adaptive computerized tests are not an adequate solution to increase the usefulness of fixed-length tests. The usefulness of measurement instruments can be increased by a computerized adaptive test (CAT) (Achtyes et al., 2015; Reise & Henson, 2000; Simms & Clark, 2005). A CAT application allows for shorter tests with fixed precision (variable length). The superiority of CAT in terms of measurement precision and practicality is enabled by the preferred measurement model.

Both classical test theory (CTT) and item response theory (IRT) are widely used measurement approaches today. However, both models' approaches and mathematical backgrounds for person-item interaction are different. In CTT, the entire set of items must be answered to measure the person's trait. It is possible that this limitation can be overcome by an IRT-based CAT implementation. The

* Assist. Prof. Dr., Kırşehir Ahi Evran University, Faculty of Education, Kırşehir-Türkiye, asalihsimsek@gmail.com, ORCID ID: 0000-0002-9764-3285
** Prof. Dr., Ankara University, Faculty of Education, Ankara-Türkiye, etavsancil@gmail.com, ORCID ID: 0000-0002-8318-2043

To cite this article:
Şimşek, A. S., & Tavşancıl, E. (2022). Applicability and efficiency of a polytomous IRT-based computerized adaptive test for measuring psychological traits. *Journal of Measurement and Evaluation in Education and Psychology, 13*(4), 328-344. https://doi.org/10.21031/epod.1148313

*Received: 25.07.2022*
*Accepted: 11.11.2022*

implementation of CAT allows for a reduction in test length by selecting items that are appropriate for each person. This implies a solution to the validity issues arising from the practicality problem of measurement instruments consisting of a large number of items.

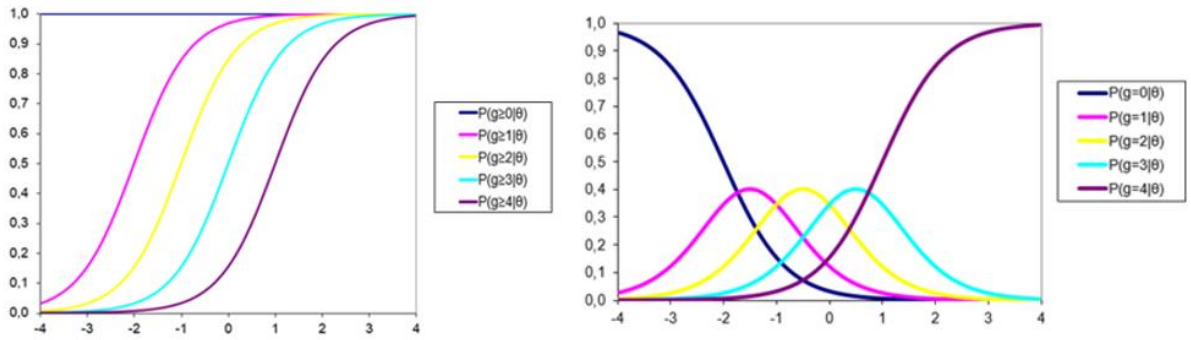
Most of the research on CAT focuses on measuring maximum performance, which mostly consists of dichotomous items (achievement, ability, etc.). However, there are relatively few CAT studies of psychological measurement instruments that require responses to polytomous items (Betz & Turner, 2011; Hol et al., 2007; Reise & Henson, 2000; Vogels et al., 2011). There are currently developed IRT models called polytomous item response theory for polytomous items (Ostini & Nering, 2006). Polytomous item response theory (PIRT) models can be described as IRT models that require responses to items that consist of ordered response categories (Schinka & Velicer, 2003). The PIRT model can be used to measure both maximum performance and psychological constructs. However, it is more commonly used with psychological measurement instruments that contain Likert-type items. One of the main research areas of CAT is the measurement instruments used to assess psychological characteristics. The fact that the PIRT models are mathematically more complex may have made them less suitable for dichotomous items compared to the IRT model (Smits et al., 2011; Waller & Reise, 1989).

The Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM) are the most commonly used PIRT models (Kang et al., 2005; Kang et al., 2009; Wang & Wang, 2002). Generally, the GRM has been favored for fitting rating scale responses (e.g., Likert-type data), whereas the GPCM has been used to score responses to items in cognitive tests (Ren et al., 2020). In the study conducted by Kang et al. (2009) on the bias of PIRT models in parameter estimation, the GRM model was found to outperform the GPCM model for data sets of 1000 or more and for Likert-type items with five points. Studies in the literature support the conclusion that GRM makes better predictions than GPCM (Hol et al., 2007; Smits et al., 2011).

The GRM model developed by Samejima (1969) has the item slope (a) and item position (bg) parameters. Since the item slope parameter is the same for each category, a category-bound characteristic function (CBCF) is created in parallel with the GRM (Fig. 1). This feature means that GRM can be used for sequential equivalent intervals, such as Likert-type items. While the relationship between the probability with which a person selects a response category and θ is modeled with the item-category characteristic curve (ICCC), the dichotomization of polytomous response categories is modeled with the CBCF (Fig. 1).

**Figure 1**

*Example of ICCC (right) and CBCF (left) for a 5-point Likert Item*




CAT design consists of three basic steps: initial theta estimation, item selection, and test termination (Thompson & Weiss, 2011). Both the theta parameter and the standard error of the estimate are updated with each response given by the person. In PIRT models, the item information function is calculated by obtaining the information functions for each category (Ostini & Nering, 2006). The item information function in the GRM is defined as the negative value of the second derivative of the logarithm of the

ICCC (Ostini & Nering, 2006). Thus, the item category information function to represent the g-category threshold for item i is as follows;

$$I_{i_g}(\theta) = -\frac{\partial^2}{\partial\theta^2} log P_{i_g}(\theta) \quad (\#1)$$

Equation (#1) shows the item category bound function (ICBF). The weighted sum of the ICBFs forms the Item Information Function (IIF) (equation #2).

$$I_i(\theta) = \sum_{g=0}^{m}(I_{i_g}(\theta).P_{i_g}(\theta)) \quad (\#2)$$

If equality (#1) and equality (#2) are considered together, the information function can be obtained in its simplest form (equation #3).

$$I_i(\theta) = \sum_{g=0}^{m}\frac{\left(P_{ig}^{*\prime}(\theta) - P_{ig+1}^{*\prime}(\theta)\right)^2}{\left(P_{ig}^{*}(\theta) - P_{ig+1}^{*}(\theta)\right)} \quad (\#3)$$

In this way, a relationship can be established between the item category information function and the item information function, similar to the relationship between the IRT item information function and the test information function for PIRT. Although the amount of information shared by each category is different, its cumulative value is the item information curve (ICC) (Fig. 3). Similar to IRT, the sum of the ICC yields the test information curve (TIC). While the ICC is very important for item selection, TIC is a very powerful method for measurement precision (Hambleton et al., 1991). In this way, all the activities performed by test specialists to configure and adapt the test to an individual can be performed via CAT implementation during testing (Linden & Glas, 2010). The CAT can overcome the problems of the practicality of fixed-length tests. Some of the advantages of CAT over PPT are listed below (Hambleton et al., 1991; Rezaie & Golshan, 2015; Wainer et al., 2000; Weiss, 1982);

a. Faster response (Rezaie & Golshan, 2015).
b. Less test time (Hambleton et al., 1991; Rezaie & Golshan, 2015; Wainer et al., 2000; Weiss, 1982).
c. Determination of measurement precision for each person (Hambleton et al., 1991; Rezaie & Golshan, 2015; Wainer et al., 2000; Weiss, 1982)
d. Faster preparation of tests with predetermined difficulty and precision (Hambleton et al., 1991; Wainer et al., 2000)
e. Flexible test applications with asynchronous test administration (Wainer et al., 2000; Rezaie & Golshan, 2015)
f. Increased practicality for retesting (Rezaie & Golshan, 2015).
g. Feedback for individual test results (Hambleton et al., 1991; Rezaie & Golshan, 2015; Wainer et al., 2000; Weiss, 1982)
h. Rapid reporting (Rezaie & Golshan, 2015).
i. Increases the security of tests (Wainer et al., 2000)
j. Effective item pool management (Hambleton et al., 1991)
k. Flexibility in the item format (Hambleton et al., 1991; Rezaie & Golshan, 2015; Wainer et al., 2000)

Although studies focusing on CAT applications that measure cognitive traits are prevalent in the literature, there are few studies on psychological traits (interest, personality, attitude, etc.) (Betz & Turner, 2011; Hol et al., 2007; Reise & Henson, 2000; Vogels et al., 2011). Depression (Achtyes et al., 2015; Fliege et al., 2005; Gardner et al., 2004; Gibbons et al., 2012; Smits et al., 2011), anxiety (Gibbons et al., 2008, Gibbons et al., 2014), Personality (Reise & Henson, 2000; Simms & Clark, 2005; Waller &

Reise, 1989), personality disorder (Simms et al., 2011), vocational interest (Aybek & Çıkrıkçı, 2018; Betz & Turner, 2011), Motivation (Hol et al., 2007), psychological problems (Stochl et al., 2016), Psychosocial Problems (Vogels et al., 2011), Attitude (Baek, 1993) are some of the CAT applications developed based on PIRT models. Besides, it is possible to divide the studies on CAT applications into simulation and live (Weiss, 2004). Among the CAT studies on psychological traits, most of the literature is about simulation studies (Betz & Turner, 2011; Fliege et al., 2005; Gardner et al., 2004; Gibbons et al., 2008; Gibbons et al., 2012; Hol et al., 2007; Smits et al., 2011). On the other hand, live application studies are rare (Achtyes et al., 2015; Baek, 1993; Smits et al., 2011; Simms & Clark, 2005; Yasuda et al., 2022).

It has already been established that CAT applications have significant advantages over paper-pencil and computerized fixed-length tests. More studies are needed in the literature so that CAT applications can be widely used. The live CAT applications, which focus on measuring psychological traits, are an important step toward this goal. Investigating the equivalence of the CAT application with the PPT application is the main goal of the current research. Vocational interest inventories are widely used, and the tests are long (i.e., they contain many items). Given the potential of CAT to make long tests more feasible, an occupational interest inventory was preferred in this study. Since this is a methodological study, details about vocational interest inventories and their measurement are not mentioned. In this context, a live CAT application of a vocational interest inventory was developed and investigated to determine whether its practicality could be increased without compromising validity.

## Method

This research is applied research because it contains information produced to overcome the usefulness problem of a measurement tool. Applied research is the research conducted to evaluate the information generated for the actual solution of the problem (Karasar, 2009).

### Participants

Data were collected from 1449 high school students (45% female), using the paper-pencil version for IRT parameter estimates and CAT simulation studies. In the Turkish education system, there are different types of high schools depending on the curriculum. Therefore, students from different types of schools were selected (60% general academic, 13% science, 13% vocational, and 14% Imam-Hatip) because the measured characteristic is vocational interest. For the equivalence study, the research group consisted of 81 students (47% female) who participated in both the paper-pencil and live CAT applications.

### Instruments

In the research, the vocational interest inventory called SCI, the Turkish version adapted by Şimşek & Tavşancıl (2022), was used to develop the CAT application. The original SCI was developed by Betz et al. (2003) as an updated version of the Strong interest inventory. The SCI paper-pencil version consists of 17 factors and 164 items. Creative Production (CS – 10 items), Cultural Sensitivity (CS – 10 items), Data Management (DM – 10 items), Helping (HE – 6 items), Leadership (LE – 10 items), Mathematics (Ma – 10 items), Mechanical (Me – 10 items), Office Services (OS – 10 items), Organizational Management (OM – 9 items), Project Management (PM – 10 items), Public Speaking (PS – 9 items), Sales (Sa – 10 items), Science (Sc – 10 items), Teaching (Te – 10 items), Teamwork (TW – 10 items), Using Technology (UT – 10 items), and Writing (Wr – 10 items) are the vocational interests measured by the SCI.

## Design and Procedure

The SCI-CAT version was developed as an RShiny web application using the shiny (0.14.1) package to avoid software or hardware issues. The main reason for choosing the R language is that it contains design components such as HTML and Bootstrap and works in harmony with the necessary packages for the CAT application. The development took into account the international standards for computer-based and Internet-transmitted testing established by ITC (2005). The SCI-CAT application consists of three main screens: Info and Instructions, Test (Fig. 2) and Result (Fig. 3). In the design of CAT, Expected a Posteriori (EAP) was used as the estimation method, unweighted Fisher information (UW-FI) as the item selection rule, and SE<.500 as the test termination rule.

**Figure 2**

*SCI-CAT Test Screen*


**Figure 3**

*SCI-CAT Result Screen*


For the live CAT application, the study group consisting of 81 volunteers was divided into two groups. Group A first participated in the live CAT application and then answered the version PPT. In group B, the reverse process was carried out as in group A.

## Data Analysis

The research data were analyzed using the R packages psych (v1.5.8; Revelle, 2015), ltm (v1.0; Rizopoulos, 2006), and catIrt (v0.5.1; Nydick, 2022). The PIRT model used for the theta estimates was selected by examining the assumptions and checking the data-model fit. Then, the item parameters were calculated using the determined PIRT model. The estimation method, item selection, and test termination rule were determined for the design of CAT through a post-hoc simulation study. Theta estimates of occupational interest for the SCI factors of participants who received both the CAT and PPT versions were obtained using the EAP method. Spearman correlation, Wilcoxon signed-rank test, and descriptive statistics were used to examine the equivalence of the CAT and PPT estimates. A significance level of .05 was determined for the hypothesis tests.

## Results

### Data-Model Fit

The unidimensionality assumption was verified by calculating the ratio of adjacent eigenvalues for each SCI factor. The results of the parallel analysis showed that the ratio of the first eigenvalue (λ1) to the second eigenvalue (λ2) varied between 3.3 and 5.6. Hambleton et al. (1991) stated that the assumption of unidimensionality is satisfied when the ratio between the first eigenvalue and the second eigenvalue is large, and there is a dominant factor. The SCI factors whose adjacent eigenvalue ratios are greater than 3 indicate unidimensionality. When the assumption of unidimensionality is met, the assumption of local independence is also met because only one factor affects the person's responses to the items (Crocker & Algina, 1986; Hambleton et al., 1991; Embretson & Reise, 2000; Thissen & Wainer, 2001; Reise & Revicki, 2015). For model selection, the -2LL values for the GRM, GRM-C, GPCM, and GPCM -C models were determined using the ltm (1.0) package (Table 1). The results showed that the lowest -2LL values were obtained for the KTM model compared to the other models. A lower value of -2LL indicates a better data-model fit (Dodd et al., 1995; Kang et al., 2005; Reise, 1990).

**Table 1**

*GRM, GRM-C, GPCM, and GPCM-C -2LL Values*

| SCI factor | GRM | GRM-C | GPCM | GPCM -C |
|---|---|---|---|---|
| Creative Production (CS) | **40617.40** | 41272.20 | 40865.40 | 41625.60 |
| Cultural Sensitivity (CS) | **41709.00** | 42075.80 | 41907.00 | 42355.20 |
| Data Management (DM) | **39619.40** | 40017.60 | 39889.40 | 40266.60 |
| Helping (HE) | **24039.40** | 24581.60 | 24306.20 | 24920.80 |
| Leadership (LE) | **39188.40** | 39268.00 | 39524.60 | 39627.20 |
| Mathematics (Ma) | **41602.60** | 42052.20 | 41874.60 | 42295.20 |
| Mechanical (Me) | **39849.20** | 40240.80 | 40098.20 | 40468.00 |
| Office Services (OS) | **36202.40** | 36380.40 | 36401.00 | 36617.40 |
| Organizational Management (OM) | **41667.00** | 41908.60 | 41793.00 | 42029.80 |
| Project Management (PM) | **39237.00** | 39380.80 | 39542.20 | 39697.60 |
| Public Speaking (PS) | **36161.60** | 36266.20 | 36381.00 | 36503.00 |
| Sales (Sa) | **38743.00** | 39294.60 | 38958.80 | 39521.60 |
| Science (Sc) | **40598.80** | 40820.40 | 40820.40 | 41094.40 |
| Teaching (Te) | **39347.40** | 39608.00 | 39605.20 | 39968.40 |
| Teamwork (TW) | **38959.00** | 39069.00 | 39203.80 | 39313.60 |
| Using Technology (UT) | **37843.00** | 44153.80 | 38139.80 | 38955.00 |
| Writing (Wr) | **40346.20** | 40516.20 | 40586.20 | 40755.80 |

The significance of the chi-square values for the item-model fit was examined using PARSCALE software. The results showed that GRM item-model fit was met for all items except six items (M037, M078, M095, M135, M147). The item parameter was estimated using the GRM for each factor of SCI. Item slope parameters of the items for each factor were analyzed descriptively (Table 2). According to Baker (2001, p.21), the item slope parameter is interpreted as low below 0.64, medium for 0.65-1.34, and high above 1.35. Although relatively low for a few factors (CS, OM, OS), the slope parameters of the SCI items are generally high.

**Table 2**

*Descriptive Statistic of Item Slope Parameter (a)*

| | k | min | max | mean (median) | std. dev. |
|---|---|---|---|---|---|
| CP | 10 | 0.60 | 2.75 | 1.71 (1.78) | 0.67 |
| CS | 10 | 0.66 | 2.50 | 1.39 (1.43) | 0.51 |
| DM | 10 | 0.96 | 2.84 | 1.77 (1.75) | 0.57 |
| He | 6 | 0.70 | 3.73 | 2.05 (1.89) | 1.05 |
| Le | 10 | 1.27 | 2.05 | 1.71 (1.74) | 0.24 |
| Ma | 10 | 0.73 | 2.45 | 1.58 (1.58) | 0.57 |
| Me | 10 | 0.99 | 2.64 | 1.72 (1.78) | 0.62 |
| OS | 10 | 0.97 | 2.13 | 1.60 (1.48) | 0.39 |
| OM | 9 | 0.73 | 1.94 | 1.33 (1.26) | 0.40 |
| PM | 10 | 1.09 | 2.26 | 1.66 (1.66) | 0.32 |
| PS | 9 | 1.15 | 2.07 | 1.68 (1.72) | 0.28 |
| Sa | 10 | 0.62 | 2.57 | 1.68 (1.76) | 0.60 |
| Sc | 10 | 1.20 | 2.45 | 1.71 (1.61) | 0.41 |
| Te | 10 | 1.04 | 2.20 | 1.64 (1.69) | 0.40 |
| TW | 10 | 1.28 | 2.23 | 1.67 (1.55) | 0.31 |
| UT | 10 | 0.96 | 3.50 | 2.20 (2.35) | 0.80 |
| Wr | 10 | 1.22 | 2.50 | 1.76 (1.67) | 0.42 |

**Post-Hoc simulation**

The post-hoc, Monte Carlo, or hybrid simulation studies are methods used to determine the CAT design (IACAT, 2016). Basically, a CAT design consists of the components of test initiation, item selection, test termination, and theta estimation (Thompson & Weiss, 2011).

Item selection; When examining the commonly used item selection rules for PIRT, it is found that Fisher Information (FI) and Kullbak-Leibler (KL) derivations are most commonly used (Choi & Swartz, 2009; He et al., 2014; Lu et al., 2012; Veldkamp, 2001). The simulation study examined the performance of unweighted Fisher information (UW-FI), Kullback-Leibler information (FP-KL), and posterior weighted Fisher information (PW-FI) for item selection.

Test termination; The standard error rule (SE) is the most commonly used test termination rule (Babcock & Weiss, 2012). Considering the relationship between SE and measurement precision, .315, .385, and .500 SE are used, corresponding to measurement precision of .90, .85, and .75, respectively (Babcock & Weiss, 2012; Kezer, 2013; Sulak & Kelecioğlu, 2019).

Estimation method; MLE and EAP methods are the leading methods used in theta estimation. It is known that the EAP estimation method can make estimates from the first item and offers significant advantages in measurement precision for short tests (Weiss, 1982). It has been observed that EAP estimation is superior to MLE in CAT applications, specifically using the GRM model (Chen et al., 1997).

The CAT designs which are generated by the item selection (UW-FI, FP-KL, PW-FI), estimation method (MLE, EAP), and test termination (SE <.315, SE <.385, SE <.500) were examined by the simulation study (Table 3). Considering that there is no prior knowledge about the individuals, the item that provides the most information in the range of θ(-1,+1) was used as the starting rule for the test.

**Table 3**

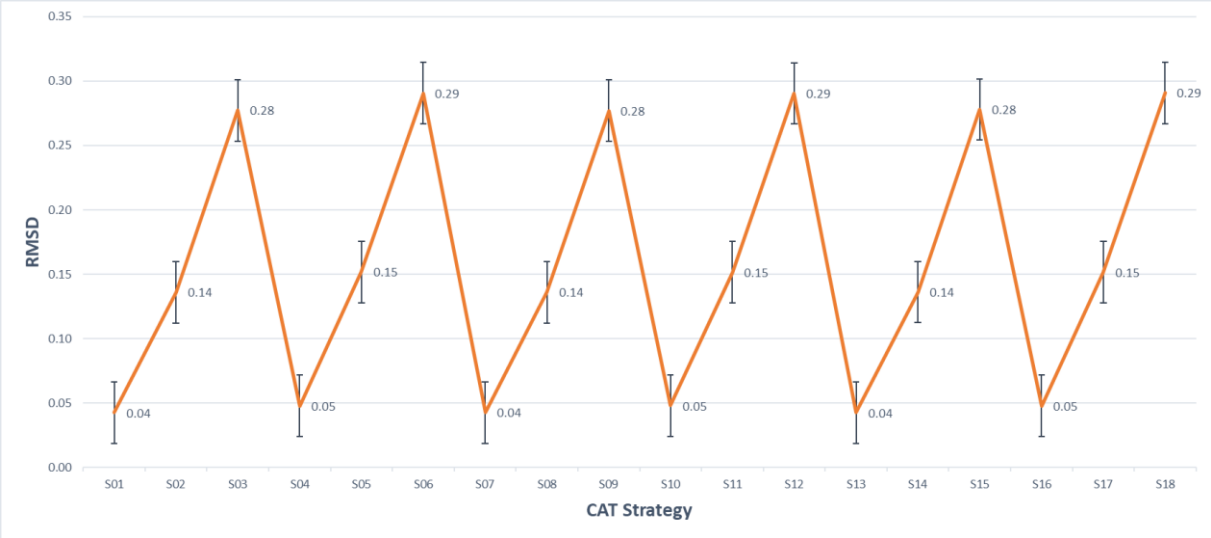
*The CAT Designs for Simulation Study*

| item selection | theta estimation | test termination | cat design |
|---|---|---|---|
| **UW-FI** | MLE | SE<.315 | S01 (UW-FI, MLE, SE<.315) |
| | | SE<.385 | S02 (UW-FI, MLE, SE<.385) |
| | | SE<.500 | S03 (UW-FI, MLE, SE<.500) |
| | EAP | SE<.315 | S04 (UW-FI, EAP, SE<.315) |
| | | SE<.385 | S05 (UW-FI, EAP, SE<.385) |
| | | SE<.500 | S06 (UW-FI, EAP, SE<.500) |
| **FP-KL** | MLE | SE<.315 | S07 (FP-KL, MLE, SE<.315) |
| | | SE<.385 | S08 (FP-KL, MLE, SE<.385) |
| | | SE<.500 | S09 (FP-KL, MLE, SE<.500) |
| | EAP | SE<.315 | S10 (FP-KL, EAP, SE<.315) |
| | | SE<.385 | S11 (FP-KL, EAP, SE<.385) |
| | | SE<.500 | S12 (FP-KL, EAP, SE<.500) |
| **PW-FI** | MLE | SE<.315 | S13 (PW-FI, MLE, SE<.315) |
| | | SE<.385 | S14 (PW-FI, MLE, SE<.385) |
| | | SE<.500 | S15 (PW-FI, MLE, SE<.500) |
| | EAP | SE<.315 | S16 (PW-FI, EAP, SE<.315) |
| | | SE<.385 | S17 (PW-FI, EAP, SE<.385) |
| | | SE<.500 | S18 (PW-FI, EAP, SE<.500) |

The performance of the CAT designs was evaluated by comparing the root mean square deviation (RMSD) and test length. Figure 4 shows that the RMSD value is sensitive to the SE value, which was set as the test termination rule. CAT Designs with less SE resulted in low RMSD. For this reason, savings in test length were reviewed for the CAT strategies (Table 4). Results show that when median scores are examined, CAT designs that use the test-stopping rule SE <.315, use almost the entire item set. This compromises the potential utility of CAT in terms of test length. When using the stopping rule SE < .500, which has sufficient measurement accuracy and the EAP estimation method, the test length with CAT has drastically decreased compared to the PPT version. The item selection method had no effect on the test length.

## Figure 4

*RMSD for the CAT Designs*




## Table 4

*Descriptive Statistics of Test Lengths for The CAT Designs*

| CAT design (item selection, theta estimation, test termination) | test length* | | |
|---|---|---|---|
| | min | max | median |
| S01 (UW-FI, MLE, SE<.315) | 5.5 | 10.0 | 9.1 |
| S02 (UW-FI, MLE, SE<.385) | 4.1 | 9.9 | 7.0 |
| S03 (UW-FI, MLE, SE<.500) | 3.6 | 6.0 | 4.4 |
| S04 (UW-FI, EAP, SE<.315) | 5.3 | 10.0 | 9.0 |
| S05 (UW-FI, EAP, SE<.385) | 4.0 | 9.4 | 6.3 |
| S06 (UW-FI, **EAP, SE<.500**) | 3.1 | 4.2 | 3.5 |
| S07 (FP-KL, MLE, SE<.315) | 5.6 | 10.0 | 9.1 |
| S08 (FP-KL, MLE, SE<.385) | 4.1 | 9.9 | 7.0 |
| S09 (FP-KL, MLE, SE<.500) | 3.6 | 6.0 | 4.4 |
| S10 (FP-KL, EAP, SE<.315) | 5.4 | 10.0 | 9.0 |
| S11 (FP-KL, EAP, SE<.385) | 4.0 | 9.4 | 6.3 |
| S12 (FP-KL, **EAP, SE<.500**) | 3.1 | 4.2 | 3.5 |
| S13 (PW-FI, MLE, SE<.315) | 5.5 | 10.0 | 9.1 |
| S14 (PW-FI, MLE, SE<.385) | 4.1 | 9.9 | 7.0 |
| S15 (PW-FI, MLE, SE<.500) | 3.6 | 6.0 | 4.4 |
| S16 (PW-FI, EAP, SE<.315) | 5.4 | 10.0 | 9.0 |
| S17 (PW-FI, EAP, SE<.385) | 4.0 | 9.4 | 6.3 |
| S18 (PW-FI, **EAP, SE<.500**) | 3.1 | 4.1 | 3.5 |

Note: The SE(θ) termination rule was employed after answering three items.

* Average of all the SCI-CAT factors

**Table 5**

*Descriptive Statistic of the Measurement Precision*

| SCI factors | T(θ) | | SE(θ) | | $1\text{-SE}(\theta)^2$ |
|---|---|---|---|---|---|
| | mean | std. dev. | mean | std. dev. | |
| CP | 9.32 | 2.85 | 0.33 | 0.05 | 0.89 |
| CS | 6.18 | 1.79 | 0.38 | 0.04 | 0.86 |
| DM | 9.45 | 3.17 | 0.32 | 0.06 | 0.90 |
| He | 8.20 | 2.90 | 0.35 | 0.07 | 0.88 |
| Le | 8.29 | 2.25 | 0.33 | 0.04 | 0.89 |
| Ma | 8.17 | 2.22 | 0.34 | 0.04 | 0.88 |
| Me | 8.86 | 3.34 | 0.34 | 0.06 | 0.88 |
| OS | 6.64 | 1.48 | 0.37 | 0.04 | 0.86 |
| OM | 5.36 | 1.21 | 0.40 | 0.03 | 0.84 |
| PM | 8.50 | 2.02 | 0.33 | 0.04 | 0.89 |
| PS | 7.79 | 2.02 | 0.34 | 0.03 | 0.88 |
| Sa | 8.33 | 3.02 | 0.35 | 0.06 | 0.88 |
| Sc | 9.08 | 2.23 | 0.32 | 0.03 | 0.90 |
| Te | 7.99 | 2.20 | 0.34 | 0.04 | 0.88 |
| TW | 8.19 | 2.08 | 0.34 | 0.04 | 0.88 |
| UT | 14.11 | 6.40 | 0.28 | 0.08 | 0.92 |
| Wr | 9.10 | 2.70 | 0.33 | 0.04 | 0.89 |

Higher test information means lower standard error and higher measurement precision during a CAT application (Embretson & Reise, 2000). Therefore, descriptive statistics of test information and standard error values were calculated to assess the measurement precision of estimates from SCI-CAT (Table 5). The results show that the level of test information for the 14 factors of SCI-CAT varies from 8 to 14. On the other hand, the level of test information for three factors (CS, OS, OM) is relatively low compared to the other factors. It has already been noted that the item slope parameters for these factors are lower than for the other factors (see Table 2). High test information values indicated high measurement precision for SCI-CAT factors. As a result, lower SE values than expected were obtained when SCI-CAT application. Hence, the result shows that the measurement precision ($1\text{-SE}^2$) is higher than expected (between .84 and .94).

### The equivalence of CAT and PPT estimates

The individuals' CAT and PPT estimates were analyzed using correlation and analysis of variance techniques. Table 6 presents that the Spearman correlation between both estimates for the 17 factors of SCI ranged from .70 to .91. The median value of the correlation coefficients drops to .85. The results show that the CAT and PPT estimates are significantly associated.

**Table 6**

*The Correlation Coefficient Between CAT and PPT Estimates*

| | CP | CS | DM | He | Le | Ma | Me | OS | OM | PM | PS | Sa | Sc | Te | TW | UT | Wr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r* | .71 | .86 | .86 | .91 | .87 | .70 | .80 | .82 | .83 | .91 | .84 | .84 | .87 | .85 | .91 | .78 | .72 |

* All correlation coefficients are significant p<.05

Because the normality assumption was not met, the Wilcoxon signed-rank test, one of the nonparametric analyses of variance techniques, was used ($p < .05$). Table 7 shows that the CAT and PPT estimations for the 15 factors of SCI were not significantly different. On the other hand, the difference between the estimates of CAT and PPT was significant for the two factors of SCI (OS and OM).

## Table 7

*PPT and CAT Estimates Wilcoxon Test Results*

| | N | mean rank* | sum of ranks | z | p |
|---|---|---|---|---|---|
| CP | 41 | 43.29 | 1775.00 | -0.539 | 0.590 |
| | 40 | 38.65 | 1546.00 | | |
| CS | 38 | 41.75 | 1586.50 | -0.032 | 0.975 |
| | 41 | 38.38 | 1573.50 | | |
| DM | 36 | 42.47 | 1529.00 | -0.619 | 0.536 |
| | 45 | 39.82 | 1792.00 | | |
| He | 43 | 34.81 | 1497.00 | -0.380 | 0.704 |
| | 32 | 42.28 | 1353.00 | | |
| Le | 45 | 40.60 | 1827.00 | -0.784 | 0.433 |
| | 36 | 41.50 | 1494.00 | | |
| Ma | 37 | 39.18 | 1449.50 | -0.638 | 0.524 |
| | 42 | 40.73 | 1710.50 | | |
| Me | 37 | 38.69 | 1431.50 | -1.078 | 0.281 |
| | 44 | 42.94 | 1889.50 | | |
| OS | 31 | 33.82 | 1048.50 | -2.741 | 0.006 |
| | 49 | 44.72 | 2191.50 | | |
| OM | 52 | 41.38 | 2152.00 | -2.552 | 0.011 |
| | 28 | 38.86 | 1088.00 | | |
| PM | 40 | 40.63 | 1625.00 | -0.220 | 0.826 |
| | 39 | 39.36 | 1535.00 | | |
| PS | 47 | 40.21 | 1890.00 | -1.295 | 0.195 |
| | 33 | 40.91 | 1350.00 | | |
| Sa | 44 | 40.83 | 1796.50 | -0.640 | 0.522 |
| | 37 | 41.20 | 1524.50 | | |
| Sc | 33 | 41.33 | 1364.00 | -1.396 | 0.163 |
| | 48 | 40.77 | 1957.00 | | |
| Te | 40 | 38.49 | 1539.50 | -0.570 | 0.569 |
| | 41 | 43.45 | 1781.50 | | |
| TW | 41 | 38.40 | 1574.50 | -0.169 | 0.866 |
| | 37 | 40.72 | 1506.50 | | |
| UT | 33 | 41.48 | 1369.00 | -1.204 | 0.228 |
| | 47 | 39.81 | 1871.00 | | |
| Wr | 36 | 40.79 | 1468.50 | -0.168 | 0.867 |
| | 41 | 37.43 | 1534.50 | | |

* : first row: *CAT<PPT*; second row: *PPT<CAT*

*Note: Z-scores were obtained for each individual's PTT and CAT estimates. Z-scores were used for the Wilcoxon test.*

## Table 8

*The mean difference between CAT and PPT estimates*

| | mean | std. dev. |
|---|---|---|
| **CP** | 0.00 | 0.69 |
| **CS** | 0.01 | 0.57 |
| **DM** | -0.03 | 0.52 |
| **He** | 0.02 | 0.38 |
| **Le** | 0.03 | 0.59 |
| **Ma** | -0.02 | 0.71 |
| **Me** | -0.03 | 0.58 |
| **OS** | -0.10 | 0.49 |
| **OM** | 0.07 | 0.41 |
| **PM** | 0.01 | 0.45 |
| **PS** | 0.02 | 0.63 |
| **Sa** | -0.01 | 0.51 |
| **Sc** | -0.05 | 0.53 |
| **Te** | -0.02 | 0.60 |
| **TW** | 0.01 | 0.46 |
| **UT** | -0.01 | 0.70 |
| **Wr** | 0.02 | 0.77 |

* The mean difference between of CAT and PPT

The average values of theta difference for both measurements of the individuals are shown in Table 8. The highest difference between the theta values of 0.10 belongs to the factor OS. Considering the theta range (±4), we can say that this difference is small enough to be neglected. This indicates that the estimates of SCI-CAT are consistent with the results of PPT. Considering the test information values given in Table 5, it was evaluated that the low measurement precision of the factors OS and OM is the cause of the difference between the estimates of CAT and PPT of the individuals.

In the PPT application, participants answered 164 items in approximately 30 minutes. In the application CAT, both the number of items answered and the response time of each participant were logged. Descriptive statistics of the number of items answered in the CAT application and the test duration can be found in Table 9. The number of items answered varies between 69 and 121, with an average of 83 (SD =12). Participants' response time is distributed with an average of 7 minutes (SD =2). The results show that SCI-CAT can save 50% of the test length and 77% of the test duration compared to the PPT version.

## Table 9

*Descriptive Statistics of Test Length and Duration of the SCI-CAT*

| | mean | std. dev. | min | max | range |
|---|---|---|---|---|---|
| Test length (number of items) | 83.2 | 11.7 | 69.0 | 121.0 | 52.0 |
| Test duration (minutes) | 6.9 | 1.9 | 4.1 | 13.2 | 9.0 |

## Discussion

The purpose of this study was to increase the practicality of a vocational interest inventory called SCI using CAT. The scale was evaluated by parallel analysis, and each factor was found to be unidimensional. Therefore, unidimensional polytomous IRT models were preferred for the parameter estimates. The fit of the model data was investigated using IRT models (GRM, GRM-C, GPCM, GPCM-

C) developed for polytomous items. A better fit of the model data was obtained with the GRM model. Previous studies support the conclusion that the GRM makes better predictions for Likert items than the GPCM (Hol et al., 2007; Smits et al., 2011). The result shows that the factors consisting of items with high discrimination have higher test information (see Table 2 and Table 5). As a result, higher measurement accuracy is obtained for these factors. This result is confirmed by previous research (Langenbucher et al., 2004; Pedraza et al., 2011).

In this study, we specifically chose to evaluate the CAT design under different theta estimation methods, item selection rules, and test termination strategies. Previous studies have shown that polytomous IRT-based CAT can handle a small item set (Dodd et al., 1995; Paap et al., 2017). In addition, some research has found that CAT can be an accurate measure even when the instrument contains only five items per dimension (Paap et al., 2019).

The simulation study showed that the EAP estimation method and the SE < .500 test termination strategy were superior compared to the other CAT designs. Item selection did not play a role in reducing test length or increasing measurement accuracy. As a result, it was found that an examinee's interests could be estimated with approximately four items. The finding that the EAP estimation method is more useful with small item pools is consistent with similar studies in the literature (Chen et al., 1997; Eroğlu & Kelecioğlu, 2015; Weiss, 1982). Similar to the literature, this study also found that the EAP estimation method was more useful than the MLE estimation method in terms of test length and theta estimation. The results show that SE<.500 is more efficient as a termination strategy in terms of test length for a CAT application. (Achtyes et al., 2015; Betz & Turner, 2011; Demir & French, 2021; Hol et al., 2007; Simms et al., 2011; Simms & Clark, 2005; Stochl et al., 2016). The results obtained in this study are consistent with those in the literature (Babcock & Weiss, 2012; Choi & Swartz, 2009; Deng et al., 2010; Eroğlu & Kelecioğlu, 2015; Gnambs & Batinic, 2011; He et al., 2014; Kezer, 2013; Linden, 2005; Ping et al., 2006; Sulak & Kelecioglu, 2019; Weiss, 1982).

Results from the live CAT application showed that estimates of CAT were strongly positively correlated with paper-pencil. With the exception of two factors, the difference between individuals' estimates obtained from both applications is not statistically significant. Consequently, the estimates from CAT are equivalent to the results from paper-pencil. This is consistent with recent studies on the equivalence of CAT (Abidin et al., 2019; Demir & French, 2021, Yasuda et al., 2022). In addition, the implementation of CAT increased the practicality compared to the fixed-length test version by reducing test length and time. Similar studies support the findings regarding the advantage of CAT in terms of test length and duration (Abidin et al., 2019; Alkhadher et al., 1998; Betz & Turner, 2011; Choi et al., 2010; Demir & French, 2021; Jodoin et al., 2006; Kezer, 2013; Paap et al., 2019; Rezaie & Golshan, 2015; Yasuda et al., 2022; Weiss, 2011).

The paper-pencil or computerized fixed-length tests are still the most popular method for psychometric measurement. It is not surprising that they are the first choice for short tests because of their ease of development and use. Based on our findings, CATs should be the first choice for long tests when it comes to measurement validity, despite the relatively difficult development process. We recommend that developers of CAT use an item pool consisting of items with high item discrimination to achieve high measurement accuracy. The results of this study can also serve as a reference for educational supervisors to use the online CAT system in large-scale examinations such as the National Career Program. It is recommended that researchers conduct more research on this topic so that CATs based on Polytomous IRT can be widely used.

## Declaration

**Author Contribution:** Author 1 - Theoretical framework, literature review, methodology, data collection, data analysis, discussion, and writing the original draft. Author 2 - Theoretical framework, methodology, discussion, supervision, and editing of the original draft.

**Conflict of Interest:** The authors did not declare a potential conflict of interest.

**Ethical Approval:** The study was ethically approved by the Ministry of National Education (research number: 81576613/605/2144292, dated 26/02/2015). This study has been produced from the dissertation of the first author that was conducted under the supervision of the second author.

## References


Abidin, A. Z., Istiyono, E., Fadilah, N., & Dwandaru, W. S. B. (2019). A computerized adaptive test for measuring the physics critical thinking skills. *International Journal of Evaluation and Research in Education*, 8(3), 376-383. http://dx.doi.org/10.11591/ijere.v8i3.19642

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: he VOCATIONS trial. *Psychiatric Services*, 1–6. http://doi.org/10.1176/appi.ps.201400390

Alkhadher, O., Clarke, D. D., & Anderson, N. (1998). Equivalence and predictive validity of paper-and-pencil and computerized adaptive formats of the differential aptitude tests. *Journal of Occupational and Organizational Psychology*, 71(3), 205–217. http://doi.org/10.1111/j.2044-8325.1998.tb00673.x

Aybek, E. C., & Çıkrıkçı, R. N. (2018). Kendini değerlendirme envanteri'nin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliği. *Turkish Psychological Counseling and Guidance Journal*, 8(50), 117-141. http://hdl.handle.net/20.500.12575/37233

Babcock, B., & Weiss, D. J. (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. http://doi.org/10.7333/1212-0101001

Baek, S. G. (1995). Computerized adaptive attitude testing using the partial credit model. *Dissertation Abstracts International*, 55(7-A), 1922. Retrieved April 10, 2022, from PsychInfo database.

Baker, F. B. (2001). *The basics of item response theory* (second edition). Retrieved July 22, 2022, from http://eric.ed.gov/?id=ED458219

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. http://doi.org/10.1177/1069072710395534

Betz, N. E., Borgen, F. H., Rottinghaus, P., Paulsen, A., Halper, C. R., & Harmon, L. W. (2003). The expanded skills confidence inventory: measuring basic dimensions of vocational activity. *Journal of Vocational Behavior*, 62(1), 76–100. http://doi.org/10.1016/S0001-8791(02)00034-9

Chen, S.-K., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population and method of theta estimation on computerized adaptive testing (CAT) using the rating scale model. *Educational and Psychological Measurement*, 57(3), 422–439. https://doi.org/10.1177/0013164497057003004

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement*, 33(6), 419–440. http://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. http://doi.org/10.1007/s11136-009-9560-5

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich

Demir, C., & French, B. F. (2021). Applicability and efficiency of a computerized adaptive test for the Washington assessment of the risks and needs of students. *Assessment*. https://doi.org/10.1177/10731911211047892

Deng, H., Ansley, T., & Chang, H. H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement*, 47(2), 202–226. http://doi.org/10.1111/j.1745-3984.2010.00109.x

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

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

Eroğlu, M. G., & Kelecioğlu, H. (2015). Bireyselleştirilmiş bilgisayarlı test uygulamalarında farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, 28(1), 31–52. https://doi.org/10.19171/uuefd.87973

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 14(10), 2277–91. http://doi.org/10.1007/s11136-005-6651-9

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(1), 13. http://doi.org/10.1186/1471-244X-4-13

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., … Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, 59(4), 361–8. http://doi.org/10.1176/appi.ps.59.4.361

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. http://doi.org/10.1001/archgenpsychiatry.2012.14

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014). Development of the CAT-ANX: A computerized adaptive test for anxiety. *American Journal of Psychiatry*, 171(2), 187–194. http://doi.org/10.1176/appi.ajp.2013.13020178

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. http://doi.org/10.1177/0013164410393956

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

He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement*, 74(4), 677–696. http://doi.org/10.1177/0013164413517503

Hol, M. A., Vorst, H. C., & 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. http://doi.org/10.1177/0146621606297314

IACAT. (2016). *Research Strategies in CAT | IACAT*. Retrieved February 2, 2019, from http://iacat.org/content/research-strategies-cat

International Test Commission. (2005). *ITC Guidelines for Translating and Adapting Tests*. Retrieved February 2, 2019, from www.intestcom.org

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. http://doi.org/10.1207/s15324818ame1903_3

Kang, T., Cohen, A. S., & Sung, H.-J. (2005). IRT model selection methods for polytomous items. *In: Annual Meeting of the National Council on Measurement in Education*, Montreal, 2005. Retrieved February 2, 2019, from https://testing.wisc.edu/

Kang, T., Cohen, A. S., & Sung, H.-J. (2009). Msodel selection indices for polytomous items. *Applied Psychological Measurement*, 33(7), 499–518. http://doi.org/10.1007/s00330-011-2364-3

Karasar, N. (2009). *Bilimsel araştırma yöntemleri*. Ankara: Nobel Yayın Dağıtım.

Kezer, F. (2013). Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması. *Eğitim Bilimleri Araştırmaları Dergisi*, 4(1), 145–175. http://doi.org/http://dx.doi.org/10.12973/jesr.2014.41.8

Langenbucher, J. W., Labouvie, E., Martin, C. S., Sanjuan, P. M., Bavly, L., Kirisci, L., & Chung, T. (2004). An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. *Journal of abnormal psychology*, *113*(1), 72. https://doi.org/10.1037/0021-843x.113.1.72

Linden, W. J. Van Der, & Glas, C. A. W. (2010). Elements of Adaptive Testing. New York, NY: Springer.

Linden, W. J. Van Der. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, 42(3), 283-302. http://dx.doi.org/10.1111/j.1745-3984.2005.00015.x

Lu, P., Zhou, D., Qin, S., Cong, X., & Zhong, S. (2012). The study of item selection method in CAT. *In: 6th International Symposium*, ISICA (pp. 403–415). Wuhan - China.

Nydick, S. (2022). catIrt: Simulate IRT-Based Computerized Adaptive Tests. R package version 0.5.1. https://CRAN.R-project.org/package=catIrt

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. SAGE.

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

Paap, M. C. S., Kroeze, K. A., Glas, C. A. W., Terwee, C. B., van der Palen, J., & Veldkamp, B. P. (2017). Measuring patient-reported outcomes adaptively: multidimensionality matters!. *Applied Psychological Measurement*, 42(5), 327–342. https://doi.org/10.1177/0146621617733954

Pedraza, O., Sachs, B. C., Ferman, T. J., Rush, B. K., & Lucas, J. A. (2011). Difficulty and discrimination parameters of Boston Naming Test items in a consecutive clinical series. *Archives of Clinical Neuropsychology*, *26*(5), 434-444. https://doi.org/10.1093/arclin/acr042

Ping, C., Shuliang, D., Haijing, L., & Jie, Z. (2006). Item selection strategies of computerized adaptive testing based on graded response model. *Acta Psychologica Sinica*, 38(03), 461. https://journal.psych.ac.cn/acps/EN/Y2006/V38/I03/461

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory (pp. 79-112).* Springer.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, 14(2), 127-137. https://doi.org/10.1177/014662169001400202

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment*, 7(4), 347–364. https://doi.org/10.1177/107319110000700404

Reise, S. P., & Revicki, D. A. (2015). *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge.

Ren, H., Choi, S.W. & van der Linden, W.J. (2020). Bayesian adaptive testing with polytomous items. *Behaviormetrika* 47, 427–449. https://doi.org/10.1007/s41237-020-00114-8

Revelle, W. (2015) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.5.8.

Rezaie, M., & Golshan, M. (2015). Computer adaptive test (CAT): Advantages and limitations. *International Journal of Educational Investigations*, 2(5), 128–137. http://www.ijeionline.com/attachments/article/42/IJEI_Vol.2_No.5_2015-5-11.pdf

Rizopoulos, D. (2006). "ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses." *Journal of Statistical Software*, 17(5), 1–25. https://doi.org/10.18637/jss.v017.i05.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 35(17), 139. http://doi.org/10.1007/BF02290599

Schinka, J. A., & Velicer, W. F. (2003). Research Methods in Psychology. In: I. B. Weiner (Ed.), *Handbook of Psychology* (Vol. 2). John Wiley & Sons, Inc.

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. http://doi.org/10.1037/1040-3590.17.1.28

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: introducing the CAT–PD project. *Journal of Personality Assessment*, 93(4), 380–389. http://doi.org/10.1080/00223891.2011.577475

Şimşek, A.S., & Tavşancıl, E. (2022). Validity and reliability of Turkish version of skills confidence inventory. *Turkish Psychological Counseling and Guidance Journal*, 12(64), 89-107. https://doi.org/10.17066/tpdrd.1096008

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. http://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. http://doi.org/10.1186/s12874-016-0158-7

Sulak, S., & Kelecioğlu, H. (2019). Investigation of Item Selection Methods According to Test Termination Rules in CAT Applications. *Journal of Measurement and Evaluation in Education and Psychology*, 315–326. https://doi.org/10.21031/epod.530528

Thissen, D., & Wainer, H. (2001). *Test Scoring*. Lawrance Erlbaum Associates.

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

Veldkamp, B. P. (2001). Item selection in polytomous CAT. In *Proceedings of the International Meeting of the Psychometric Society* IMPS2001 (pp. 207–214). Osaka - Japan.

Vogels, A. G. C., Jacobusse, G. W., & Reijneveld, S. A. (2011). An accurate and efficient identification of children with psychosocial problems by means of computerized adaptive testing. *BMC Medical Research Methodology*, 11, 111. http://doi.org/10.1186/1471-2288-11-111

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R., Thissen, D. (2000). *Computerized adaptive testing: A primer* (Second Ed). Lawrence Erlbaum Assocaiates.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: an illustration with the Absorption scale. *Journal of Personality and Social Psychology*, 57(6), 1051–1058. http://doi.org/10.1037/0022-3514.57.6.1051

Wang, S., & Wang, T. (2002). *Relative precision of ability estimation in polytomous CAT: a comparison under the generalized partial credit model and graded response model*. American Educational Research Association.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6(4), 473–492. https://doi.org/10.1177/014662168200600408

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, 37(2), 70–84. Retrieved from http://www.psych.umn.edu/psylabs/catcentral/pdf files/we04070.pdf

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. Retrieved from https://www.assess.com/docs/Weiss(2011)_CAT.pdf

Yasuda, J. I., Hull, M. M., & Mae, N. (2022). Improving test security and efficiency of computerized adaptive testing for the Force Concept Inventory. *Physical Review Physics Education Research*, *18*(1), 010112. https://doi.org/10.1103/PhysRevPhysEducRes.18.010112