Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575

Kış 2022
Winter 2022

Cilt: 13-Sayı: 4
Volume: 13-Issue: 4

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayımlanan hakemli uluslararası bir dergidir. Yayımlanan yazıların tüm sorumluluğu ilgili yazarlara aittir.

Journal of Measurement and Evaluation in Education and Psychology (JMEEP) is a international refereed journal that is published four times a year. The responsibility lies with the authors of papers.

İletişim
e-posta: epodderdergi@gmail.com
Web: https://dergipark.org.tr/tr/pub/epod

Contact
e-mail: epodderdergi@gmail.com
Web: http://dergipark.org.tr/tr/pub/epod

Dizinleme / Abstracting & Indexing
Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), SCOPUS, TÜBİTAK TR DIZIN Sosyal ve Beşeri Bilimler Veri Tabanı (ULAKBİM), Tei (Türk Eğitim İndeksi), EBSCO

Hakem Kurulu / Referee Board

Hakem Kurulu / Referee Board

*Ada göre alfabetik sıralanmıştır. / Names listed in alphabetical order.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Aralık 2022, Sayı: 13-4)

Journal of Measurement and Evaluation in Education and Psychology (December 2022, Issue: 13-4)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(4); 282-304

Analyzing the Interaction of Item Position Effect and Student Characteristics within Explanatory IRT Models*

Sinem DEMİRKOL** Hülya KELECİOĞLU***

Abstract

The location of the same item in different positions among booklets leads to a biased estimation of item parameters. This undesirable effect on the probability of answering the items correctly is referred as the item position effect. The purpose of this study is to examine the items that are more sensitive to the item position effect and to investigate the student characteristics related to the item position effect. In the study, the items in the PISA 2015 reading domain are used. The study group consists of 2418 students who responded to the items in the reading domain from PISA 2015 Turkey Sample. Explanatory IRT models are used in the analysis of the research. According to the results, 42% of the items are affected by the item position. The most important characteristic related to item position is the SES level of students. In addition, male students are more affected by item position than female students.

Keywords: Item position effect, Explanatory IRT models, student characteristics reading domain

Introduction

In general, tests are tools that measure and understand the natural or learned abilities, knowledge, and characteristics of a person or community (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014). The general purpose of these tools is to reveal the properties of individuals to be measured through observable response behaviors to a series of items (Rose et al., 2019). Items in a booklet are always given in a context. As the test and item properties change, the context of the item in the test changes. These adjustments are possible sources of construct-irrelevant variability in test scores. Such undesirable variability in response behaviors, which cannot be explained by the ability level of students and stem from the context of the test, is expressed as context effect (Leary & Dorans, 1985; Yen, 1980). According to Brennan (1992), although there is no full definition of context effects, differences in one or more statistics should be taken as evidence for context effects.

As a type of context effects, the effect of the same item in different locations among the booklets on the probability of answering the item correctly is expressed as the item position effect (Leary & Dorans, 1985). According to Weirich et al. (2017), item position effects can be considered a special case of context effects, since the position in a test is part of the context. Therefore, although context and position effects are tried to separate from each other, it is not correct to assert the cause of the construct-irrelevant effect as the context or the position effect only (Brennan, 1992; Leary & Dorans, 1985).

The fact that the same item is in different positions among test forms affects the item parameters (Bulut, Quo & Gierl, 2017; Mollenkopf, 1950; Qian, 2014). Studies generally examined the effect of item position on item difficulty (Alexandrowicz & Matschinger, 2008; Guertin, 1954; Wise et al., 1989; Harting & Buchholz, 2012, Hahne, 2008; MacNicol, 1956). Items at the end of the test may become easier or more difficult than items at the beginning of the test. There are usually two possible

[*] The present study is a part of PhD Thesis conducted under the supervision of Prof. Dr. Hülya KELECİOĞLU and prepared by Sinem DEMİRKOL.

[**] PhD., Hacettepe University, Faculty of Education, Ankara-Türkiye, dmrklsinem@gmail.com, ORCID ID: 0000-0002-9526-6156

[***] Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

To cite this article:
Demirkol, S., & Kelecioğlu, H. (2022). Analyzing the interaction of item position effect and student characteristics within explanatory IRT models. *Journal of Measurement and Evaluation in Education and Psychology, 13*(4), 282-304. https://doi.org/10.21031/epod.1126368

Received: 5.06.2022
Accepted: 3.11.2022

explanations, depending on the direction of the effect on item difficulty. The increase in item difficulty can be interpreted as an effect of students' fatigue or low motivation to answer. On the other hand, the decrease in item difficulty can be interpreted as a practice or learning effect (Kingston & Dorans, 1984). The learning effect may be due to test takers becoming more familiar with the items during the test. (Hohensinn et al., 2011). Although both explanations seem plausible, in order to examine the position effects, item or individual properties should be included in the models (Debeer & Janssen, 2013). Investigating such variables will help to reduce the item position effects; thus, the reliability and validity of the measurement tool will improve.

The assumption that the item and person parameters are not affected by the booklet selection is violated by the item position effect (Albano, 2013). If item responses are not independent under one dimension, another dimension may cause dependency. The effect of the item position on the item parameters violates the local independence assumption of the item response theory (IRT) in particular (Hahne, 2008). This situation causes biased results in the estimation of the item parameters, and thus in the ability estimations of the individuals (Whiteley & Davis, 1976). One of the most important advantages of the models of IRT is that the item parameters are independent of the latent trait. (DeMars, 2016; Embretson & Reise, 2000). In cases where IRT assumptions are met, item parameter invariance is also ensured (Hambleton & Swaminathan, 1985). As mentioned above, item position effects violate the local independence assumption and item parameter invariance assumption, which is shown as one of the most important differences between classical test theory (CCT) and IRT. Violation of this assumption causes problems, especially in equating studies using common items (Angoff, 1971; Meyers, Miller & Way, 2009).

The item position effect differed among students, that is, the students were not exposed to the same level of item position effect (Christiansen & Janssen, 2020; Deeber & Janssen, 2013; Demirkol & Kelecioğlu, 2022). These results raised the question of what individual characteristics might be related to the item position effect, and the relationship between different student characteristics and the item position effect was investigated (Smouse & Munz, 1968; Nagy et al., 2018; Qian, 2014; Weirich et al., 2017; Wu et al., 2019). Therefore, in this study, the relationship between motivation, anxiety levels, gender, and SES, which are thought to have the most effect on students' response behaviors, and the item position effect are discussed.

When the literature is examined, it has been emphasized that the item position effect may be caused by the differences in the motivation level of the test takers (Kingston & Dorans, 1982; Albano, 2013; Debeer & Janssen, 2013). Wu et al. (2019) examined the relationship between students' motivation levels and item position effect to explain individual differences in item position effects. In the study, students' motivation levels were represented by the variables; enjoyment and interests, effort thermometer, and perseverance. According to the results of the research, in most countries, students who enjoyed reading had higher persistence levels in 2009 PISA (The Programme for International Student Assessment), but this relationship was not observed in 2006 PISA science and 2012 PISA mathematics. In addition, in some countries, it was stated that students' test-solving efforts had an effect on persistence in the 2006 and 2012 PISA. However, it was emphasized that a general motivational effect was not consistently associated with item position in all countries and PISA cycles. Weirich et al. (2017) stated that even when the initial motivation levels and variability of the motivation levels of the students are controlled, the position effects continue in an “ideal” group of highly motivated students during the test.

One of the most important variables associated with students' academic achievement is considered to be SES (Taylor, 2005). Students with higher SESs may have more chance to focus on their studies and maintain their attention (Sirin, 2005). This can make students more advantageous during the exam. Therefore, SES affects both academic skills and motivation of students (Duncan & Magnuson, 2005). Nagy et al. (2018), using PISA 2006 data, examined the relationship of the item position effect with student characteristics with structural equation models at school and student levels. In the reading field, it was found that the most important student characteristic related to item position was SES, and students with high SES levels were less affected by the item position effect by 32%. Contrary to this study, Wu et al., (2019) stated that in most of the countries they examined, there was no significant relationship between the item position effect and SES.

Grandy (1987) stated that the gender factor is effective in students' response behaviors. When the literature is examined, female students have more effort and their attention levels are higher than male students in low-stake exams (Butler & Adams, 2007; Eklöf, 2007). It can be thought that the motivation, attention, and effort of male students' change/decrease more than female students during the test, and therefore they are exposed to more item position effect. When the relationship between item position effect and gender was examined, Qian (2014) found that male students were more affected by item position in the 2007 NAEP (National Assessment of Educational Progress) writing data. Nagy et al. (2018) stated that gender was related to the item position effect in the domains of mathematics, reading, and science, and that male students were more affected by the item position effect than female students. Again, Wu et al. (2019) stated that male students were more affected by the item position effect than female students.

The effects of item position effect and test anxiety on achievement test scores were first studied by Smouse and Munz (1968). A multiple-choice test consisting of 100 items prepared in the field of psychology was administered to 113 undergraduate students. Three different forms of the test were prepared as items from easy to difficult, difficult to easy, and at random, and students were randomly divided into two groups. The first group was given information to increase the anxiety level of the students, while the normal test atmosphere was maintained in the second group. As a result of the analysis of variance, the ordering of the items according to different difficulty levels did not have any interaction with anxiety. Later, the research was expanded and the study was repeated with 40 students with the lowest and highest anxiety levels using the Achievement Anxiety Test. According to the results of this study, a significant interaction was found between item position and anxiety. This result was interpreted as students with very low or high anxiety levels might be affected by the item position. On the other hand, Berger et al. (1969) and Towle & Merrill (1975) found that there was no significant relationship between the item position effect and anxiety.

Purpose of the Study

In many test programs, it is assumed that context or position effects have no or negligible effect on students' responses. Violation of this assumption may lead to biased estimates in item and person parameters. Rather than a general item position effect, it will be valuable for test developers and practitioners to investigate the problematic items that are significantly affected by the item position and to find solutions for the undesirable effects that occur in these items (Albano, 2013). Depending on the psychological events that may occur during the test (such as frustration, excitement, fatigue) or individual characteristics (such as gender, SED), the position of an item contributes to the probability of correct answers. This result gave rise to the question, "Which characteristics of individuals related to the item position effect?" (Nagy et al. 2018; Bulut et al., 2017) The purpose of this study is to examine items that are more sensitive to item position rather than a general item position effect and individual characteristics that are thought to be related to item position effect. For this purpose, the relationship between the motivation, anxiety, SES, and gender variables of students and the effect of item position are examined on an item basis. In this study, it is aimed to answer the following questions.

1. How is the item position interaction at item level?
2. How does the item position effect interact with the SES?
3. How does the item position effect interact with gender?
4. How does the item position effect interact with students' test anxiety?
5. How does the item position effect interact with students' achievement motivations?

Method

Working Group

5895 (50.2% female, 49.5% male) students participated in the PISA 2015 in Turkey. Since the research focused on the items in the reading domain, all students who answered the items in the reading were included in the research. Therefore, the study group of the research consists of 2418 (49.8% female, 50.2% male) students who were drawn from the PISA 2015 Turkey sample and answered the items in the reading domain.

Data Collection Methods

PISA, financed by the OECD (Organization for Economic Cooperation and Development), is an international research to measure and evaluate the purposes of education (OECD, 2017). PISA mainly assesses students' literacy in science, math, and reading fields. Knowledge about students' motivations, their opinions about themselves, their psychological characteristics about learning processes, and school environments are collected via student, teacher, and school questionnaires.

PISA 2015 was carried out on computer based in Turkey. For computer-based assessment, 66 different main booklets were used. There were 12 different clusters in science, 6 in mathematics, 6 in reading, and 4 in collaborative problem-solving. The booklets used in PISA 2015 were created to include four of these clusters prepared in the fields of science, mathematics, reading, and collaborative problem-solving. The positions of the clusters in the booklets differ, but the positions of the items in the clusters are fixed. Therefore, the item position effect was examined on cluster-based in this study. Since the test forms consisted of four different clusters, the first cluster was coded as 0, the second cluster 1, the third cluster 2, and the fourth cluster 3. In this way, the position variable is a variable that takes a value between 0 and 3.

In this study, the effect of item position on reading items was investigated. Thirty-six booklets containing reading items were used. There were 88 items in the reading area, and all items were included in the analysis. Since the analysis model used was in accordance with the dichotomous scored item format, the partially scored (7 items) items were converted into dichotomous scoring with 0 for "incorrect" and "partially correct" items and 1 for "fully correct" items. In addition, omitted items were considered incorrect and inaccessible items or missing items due to other reasons were considered missing.

Variables Used in the Research

Gender. The gender variable is a two-category variable, coded as 0 for females and 1 for males.

Socio-economic status (SES). SES index is built by the PISA study team via principal components analysis using parent education (PARED), highest parent occupation (HISEI), and home possessions (HOMEPOS). This variable is standardized to have a mean of 0 and a standard deviation of 1. In the sample of Turkey, the minimum value of the SES variable is -5.131, and the maximum value is 3.123. In the PISA technical report, factor loadings of the variables used in the SES are given for each country. For the Turkish sample, the factor load of HISEI is calculated as 0.82, the factor load of PARED as 0.79, and the factor load of HOMEPOS as 0.77. In addition, the scale reliability (Cronbach's alpha) of the SES variable for the Turkish sample is estimated as 0.68.

Test anxiety. In order to investigate the test anxiety of test takers, there are five items in the PISA 2015 student questionnaire. Students answer these items in the categories of "strongly agree", "agree", "disagree", and "strongly disagree". The scores obtained from these categories are included in the student questionnaire with the ANXTEST code. This variable is standardized to have a mean of 0 and a standard deviation of 1. In the sample of Turkey, the minimum value of the anxiety variable is -2.505, and the maximum value is 2.549. For the Turkish sample, the reliability coefficient (Cronbach Alpha) of the "test anxiety" scale is estimated as 0.825.

Achievement motivation. There are five items in the PISA 2015 student questionnaire in order to investigate the achievement motivation of test takers. Students answered these items in the categories of "strongly agree", "agree", "disagree", and "strongly disagree". The scores obtained from these categories are included in the student questionnaire with the MOTIVAT code. This variable is standardized to have a mean of 0 and a standard deviation of 1. The minimum value of this variable is -3.087, and the maximum value is 1.854. For the Turkish sample, the reliability coefficient (Cronbach Alpha) of the scale is estimated as 0.825. For detailed information, the PISA 2015 technical report can be viewed (OECD, 2017).

Analysis of Data

A two-stage procedure can be used when the purpose of research is to explain the differences in person and item parameters with traditional IRT models. In the first stage, the abilities and item parameters are estimated, and in the second stage, the variables that are thought to cause differences in ability and item parameters can be modeled with various methods (De Boeck & Wilson, 2004; Atar & Çobanoğlu Aktan, 2013).

When IRT models are considered within the framework of generalized linear or non-linear mixed models, they can simultaneously estimate the variation among individual and item parameters by including the predictors of the individual, item, or interaction of both in the model. These models are called Explanatory IRT models (De Boeck & Wilson, 2004). The main advantage of explanatory IRT models is that they provide flexibility to analyse the covariance between these parameters simultaneously while estimating item and individual parameters (Briggs, 2008).

In explanatory IRT models, the responses of items are considered repeated measurements. So item answers are embedded within students. Examining the responses to the items in a multi-level framework also allows to consider the effect of the explanatory variables as a fixed or random effect across the levels (De Boeck & Wilson, 2004). Within the framework of Explanatory IRT models, the item position variable can be included in the model as a predictor variable to explain the difference between the difficulty of the items (Atar, 2011; De Boeck & Wilson, 2004; Debeer & Janssen, 2013).

There are four explanatory IRT models that are widely used. These are the Rasch model, the latent regression Rasch model (LRRM), the linear logistic test model (LLTM), and the latent regression LLTM. In the Rasch Model, while estimating the difficulties of the items and the abilities of the persons, there are no item or person properties to explain the differences in these parameters. Hence this model is referred to as "doubly descriptive model". The Latent Regression Rasch Model (LRRM) is a model in which person properties are included as explanatory variables in order to explain the differences among persons' ability levels, but no explanatory properties are included at the item level. Since this model only includes explanatory variables at the person level, it is also known as "person explanatory model". The Linear Logistic Test Model (LLTM), unlike the LRRM, includes item-level explanatory variables to explain the differences in the difficulty of items, but no explanatory properties are included at the person level. It is also known as "item explanatory model". Finally, the Latent Regression LLTM includes item and person properties simultaneously to explain the differences in item and person parameters. In addition, in this model, interactions of item and person properties can be added. This model is also called the "double explanatory model" as it includes explanatory variables at both the item and person level (De Boeck & Wilson, 2004).

The purpose of this study is to investigate items that are more sensitive to the item position effect rather than an overall effect. Therefore, the study started by examining the interaction between the item and item position on an item basis. Then, the relationship between these interactions (item-position) and individual characteristics was examined. The interaction model between the item and the position is given in equation 1.

$$logit\left[P\left(Y_{pik}=1\right)\right]=\theta_p+\beta_i+\delta_p\left(k_{pi}-1\right)+\gamma_{1i}\left(k_{pi}-1\right) \tag{1}$$

$P\left(Y_{pik}=1\right)$, is the probability of the person p giving the correct answer to the item i in the k position, β_i is the ease of item i in the reference position, k_{pi} is the position in which p person takes item i, γ_{1i} is the item-position interaction effect of item i. In the models, the random effect of θ_p is person ability and it is assumed to have a normal distribution ($\theta_p \sim N(0, \sigma_\theta^2)$). β_i is fixed effect of items, δ_p is the random effect of item position among persons. In other words, it is the deviation of the person p from the general position effect and it is assumed to have a normal distribution $((\delta_p \sim N(0, \sigma_\delta^2))$.

When individual characteristics are added to the item-position interaction, equation 1 is extended as equation 2.

$$logit\left[P\left(Y_{pik}=1\right)\right]=\theta_p+\beta_i+\delta_p\left(k_{pi}-1\right)+\gamma_{1ip}\,Z_p\left(k_{pi}-1\right) \tag{2}$$

In Equation 2, Z_p is the value of the Z property of person p (person level covariate). γ_{1ip} is the interaction of item, position, and person properties of item i.

The item position interactions were examined on an item basis with the model (Model 1) given in equation 1. Then, with equation 2, the interaction of the item position and the variables of gender (Model 2), SES (Model 3), anxiety (Model 4), and motivation (Model 5) were examined. The analysis of the study was carried out in the R program, within the framework of GDKMs, with the glmer function of the lme4 (Bates et al., 2014) package and the eirm (Bulut, 2021) package suitable for the analysis of Explanatory IRT models. The R codes used in the models are given in Table 1. Maximum Likelihood (Laplace Approximation) method was used in the estimations of the models.

Table 1

R Codes used in Models

Models	R codes used in lme4
Model 1	responses ~ -1 + items + position:items + (1 + position | id)
Model 2	responses ~ -1 + items + position:items:gender + (1 + position | id)
Model 3	responses ~ -1 + items + position:items:SES + (1 + position | id)
Model 4	responses ~ -1 + items + position:items:anxiety + (1 + position | id)
Model 5	responses ~ -1 + items + position:items:motivation + (1 + position | id)

Results

Model fit indices are given in Table 2. When AIC, BIC, and log-likelihood indexes were examined, it was seen that M3 was the best model with model data fit. This result can be interpreted as that the interaction between item position and SES is more than other variables.

Table 2

Model Fit Indices

Models	AIC	BIC	logLik	deviance
M1	64540	66145	-32091	64182
M2	64546	66151	-32094	64188
M3	64535	66140	-32088	64177
M4	64664	66269	-32153	64304
M5	64653	66258	-32147	64295

First, the interaction of each item with the position was examined. According to the results, item-position interactions were significant in 37 of 88 items and these effects were in the range of -0.549 to -0.162 logit. When the direction of these effects was examined, the significant effect in all items was negative. Locating these items one cluster later reduces probability of answering the item correctly. Table 3 shows statistically significant item-position interactions. Item-position interaction for all items are given in Appendix A.

Table 3

Item-Position Interactions

Item No	Estimate	Item No	Estimate	Item No	Estimate
3	-0.307 (0.081)***	27	-0.228 (0.080)**	55	-0.225 (0.087)***
4	-0.162 (0.079)*	29	-0.242 (0.117)*	57	-0.264 (0.081)***
5	-0.254 (0.083)**	30	-0.404 (0.078)***	64	-0.176 (0.086)*
6	-0.184 (0.077)*	31	-0.368 (0.157)*	69	-0.234 (0.095)*
7	-0.310 (0.083)***	35	-0.177 (0.083)*	72	-0.271 (0.082)**
9	-0.298 (0.079)***	37	-0.276 (0.077)***	73	-0.211 (0.078)**
10	-0.196 (0.080)*	39	-0.229 (0.086)**	74	-0.274 (0.088)**
12	-0.336 (0.082)***	40	-0.240 (0.077)**	76	-0.169 (0.081)**
13	-0.444 (0.101)***	41	-0.250 (0.080)**	79	-0.181 (0.079)*
18	-0.549 (0.192)**	51	-0.189 (0.095)*	82	-0.224 (0.090)*
20	-0.238 (0.087)**	52	-0.172 (0.085)*	83	-0.212 (0.094)*
22	-0.249 (0.079)**	54	-0.312 (0.103)**	85	-0.224 (0.088)*
23	-0.172 (0.077)*				

* p< .05; ** p< .01; *** p<.001. Standard errors of estimates are shown in parentheses.

For example, responding to the 18th item one cluster later reduces the probability of answering the item correctly by 0.549 logits. When the exponential of 0.549 on the logit scale is taken (exp(0.549)), the obtained value of 1.731 gives the odd ratio. Answering item 18 after one cluster will reduce the odds ratio to approximately 1.731. If the probability of answering this item correctly is 0.50 at the reference position, answering it one cluster later reduces the probability of answering correctly to approximately 0.36.

Figure 1 shows the relationship between the probability of answering the items correctly and the item position. Each quadruple dot (Red-Blue-Green-Purple) in the graphs shows the variation in the probability of responding to items correctly according to the item position.

Figure 1

Item-position interactions

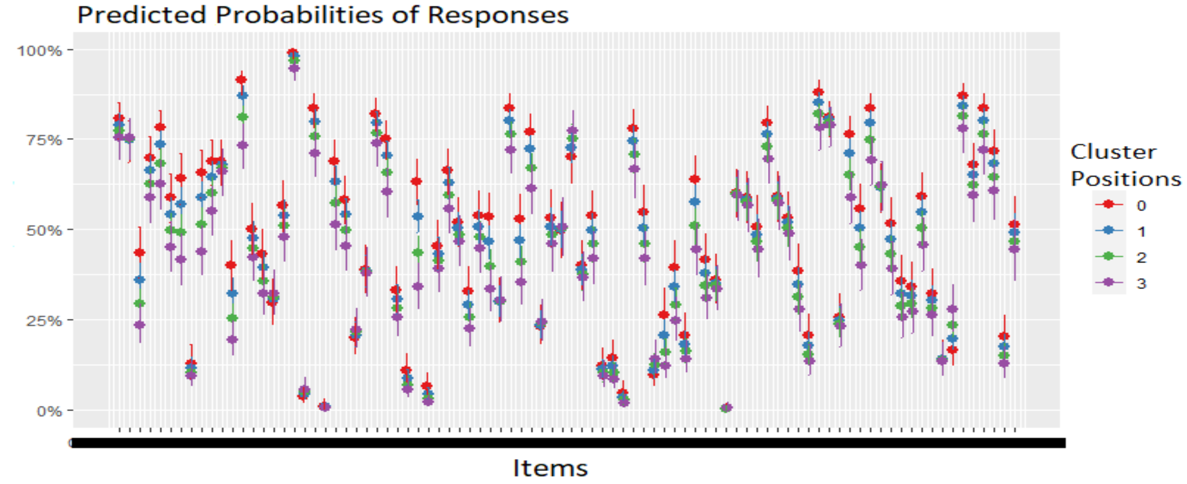


Figure 1 shows that, generally, the red dots (reference position) are at the top and the purple dots are at the bottom. In other words, the items in the reference position are more likely to be answered correctly, while the items in the last positions are less likely to be answered correctly. The points that overlap or are very close to each other can be interpreted as the probability of answering the item correctly is not affected by its position in the test. In addition, in appendix B, the graphs of the items with significant item position interactions are shown on an item basis.

The relationship between item-position interaction and gender was examined. The purpose here is to examine how the probability of answering the item changes according to the item position for different gender groups. Since females are coded as 0 and males as 1, the reference group is females. Table 4 shows the estimates for items that have significant interactions. Item-position-gender interactions for all items are given in Appendix A.

Table 4

Item-Position-Gender Interaction

ItemNo	Estimate	ItemNo	Estimate	Item No	Estimate
4	-0.167 (0.078)*	32	-0.174 (0.078)*	67	-0.386 (0.097)***
7	-0.225 (0.081)**	37	-0.198 (0.077)*	74	-0.209 (0.083)*
9	-0.153 (0.079)*	44	-0.203 (0.088)*	79	-0.253 (0.082)**
12	-0.182 (0.083)*	47	-0.210 (0.083)*	80	-0.188 (0.090)*
14	-0.155 (0.076)*	54	-0.428 (0.128)***	85	-0.199 (0.083)*
18	-0.352 (0.154)*	55	-0.183 (0.092)*	86	-0.197 (0.085)*
30	-0.169 (0.080)*	57	-0.173 (0.084)*		

* p< .05; ** p< .01; *** p<.001. Standard errors of estimates are shown in parentheses.

In 20 of the 88 items, the interaction of item, position, and gender is significant. The interaction values range from -0.428 to -0.153 logit. All significant interaction values are negative. The items that have statistically significant interaction values are answered one cluster later by male students reduces the probability of answering the items correctly. Figure 2 shows the effect of students' gender and item-position interaction on the probability of answering the item correctly.

Figure 2

Item-position-gender interaction

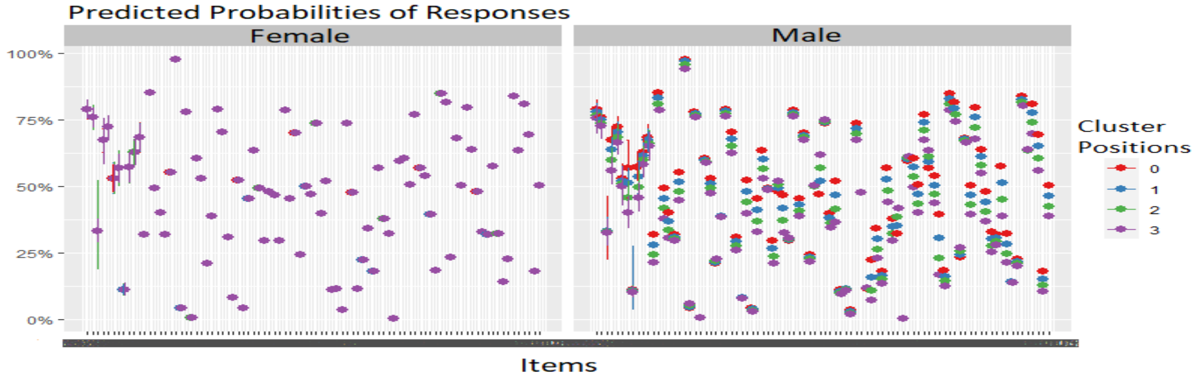


Figure 2 shows that if an item is located in the last position (purple dot), the probability of being answered correctly decreases compared to the first position (red dot). On the other hand, in items where the colors overlap, it can be interpreted that the item position effect does not differ according to gender. When female students are taken as the reference group, male students' answering the items in later positions decreases the probability of answering the items correctly. In other words, male students are more affected by item-position interaction than female students. In addition, in appendix C, the graphs of the items with significant item position and gender interactions are shown on an item basis.

The relationship between the item-position interaction and the SES levels of the students was examined. The purpose here is to examine how the probability of correct answers to the items varies according to the item position for students from different SES levels. Table 3 shows the estimates of the items that have significant interactions. Item-position-SES interactions for all items are given in Appendix A.

Table 5

Item-Position-SES Interactions

Item No	Estimate	Item No	Estimate	Item No	Estimate
3	0.144 (0.034)***	43	0.076 (0.038)*	67	0.117 (0.040)**
5	0.099 (0.031)**	47	0.087 (0.034)*	70	0.083 (0.036)*
10	0.089 (0.032)**	49	0.125 (0.052)*	72	0.092 (0.033)**
12	0.111 (0.035)***	52	0.120 (0.036)***	73	0.086 (0.033)**
13	0.092 (0.039)*	54	0.113 (0.048)*	74	0.099 (0.035)**
16	0.065 (0.033)*	55	0.084 (0.038)*	76	0.092 (0.035)**
29	0.111 (0.054)*	56	0.092 (0.045)*	77	0.095 (0.037)*
30	0.095 (0.032)**	57	0.083 (0.033)*	78	0.103 (0.038)**
36	0.083 (0.031)**	62	0.126 (0.036)***	84	0.063 (0.032)*
41	0.082 (0.032)**	66	0.087 (0.036)*	87	0.146 (0.046)**

* p< .05; ** p< .01; *** p<.001. Standard errors of estimates are shown in parentheses.

Table 5 shows that in 30 of the 88 items, the interaction of item, position, and SES is significant and the interaction estimates are in the range from 0.063 to 0.146 logit. All significant interactions are positive. Students with higher SES will be more likely to answer the item correctly as the item position increases (when the item is located in the later parts of the test).

Figure 3

Item-position-SES interaction at different SES levels

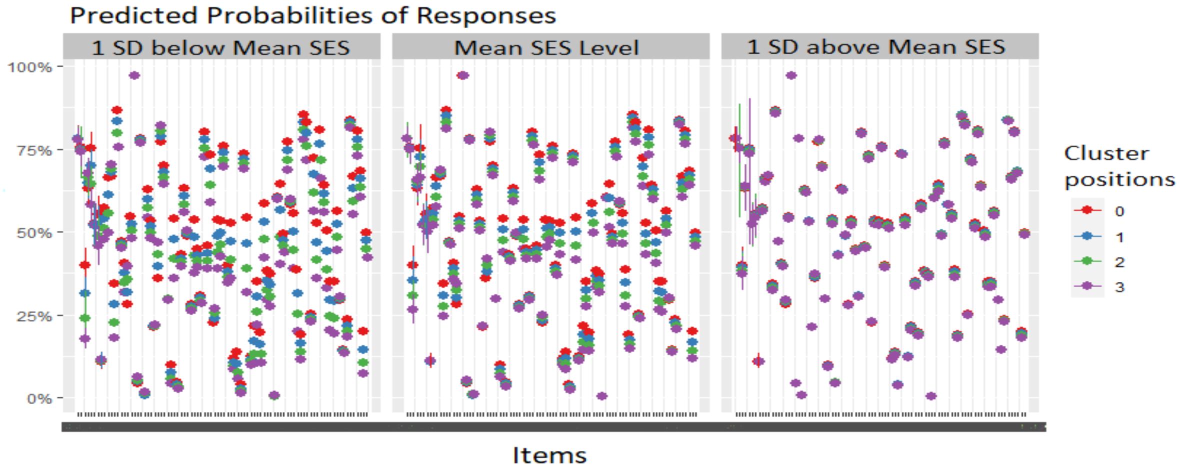


Figure 3 shows that in groups whose SES level is one standard deviation below the mean SES level, answering the item one cluster later reduces the probability of a correct answer. As the SES level of the individuals increases, the probability of answering the item correctly in the subsequent positions increases and there is almost no item position effect above one standard deviation of the mean SES level. In addition, in appendix D, the graphs of the items with significant item position and SES interactions are shown on an item basis.

How the probability of correct answers to the items changes according to the item position was examined for individuals who have different anxiety levels. Table 6 shows the estimates of the items with significant interactions. Item-position-anxiety interactions for all items are given in Appendix A.

Table 6

Item-Position-Anxiety Interaction

Item No	Estimate	Item No	Estimate	Item No	Estimate
3	-0.086 (0.043)*	13	-0.108 (0.050)*	41	-0.105 (0.043)*
5	-0.098 (0.044)*	18	-0.173 (0.086)*	53	0.141 (0.058)*
9	-0.088 (0.042)*	30	-0.103 (0.044)*	65	0.124 (0.049)*
12	-0.105 (0.049)*				

* p< .05; ** p< .01; *** p<.001. Standard errors of estimates are shown in parentheses.

Table 6 shows that in 10 of the 88 items, the interaction of item, position, and anxiety is significant and the interaction estimates ranged from -0.173 to 0.144. 8 of the 10 items that have significant interaction are negative, and 2 of them are positive. Student with higher anxiety level responding to the item one set later decreases the probability of answering correctly in 8 out of 10 items, and increases it in 2 of 10 items (items 53 and 65).

Figure 4

Item-position-anxiety interaction at different anxiety levels

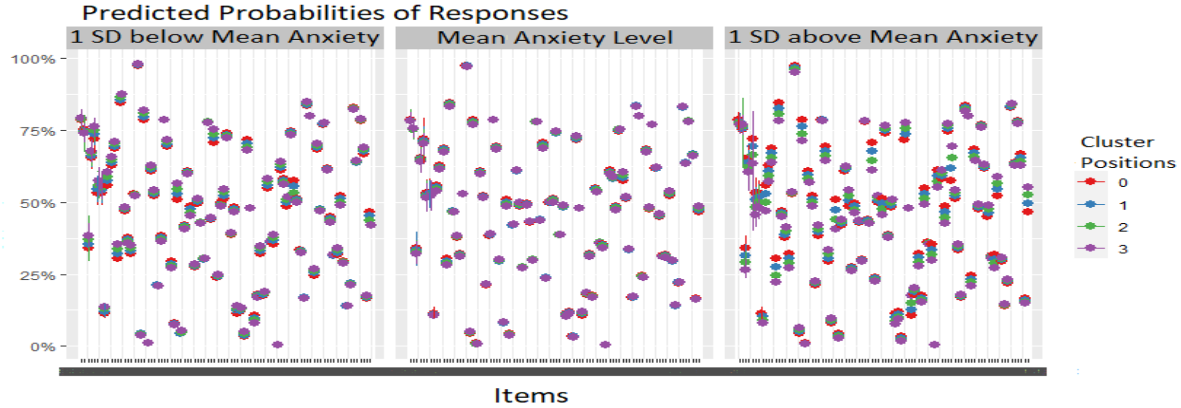


Figure 4 shows that the item position effect is less when the student's anxiety level is average, and the item position effect increases when the student's anxiety level is high. This result can be interpreted as that when the anxiety level is average, the student maintains their attention more during the test and is less affected by the effect of fatigue. In addition, in appendix E, the graphs of the items with significant item position and anxiety interactions are shown on an item basis.

How the probability of correct answers to the items changes according to the item position was examined for individuals with different motivation levels. Table 7 shows the estimates of the items that have significant interactions. Item-position-motivation interactions for all items are given in Appendix A.

Table 7

Item-Position-Motivation Interaction

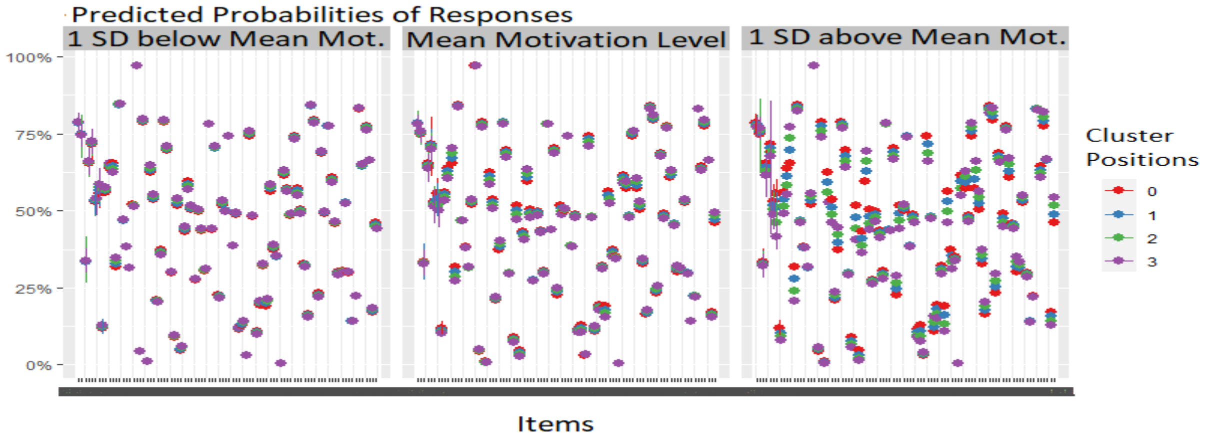
Item no	Estimate	Item no	Estimate	Item no	Estimate
7	-0.113 (0.043)**	30	-0.091 (0.041)*	33	0.089 (0.043)*
11	0.188 (0.046)*	31	-0.266 (0.078)***	56	-0.133 (0.059)*
12	-0.119 (0.046)**				

* p< .05; ** p< .01; *** p<.001, Standard errors of estimates are shown in parentheses.

In 7 of the 88 items, the interaction of item, position, and motivation was significant, and the interaction values ranged from -0.266 to 0.188. In 5 of the 7 items (items 7, 12, 30, 31, and 56), the direction of the interaction is negative. Answering these items in later positions reduces the probability of answering the item correctly. In 2 of the 7 items (items 11 and 33), the direction of the interaction is positive. Answering these items in later positions increases the probability of answering the items correctly.

Figure 5

Item-position-motivation interaction at different motivation levels



While there were 37 items with statistically significant item-position interactions, the number of items with significant interactions decreased to 7 when the motivation levels of individuals were included in this interaction. In addition, when the direction of these interactions was examined, it was seen that consistent results were not obtained. This result suggests that, among other variables used in the study, the motivation levels of individuals are less related to the item-position interaction. In addition, in appendix F, the graphs of the items with significant item position and motivation interactions are shown on an item basis.

Discussion and Conclusion

A general item position effect has been investigated in many studies in the literature (Christiansen & Janssen, 2020; Debeer & Janssen, 2013; Hahne, 2008; Meyers et al.,2009; Nagy et al., 2018; Weirich et al., 2017). Demirkol & Kelecioğlu (2022) found that there is a general item position effect in the PISA 2015 reading and mathematics data, and the probability of correct answers decreases when the items are located in later positions. But it is precious, especially for test developers and practitioners, to examine items that are more sensitive to item position effect (Albano, 2013, Bulut et al., 2017). Therefore, the purpose of this study is to examine the items that are more sensitive to the item position rather than a general item position and to investigate the relationship of item position effect with student characteristics. For this purpose, the interaction between individual characteristics and the effect of item position is examined on the basis of items, and a more detailed picture is tried to be provided. According to the results, when the item-position interaction is examined at the item level, the change in the positions of approximately 42% of the items in the test significantly affects the probability of answering the item correctly. Answering these items one cluster later reduces the probability of correct answers.

SES can be built using different variables. In the study, it was built by the variables of parental education level, home possessions, and highest parental occupation via principal component analysis. According to the results, the most important variable related to item position among the variables discussed in this study is the SES level of students. In approximately 34% of the items in the study, the item-position interaction is related to the SES level of the students. The relationship which occurs in these items increases the probability of correct answers. That is, there is a learning effect in students with high SES. In the graphs shows that the item position effect has less effect on students whose SES level is 1 standard deviation higher than the mean SES level. Given that this variable is continuous, it can be said that learning effects occur as the SES level of individuals increases; students become more familiar with the items or increase their attention levels during the test. While there are studies in the literature that support a relationship between the item position effect and SES (Nagy et al., 2018), there are also studies in which there is no stable relationship between the item position effect and SES (Wu et al., 2019).

When examining how the probability of answering the item correctly changes according to the item position for different gender groups, item-position interactions are associated with gender in 23% of the items. In the study, it is found that male students are more affected by item position than female students. This result is especially important for education politicians. The reasons for this difference between male and female students should be investigated, and necessary studies should be focused on. In studies examining the relationship between the item position effect and gender, it is found that male students are more affected by the item position effect than female students (Qian, 2014; Nagy et al., 2018; Wu et al., 2019).

When the interaction of the anxiety levels of the students with the item position is examined, it is seen that this interaction is significant in 11% of the items, and students with average anxiety levels are less affected by the item position. Increasing the level of anxiety reduces the probability of correctly answering the items in the later parts of the test. Many students are a bit anxious because of their high motivation to get a "good" score. During the test, inevitably, students' anxiety levels may increase as items become too difficult or unclear for them to answer. Therefore, when moving on to the next items, the anxiety and disappointment caused by the unsolved item may decrease the student's performance and the probability of correctly answering the items (McKeachie et al., 1955; Stanley, 1961; Cronbach, 1984). Smouse and Munz (1968) stated that students with very low or high anxiety levels are affected by item position. On the other hand, some studies found that there are no significant interactions between anxiety and item position (Berger et al., 1969; Towle & Merrill, 1975).

When the relationship between the item position effect and the motivation level of the students is examined, in approximately 8% of the items, the item-position interaction is associated with the motivation level of the students, and this relationship generally increases the probability of the correct answer to the items. However, when compared with other individual characteristics investigated in this study, the least interaction with the item position effect was seen in the motivation level of the student. The motivation variable used in this study is based on the information given by students about their own motivation levels. The scores obtained from such scales have limitations that may arise from the fact that students have deviated from the real situation (Finn, 2015; Wise & Kong, 2014). In addition, although the motivation levels of the students when they start the test are important, the item position effect may be more related to the motivation levels of the student during the test (Weirich et al., 2017). In some studies, it has been found that the most important variable related to the item position effect is motivation (Qian, 2014), while in some studies, the relationship between the item position effect and the motivation levels of students is not very clear (Wu et al., 2019).

Context and position effects are possible sources of scores independent of the test structure (Brennan, 1992). Therefore, the position of an item should be included in the measurement model as a predictor in order to examine whether the item's probability of answering the item correctly depends on the item position (Leary & Dorans, 1985; Pomplun & Ritchie, 2004). Brennan (1992) stated that “models that include the probability of the existence of context effects should be developed”. Kingston and Dorans (1984) suggested that “more general models with item position parameters should be developed”. Davey and Lee (2011) stated that “a possible direction for future analysis is to use some IRT models that can incorporate item position as a predictor”. PISA 2015 reading data was used in this study. Compared to low-stakes exams such as PISA and TIMSS, students' motivation level is higher in high-stakes exams where important decisions are made for the future of the student (Wise & DeMars, 2005). For this reason, the effect of item position on high-stakes exams can be investigated in future studies.

PISA is an exam assessing 15-year-old students. In future studies, it can be examined whether the developmental characteristics of students are effective in the item position effect. For example, whether the student's developmental characteristics (child-early-adult) are related to the item position effect can be examined with a longitudinal study. Debeer and Janssen (2013) found that linearly modeling item position effects have better model-data fit than non-linear models. Therefore, in this study, the item position effect was modeled linearly. However, an item at the beginning of the test and an item at the end of the test may not be affected by the item position effect of the same level. That is, the item position effect may increase or decrease during the test. In future studies, the item position effect can be modeled more flexibly by non-linear models.

This study has some limitations. Partially scored items in the analyses were converted to dichotomous scoring. The analysis method used in the study is suitable for dichotomously scored items. When the studies on the item position effect were examined, it was seen that the partially scored items were converted to dichotomous scoring, and the analyses were carried out in this way (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Wu et al., 2019). Therefore, converting the partially scored items to dichotomous scoring is a limitation of the study. It was observed that 42% of the items in this study were significantly affected by the item position effect. This may be due to the analysis method and scoring procedures used. In future studies, results obtained using other analysis methods and scoring procedures can be compared. In addition, item difficulties and item discriminations are estimated in PISA. However, in this study, only the effect of item position on item difficulty was investigated. In future studies, the effect of the item position on item difficulty and the effect on item discrimination can be examined.

Declarations

Author Contribution: Sinem Demirkol-Conceptualization, methodology, analysis, writing & editing, visualization. Hülya Kelecioğlu-Conceptualization, methodology, writing & editing, supervision.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

References

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50*(4), 408–426. https://doi.org/10.1111/jedm.12026

Alexandrowicz, R., & Matschinger, H. (2008). Estimating item location effects by means of a generalized logistic regression model. *Psychology Science Quarterly*, *50*(1), 64-74. https://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/1-2008/06_Alexandrowicz.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. American Psychological Association.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). American Council on Education.

Atar, B. (2011). An application of descriptive and explanatory item response models to TIMSS 2007 Turkey mathematics data. *Education and Science*, *36*(159), 256-259. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/811

Atar, B., & Cobanoglu Aktan, D. (2013). Person explanatory item response theory analysis: Latent regression two parameter logistic model. *Education and Science*, *38*(168), 59–68. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/942

Bates, D., Maechler, M., Bokler, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berger, V. F., Munz, D. C., Smouse, A. D., & Angelino, H. (1969). The effects of item difficulty sequencing and anxiety reaction type on aptitude test performance. *Journal of Psychology*, *71*(2), 253–258. https://doi.org/10.1080/00223980.1969.10543091

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, *5*(3), 225–264. https://doi.org/10.1207/s15324818ame0503_4

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89-118. https://doi.org/10.1080/08957340801926086

Bulut, O. (2021). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses, R package* (version 0.3.0) [Computer software]. http://CRAN.R-project.org/package=eirm

Bulut, O., Quo, O., & Gierl, M. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. *Large Scale in Assessments in Education, 5*(8), 1-20. https://doi.org/10.1186/s40536-017-0042-x

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement, 8*(3), 279–304. https://pubmed.ncbi.nlm.nih.gov/17804895/

Christiansen, A., & Janssen, R. (2020). Item position effects in listening but not in reading in the European Survey of Language Competences. *Educational Assessment, Evaluation and Accountability, 33*(3), 49–69. https://doi.org/10.1007/s11092-020-09335-7

Cronbach, L. J. (1984). *Essentials of psychological testing*. Harper and Row.

Davey, T., & Lee, Y. H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised general test* (RR-11-26). ETS.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Statistics for Social Science and Public Policy. Springer.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164-185. https://doi.org/10.1111/jedm.12009

DeMars, C. (2016). *Madde tepki kuramı* (Çev. Ed. H. Kelecioğlu). Nobel.

Demirkol, S., & Kelecioğlu, H. (2022). Investigating the effect of item position on person and item parameters: PISA 2015 Turkey sample. *Journal of Measurement and Evaluation in Education and Psychology, 13*(1), 69-85. https://doi.org/10.21031/epod.958576

Duncan, G. J., & Magnuson, K. A. (2005). Can family socioeconomic resources account for racial and ethnic test score gaps? *The Future of Children, 15*(1), 35-54. https://doi.org/10.1353/foc.2005.0004

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*(3), 311–326. https://doi.org/10.1080/15305050701438074

Embretson, S. E., & Reise, S. P. (2000). *Item response theory* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410605269

Finn, B. (2015). *Measuring motivation in low-stakes assessments* (RR-15–19). Educational Testing Service. https://doi.org/10.1002/ets2.12067

Grandy, J. (1987). *Characteristics of examinees who leave questions unanswered on the GRE general test under rights-only scoring* (RR-87-38). Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00242.x

Guertin, W. H. (1954). The effect of instructions and item order on the arithmetic subtest of the Wechsler-Bellevue. *Journal of Genetic Psychology, 85*(1), 79–83. https://doi.org/10.1080/00221325.1954.10532863

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*(3), 379-390. http://www.fyhe.com.au/past_papers/2006/Papers/Taylor.pdf

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

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54*(4), 418-431. https://www.proquest.com/openview/075b5103d499407933e3c62cca521618/1?pq-origsite=gscholar&cbl=43472

Hohensinn, C., Kubinger, K., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17*(6), 497-509. https://doi.org/10.1080/13803611.2011.632668

Kingston, N. M., & Dorans, N. J. (1982). The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory (RR-79-12bP). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01308.x

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*(2), 147–154. https://doi.org/10.1177/014662168400800202

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*(3), 387–413. https://www.jstor.org/stable/1170392

MacNicol, K. (1956). *Effects of varying order of item difficulty in an unspeeded verbal test*. Unpublished manuscript, Educational Testing Service.

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38-60. https://doi.org/10.1080/08957340802558342

McKeachie, W. J., Pollie, D., & Speisman, J. (1955). Relieving anxiety in classroom examinations. *Journal of Abnormal and Social Psychology*, *50*(1), 93-98. https://doi.org/10.1037/h0046560

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, *15*(3), 291–315. https://doi.org/10.1007/BF02289044

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2018). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 422–443. https://doi.org/10.1080/0969594X.2018.1449100

Organisation for Economic Co-operation and Development [OECD]. (2017). *PISA 2015 technical report.* OECD. https://www.oecd.org/pisa/data/2015-technical-report/

Pomplun, M., & Ritchie, T. (2004). An investigation of context effects for item randomization within testlets. *Journal of Educational Computing Research*, *30*(3), 243–254. https://doi.org/10.2190/Y4FU-45V7-74UN-HW4T

Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement*, *38*(7), 518–534. https://doi.org/10.1177/0146621614534312

Rose, N., Nagy, G., Nagengast, B., Frey, A., & Becker, M. (2019). Modeling multiple item context effects with generalized linear mixed models. *Frontiers in Psychology*, *10*(248), 1-13. https://doi.org/10.3389/fpsyg.2019.00248

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Smouse, A. D., & Munz, D. C. (1968). The effects of anxiety and item difficulty sequence on achievement testing scores. *Journal of Psychology, 68*(2), 181-184. https://doi.org/10.1080/00223980.1968.10543421

Stanley, J. C. (1961). Studying status vs. manipulating variables. In R. O. Collier & S. M. Elam (Eds.), *Research design and analysis*. Phi Delta Kappan.

Taylor, J. A. (2005). Poverty and student achievement. *Multicultural Education*, *12*(4), 53-55. https://www.proquest.com/openview/3c9d66c77504e21370c0b718cf66e27f/1?pq-origsite=gscholar&cbl=33246

Towle, N. J., & Merrill, P. F. (1975). Effects of anxiety type and item-difficulty sequencing on mathematics test performance. *Journal of Educational Measurement*, *12*(4), 241–249. https://www.jstor.org/stable/1434151

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, *41*(2), 115–129. https://doi.org/10.1177/0146621616676791

Whitely, E., & Dawis, R. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement*, *36*(2), 329–337. https://doi.org/10.1177/001316447603600211

Wise, L. L., Chia, W. J., & Park, R. (1989, March 27-31). *Item position effects for test of word knowledge and arithmetic reasoning*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Wu, Q., Debeer, D. Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large Scale Assessment in Education*, *7*(5), 1-20. https://doi.org/10.1186/s40536-019-0073-6

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*(4), 297–311. http://www.jstor.org/stable/1434871

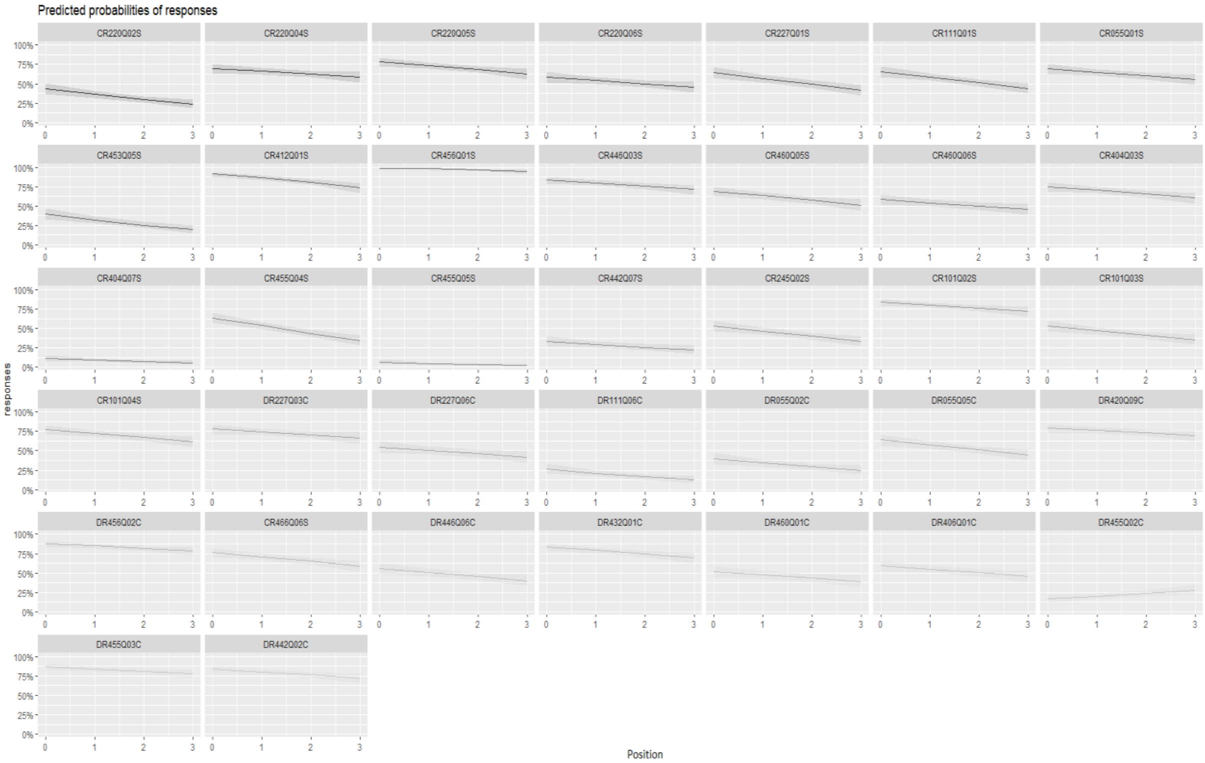
Appendix A

Item No	Item Code	Item-Position Interactions	Item-Position Gender Interactions	Item-Position SES Interactions	Item-Position Anxiety Interactions	Item-Position Motivation Interactions
1	CR067Q01S	-0.101 (0.087)	-0.068 (0.085)	0.003 (0.034)	-0.022 (0.047)	-0.015 (0.044)
2	CR102Q07S	0.014 (0.084)	-0.061 (0.083)	0.009 (0.033)	0.024 (0.046)	0.020 (0.042)
3	CR220Q02S	-0.307 (0.081)***	-0.011 (0.079)	0.144 (0.034)***	-0.086 (0.043)*	-0.012 (0.040)
4	CR220Q04S	-0.162 (0.079)*	-0.167 (0.078)*	-0.025 (0.031)	-0.051 (0.042)	-0.033 (0.040)
5	CR220Q05S	-0.254 (0.083)**	-0.093 (0.080)	0.099 (0.031)**	-0.098 (0.044)*	-0.039 (0.041)
6	CR220Q06S	-0.184 (0.077)*	-0.040 (0.075)	0.000 (0.030)	-0.074 (0.041)	-0.034 (0.039)
7	CR227Q01S	-0.310 (0.083)***	-0.225 (0.081)**	0.050 (0.032)	-0.049 (0.043)	-0.113 (0.043)**
8	CR227Q02S	-0.111 (0.111)	-0.029 (0.114)	-0.005 (0.044)	-0.095 (0.062)	-0.088 (0.060)
9	CR111Q01S	-0.298 (0.079)***	-0.153 (0.079)*	0.047 (0.031)	-0.088 (0.042)*	-0.056 (0.041)
10	CR055Q01S	-0.196 (0.080)*	-0.063 (0.078)	0.089 (0.032)**	-0.057 (0.041)	-0.070 (0.041)
11	CR453Q01S	-0.043 (0.080)	-0.050 (0.078)	-0.020 (0.033)	-0.053 (0.044)	0.118 (0.046)*
12	CR453Q05S	-0.336 (0.082)***	-0.182 (0.083)*	0.111 (0.035)**	-0.105 (0.049)*	-0.119 (0.046)**
13	CR412Q01S	-0.444 (0.101)***	-0.146 (0.093)	0.092 (0.039)*	-0.108 (0.050)*	-0.028 (0.052)
14	CR412Q05S	-0.106 (0.076)	-0.155 (0.076)*	0.010 (0.031)	-0.021 (0.043)	-0.004 (0.042)
15	CR412Q06S	-0.152 (0.077)	-0.141 (0.077)	0.048 (0.032)	0.037 (0.045)	-0.002 (0.043)
16	CR437Q01S	0.039 (0.079)	-0.038 (0.079)	0.065 (0.033)*	-0.059 (0.047)	0.004 (0.044)
17	CR437Q06S	-0.116 (0.076)	-0.140 (0.075)	0.033 (0.031)	0.008 (0.043)	0.030 (0.042)
18	CR456Q01S	-0.549 (0.192)**	-0.352 (0.154)*	0.008 (0.067)	-0.173 (0.086)*	-0.008 (0.091)
19	CR466Q03S	0.153 (0.146)	0.114 (0.149)	-0.048 (0.056)	0.095 (0.093)	0.051 (0.083)
20	CR446Q03S	-0.238 (0.087)**	-0.038 (0.086)	0.010 (0.033)	-0.095 (0.049)	-0.054 (0.047)
21	CR432Q06S	-0.169 (0.302)	-0.049 (0.343)	-0.135 (0.107)	-0.101 (0.214)	-0.176 (0.162)
22	CR460Q05S	-0.249 (0.079)**	-0.024 (0.080)	0.045 (0.030)	-0.039 (0.046)	-0.083 (0.043)
23	CR460Q06S	-0.172 (0.077)*	-0.076 (0.079)	0.027 (0.030)	-0.035 (0.046)	-0.061 (0.043)
24	CR424Q02S	0.045 (0.087)	0.031 (0.090)	-0.003 (0.035)	0.016 (0.048)	0.031 (0.047)
25	CR424Q03S	-0.012 (0.077)	-0.007 (0.081)	-0.056 (0.031)	0.038 (0.042)	0.062 (0.041)
26	CR424Q07S	-0.163 (0.086)	-0.051 (0.089)	-0.038 (0.036)	-0.003 (0.047)	-0.024 (0.045)
27	CR404Q03S	-0.228 (0.080)**	-0.116 (0.082)	0.031 (0.032)	-0.052 (0.044)	-0.052 (0.042)
28	CR404Q06S	-0.118 (0.080)	-0.081 (0.085)	0.002 (0.033)	0.045 (0.044)	-0.007 (0.043)
29	CR404Q07S	-0.242 (0.117)*	-0.019 (0.120)	0.111 (0.054)*	0.048 (0.063)	-0.100 (0.062)
30	CR455Q04S	-0.404 (0.078)***	-0.169 (0.080)*	0.095 (0.032)**	-0.103 (0.044)*	-0.091 (0.041)*
31	CR455Q05S	-0.368 (0.157)*	-0.119 (0.165)	0.072 (0.068)	-0.098 (0.088)	-0.266 (0.078)***
32	CR083Q01S	-0.085 (0.077)	-0.174 (0.078)*	0.018 (0.031)	0.020 (0.041)	-0.058 (0.041)
33	CR083Q03S	-0.147 (0.077)	-0.143 (0.076)	0.038 (0.031)	0.017 (0.041)	0.089 (0.043)*
34	CR083Q04S	-0.072 (0.076)	-0.009 (0.075)	-0.008 (0.031)	0.057 (0.041)	-0.054 (0.041)
35	CR442Q07S	-0.177 (0.083)*	-0.150 (0.084)	0.012 (0.033)	-0.017 (0.044)	-0.013 (0.046)
36	CR245Q01S	-0.120 (0.076)	0.049 (0.075)	0.083 (0.031)**	-0.032 (0.041)	-0.025 (0.041)
37	CR245Q02S	-0.276 (0.077)***	-0.198 (0.077)*	0.029 (0.031)	0.006 (0.041)	-0.018 (0.041)
38	CR101Q01S	0.006 (0.080)	0.013 (0.079)	0.014 (0.033)	-0.008 (0.043)	-0.025 (0.044)
39	CR101Q02S	-0.229 (0.086)**	-0.043 (0.084)	0.054 (0.034)	0.006 (0.046)	0.003 (0.045)
40	CR101Q03S	-0.240 (0.077)**	-0.092 (0.076)	0.036 (0.031)	-0.013 (0.041)	-0.004 (0.041)
41	CR101Q04S	-0.250 (0.080)**	-0.041 (0.079)	0.082 (0.032)**	-0.105 (0.043)*	-0.033 (0.042)
42	CR101Q05S	0.025 (0.084)	-0.047 (0.084)	-0.028 (0.033)	-0.020 (0.045)	0.063 (0.046)
43	DR219Q01EC	-0.092 (0.088)	0.003 (0.087)	0.076 (0.038)*	0.024 (0.051)	-0.061 (0.048)
44	DR219Q01C	0.013 (0.087)	0.203 (0.088)*	0.054 (0.037)	-0.055 (0.049)	0.014 (0.049)
45	DR219Q02C	0.126 (0.093)	0.018 (0.090)	0.039 (0.037)	0.035 (0.052)	-0.003 (0.048)
46	DR067Q04C	-0.048 (0.080)	-0.073 (0.082)	0.028 (0.033)	-0.009 (0.046)	-0.000 (0.044)
47	DR067Q05C	-0.157 (0.080)	-0.210 (0.083)*	0.087 (0.034)*	0.027 (0.047)	-0.022 (0.044)
48	DR102Q04C	-0.095 (0.114)	-0.068 (0.126)	0.045 (0.050)	-0.109 (0.073)	-0.062 (0.065)
49	DR102Q05C	-0.190 (0.106)	-0.022 (0.115)	0.125 (0.052)*	-0.064 (0.065)	-0.115 (0.061)
50	CR220Q01S	-0.290 (0.184)	-0.193 (0.231)	0.140 (0.096)	-0.167 (0.120)	0.052 (0.103)
51	DR227Q03C	-0.189 (0.095)*	-0.102 (0.095)	0.028 (0.038)	0.079 (0.054)	-0.080 (0.053)
52	DR227Q06C	-0.172 (0.085)*	-0.002 (0.085)	0.120 (0.036)***	0.001 (0.045)	-0.006 (0.047)

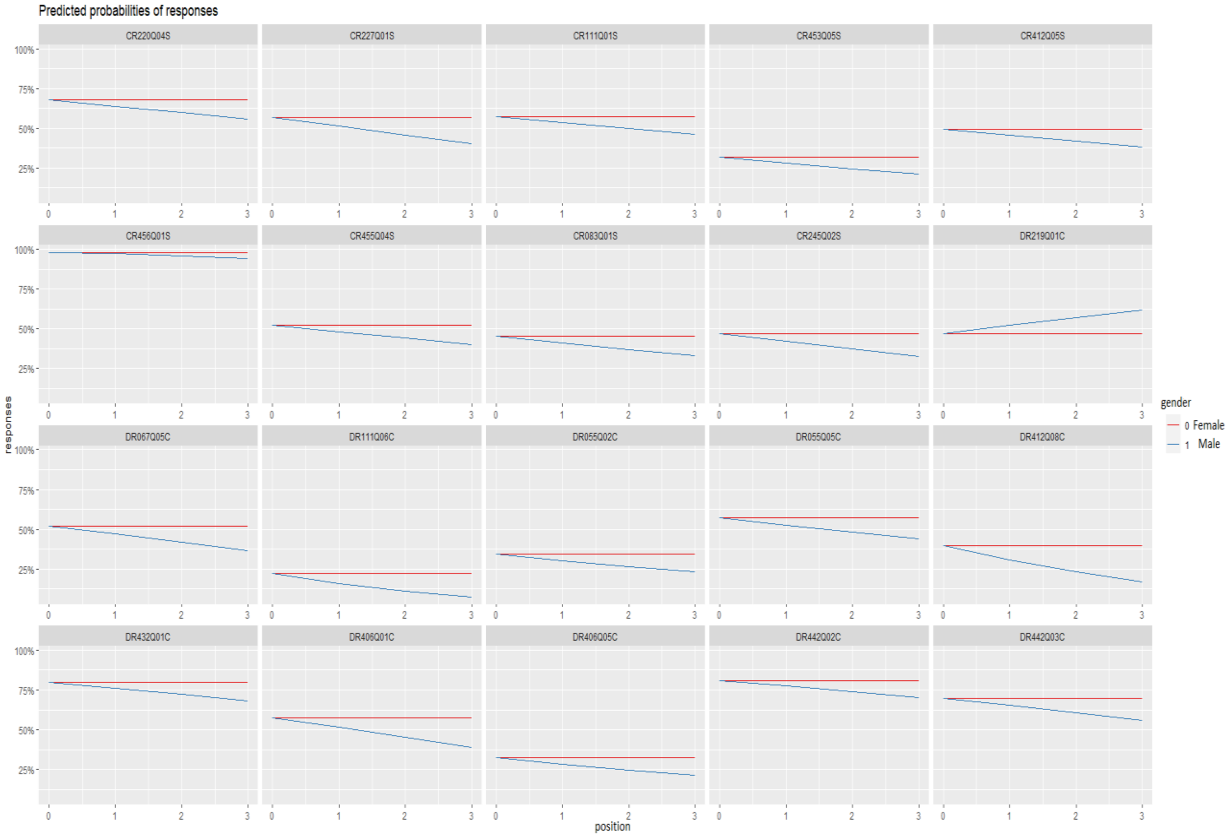
53	DR111Q02BC	0.145 (0.110)	0.002 (0.117)	0.032 (0.048)	0.141 (0.058)*	0.091 (0.063)
54	DR111Q06C	-0.312 (0.103)**	-0.428 (0.128)***	0.113 (0.048)*	0.035 (0.056)	-0.064 (0.060)
55	DR055Q02C	-0.225 (0.087)**	-0.183 (0.092)*	0.084 (0.038)*	-0.053 (0.049)	-0.025 (0.049)
56	DR055Q03C	-0.148 (0.097)	-0.111 (0.108)	0.092 (0.045)*	-0.042 (0.056)	-0.133 (0.059)*
57	DR055Q05C	-0.264 (0.081)**	-0.173 (0.084)*	0.083 (0.033)*	-0.057 (0.044)	-0.083 (0.046)
58	CR104Q01S	-0.150 (0.080)	-0.120 (0.082)	0.049 (0.034)	-0.006 (0.044)	-0.055 (0.045)
59	CR104Q02S	-0.037 (0.079)	0.140 (0.079)	0.059 (0.034)	-0.063 (0.044)	-0.012 (0.045)
60	CR104Q05S	0.143 (0.430)	0.073 (0.422)	-0.057 (0.168)	-0.053 (0.255)	-0.043 (0.253)
61	DR420Q02C	-0.008 (0.078)	0.024 (0.079)	0.003 (0.033)	-0.059 (0.044)	-0.055 (0.043)
62	DR420Q10C	-0.029 (0.082)	-0.147 (0.084)	0.126 (0.036)***	0.032 (0.050)	0.047 (0.048)
63	DR420Q06C	-0.080 (0.095)	-0.142 (0.092)	0.029 (0.039)	-0.059 (0.051)	-0.016 (0.050)
64	DR420Q09C	-0.176 (0.086)*	-0.161 (0.083)	0.055 (0.035)	0.036 (0.048)	0.047 (0.047)
65	DR453Q04C	-0.023 (0.081)	0.088 (0.085)	-0.009 (0.035)	0.124 (0.049)*	0.076 (0.047)
66	DR453Q06C	-0.060 (0.079)	-0.136 (0.081)	0.087 (0.036)*	0.026 (0.046)	0.048 (0.046)
67	DR412Q08C	-0.161 (0.086)	-0.386 (0.097)***	0.117 (0.040)**	0.022 (0.050)	0.042 (0.050)
68	DR437Q07C	-0.172 (0.095)	-0.147 (0.103)	0.074 (0.044)	0.014 (0.055)	0.054 (0.057)
69	DR456Q02C	-0.234 (0.095)*	-0.138 (0.090)	0.065 (0.037)	-0.037 (0.046)	-0.029 (0.049)
70	DR456Q06C	-0.041 (0.089)	-0.138 (0.087)	0.083 (0.036)*	0.001 (0.045)	0.052 (0.048)
71	DR466Q02C	-0.047 (0.088)	0.062 (0.090)	0.014 (0.038)	-0.053 (0.051)	0.067 (0.053)
72	CR466Q06S	-0.271 (0.082)**	-0.024 (0.080)	0.092 (0.033)**	-0.043 (0.042)	-0.029 (0.045)
73	DR446Q06C	-0.211 (0.078)**	-0.148 (0.079)	0.086 (0.033)**	0.016 (0.041)	-0.034 (0.043)
74	DR432Q01C	-0.274 (0.088)**	-0.209 (0.083)*	0.099 (0.035)**	-0.017 (0.044)	-0.013 (0.047)
75	DR432Q05C	0.008 (0.078)	-0.124 (0.078)	0.043 (0.032)	0.014 (0.041)	0.054 (0.044)
76	DR460Q01C	-0.169 (0.081)*	-0.154 (0.084)	0.092 (0.035)**	0.036 (0.044)	-0.015 (0.047)
77	DR404Q10AC	-0.155 (0.085)	-0.124 (0.091)	0.095 (0.037)*	-0.010 (0.046)	0.044 (0.048)
78	DR404Q10BC	-0.108 (0.084)	-0.059 (0.090)	0.103 (0.038)**	-0.051 (0.047)	0.026 (0.048)
79	DR406Q01C	-0.181 (0.079)*	-0.253 (0.082)**	0.061 (0.031)	0.063 (0.042)	0.017 (0.044)
80	DR406Q05C	-0.098 (0.083)	-0.188 (0.090)*	-0.006 (0.033)	0.016 (0.045)	-0.012 (0.047)
81	DR406Q02C	-0.019 (0.102)	-0.005 (0.109)	0.008 (0.043)	0.010 (0.057)	-0.007 (0.060)
82	DR455Q02C	-0.224 (0.090)*	-0.053 (0.094)	0.041 (0.037)	0.018 (0.047)	-0.000 (0.050)
83	DR455Q03C	-0.211 (0.093)*	-0.083 (0.095)	0.021 (0.037)	0.023 (0.049)	-0.005 (0.051)
84	CR083Q02S	-0.119 (0.078)	0.003 (0.076)	0.063 (0.032)*	-0.011 (0.044)	-0.030 (0.042)
85	DR442Q02C	-0.224 (0.088)*	-0.199 (0.083)*	0.054 (0.034)	-0.013 (0.048)	0.057 (0.047)
86	DR442Q03C	-0.165 (0.085)	-0.197 (0.085)*	0.044 (0.034)	-0.045 (0.048)	0.004 (0.047)
87	DR442Q05C	-0.184 (0.097)	-0.202 (0.113)	0.146 (0.046)**	-0.030 (0.057)	-0.069 (0.054)
88	DR442Q06C	-0.093 (0.090)	-0.157 (0.096)	0.038 (0.037)	0.085 (0.052)	0.067 (0.051)

Appendix B

Item-Position Interaction on Item Basis

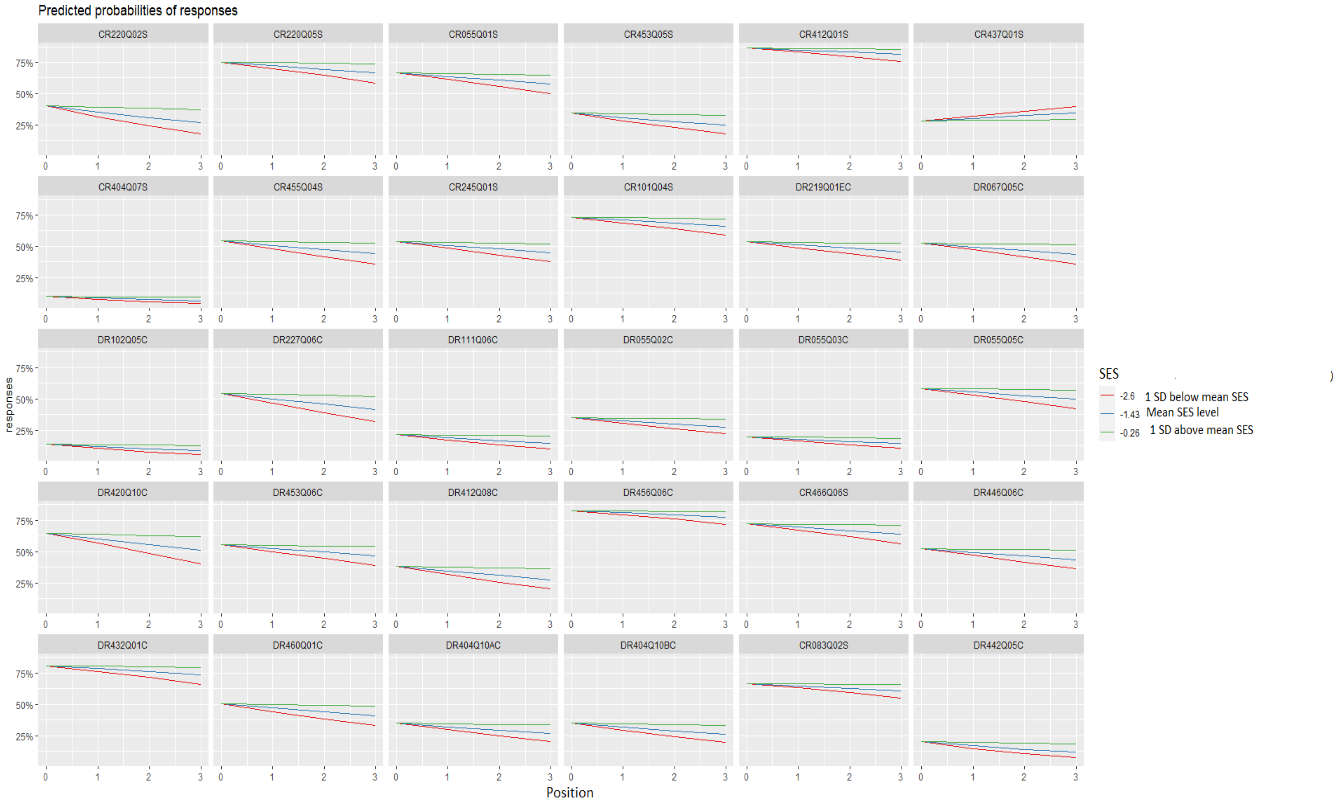


Appendix C


Predicted probabilities of responses
CR220Q04S
CR227Q01S
CR111Q01S
CR453Q05S
CR412Q05S
CR456Q01S
CR455Q04S
CR083Q01S
CR245Q02S
DR219Q01C
DR067Q05C
DR111Q06C
DR055Q02C
DR055Q05C
DR412Q08C
DR432Q01C
DR406Q01C
DR406Q05C
DR442Q02C
DR442Q03C
100%
75%
50%
25%
0
1
2
3
responses
position
gender
0 Female
1 Male

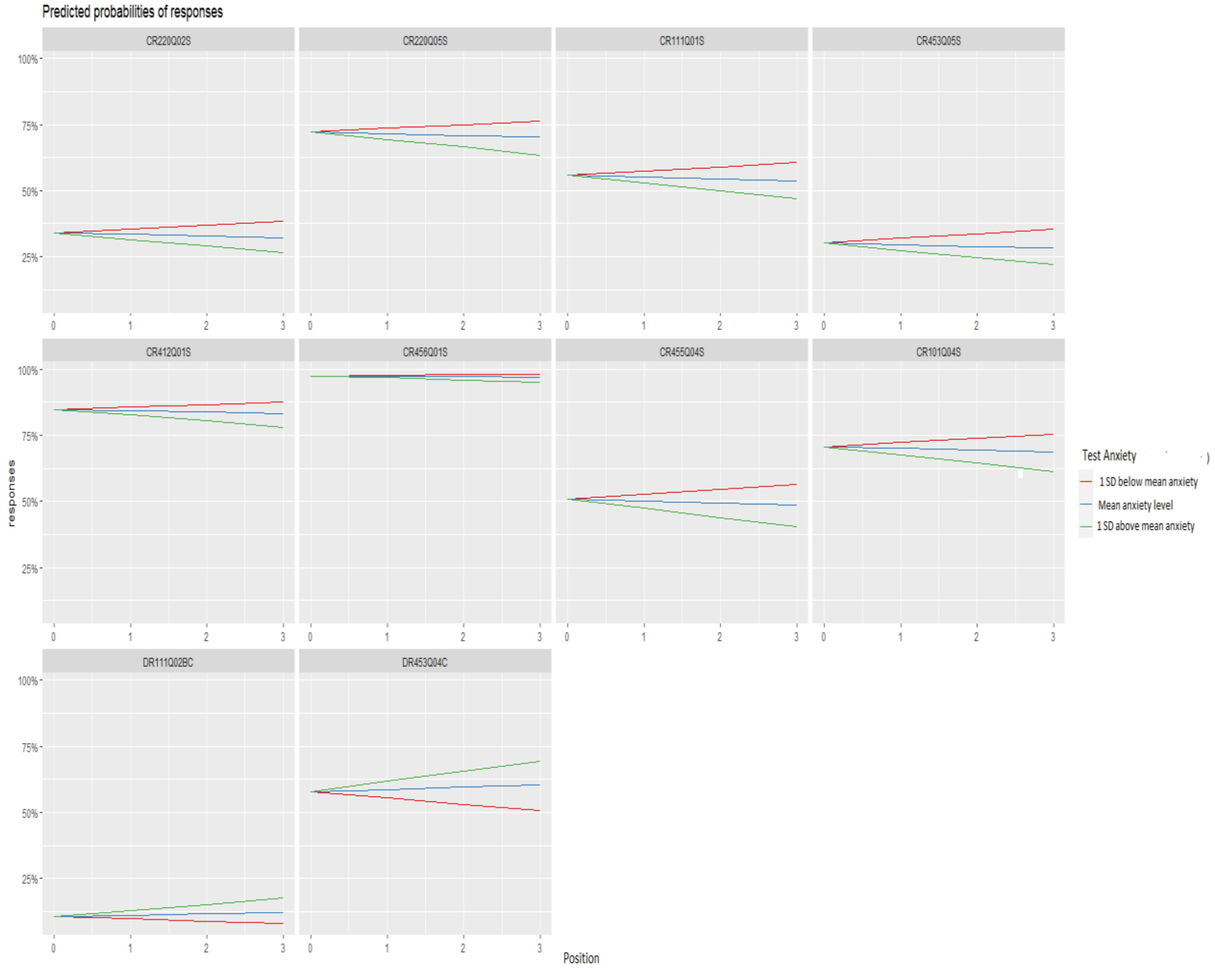
Appendix D

Item-position-SES interaction on item basis



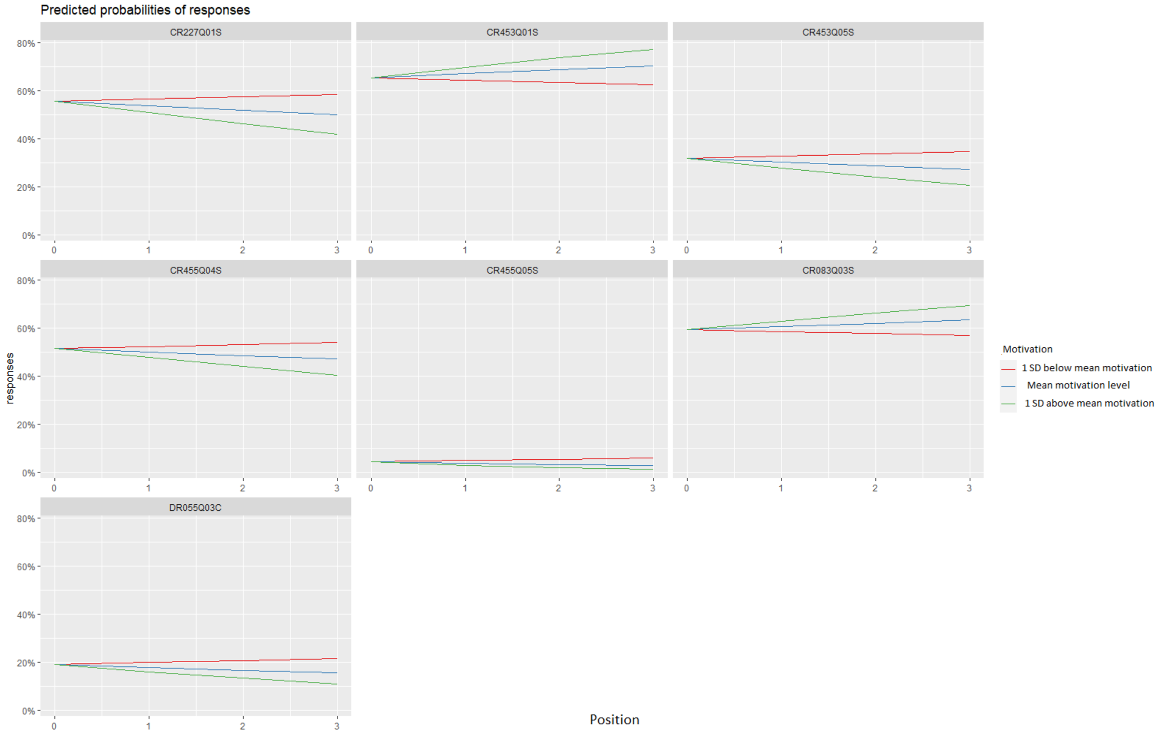
Appendix E

Item-position-anxiety interaction on item basis



Appendix F

Item-position-motivation interaction on item basis



ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(4); 305-327

Validation of the Vocabulary Size Test

Mustafa GÖKCAN * Derya ÇOBANOĞLU AKTAN **

Abstract

The Vocabulary Size Test (VST) is one of the most commonly used assessment tools for measuring English vocabulary size in the field of language testing. Despite its common usage, only a limited number of validity and reliability studies have been carried out with regard to the VST. Besides, they were mostly predicated on the Rasch model. This validation study has attempted to reveal evidence for construct validity for the VST, and to this end, item response theory (IRT) analyses were performed based on the three-parameter logistic model (3PLM). The assumptions of IRT were investigated via factor analysis (unidimensionality) and Yen's Q3 statistic (local independence). Detailed differential item functioning (DIF) analyses were conducted with Mantel-Haenszel, Lord's chi-square test, and Logistic regression methods to add evidence based on internal structure and to check fairness as a lack of measurement bias. The validation results with IRT showed that the 3PLM fitted the data better than the one- and the two-parameter logistic models. DIF results indicated that 10 items exhibited large DIF (seven favoring males and three favoring females). The results further showed that the guessing effect was not negligible for the VST.

Keywords: Language testing, vocabulary assessment, Vocabulary Size Test, item response theory, differential item functioning

Introduction

Vocabulary size is of pivotal importance in almost every aspect of learning a foreign language (Daller et al., 2007). As also echoed by Alderson (2005), "language ability is to quite a large extent a function of vocabulary size" (p. 88). Despite its importance, vocabulary size has been an oft-neglected aspect of language learning (Meara, 1980), and only recently has it drawn attention in applied linguistics and language teaching (Nation, 2013). A growing body of studies conducted on the vocabulary size of English learners indicated that it is a significant indicator of language ability (Milton, 2009). Significant positive correlations were found between English vocabulary size and listening (Li, 2019; Noreillie et al., 2018), reading (Zhang & Zhang, 2020), speaking and writing skills in English (Milton, 2013; Miralpeix & Muñoz, 2018) and, especially related to reading comprehension, it was emphasized that vocabulary size was the most significant predictor (Stæhr, 2008). Although the number of studies investigating vocabulary size has recently seen a significant increase, new assessment tools for measuring English vocabulary are rarely seen in the field (Mizumoto et al., 2019). There are also very few studies examining the validity and practicality of the available tools.

According to Standards for Educational and Psychological Testing, validity means "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." (AERA, APA & NCME, 2014, p. 11). Moreover, validity is the most fundamental characteristic of a test. It is also important to note that validity is a property of the test scores, not the test itself; in other words, having a particular validity study for a specific test does not guarantee validity in other contexts. For instance, a test can have high validity for a certain group of examinees but can have lower-level validity for other groups. It has been suggested that researchers should collect evidence for validity before they use a particular test's results for their research purposes (AERA, APA & NCME, 2014).

In language testing, the importance conferred to test validity has increased considerably in recent years. Schmitt et al. (2020) put forward that early examples of vocabulary tests generally lack appropriate

* Research Assistant, Hacettepe University, Faculty of Education, Ankara-Türkiye, gokcan.m@gmail.com, ORCID ID: 0000-0002-2284-9967
** Assistant Professor, Hacettepe University, Faculty of Education, Ankara-Türkiye, coderya@gmail.com, ORCID ID: 0000-0002-8292-3815

To cite this article:
Gökcan, M., & Çobanoğlu-Aktan, D. (2022). Validation of the vocabulary size test. *Journal of Measurement and Evaluation in Education and Psychology, 13*(4), 305-327. https://doi.org/10.21031/epod.1144808

Received: 18.07.2022
Accepted: 7.11.2022

validation examinations. They also state that “the typical practice seems to be to develop a test, get a journal article published on it, and then move on to the next project” (p. 2). If tests developed in this manner are utilized in low-stakes contexts, a lack of validation studies may not lead to any significant problems. However, since these tests are used in studies focusing on second language and foreign language acquisition, they affect theoretical and pedagogical developments (Schmitt et al., 2020).

Item response theory and language testing

Briefly stated, item response theory (IRT) models show the connection between a test item and an ability or a latent trait (indicated by the symbol “θ”) measured by that test (DeMars, 2010). The first IRT model is the normal ogive model, and the response function used in this model is given in equation 1. Birnbaum (1968) changed the normal ogive function given in equation 1 with the logistic model (equation 2), which is more statistically applicable (van der Linden & Hambleton, 1997). Working with logistic functions is easier than the normal ogive ones because the latter require mathematical integration (De Mars, 2010).

$$P_i\,(\theta)= \int_{-\infty}^{a_i(\theta-b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/z} dz. \qquad (1)$$

$$P_i\,(\theta)= \frac{1}{1+\exp\{-a_i(\theta-b_i)\}}\,. \qquad (2)$$

In the function given in equation 2, P indicates the probability of responding correctly to item “i” at a given “θ” ability level. Parameter b is the point on the ability scale where the probability of a correct response is 0.5. This parameter is also called “location parameter” and shows the position of the item characteristic curve (ICC) on the ability scale. The parameter “a” gives the curve of the ICC at the point where parameter b is located on the ability scale.

Later on, factor D was added to equation 2, and the function was formed like in equation 3. Factor D is a scaling factor and is used to make the logistic function estimates as similar as possible to normal ogive function estimates (de Ayala, 2009; Hambleton et al., 1991). If the value of factor D is equated to the constant 1.7, the logistic function is located on the same metric with the normal ogive function. By this way, for all values of θ, it is possible to get estimates differing in absolute value by less than 0.01 (Camilli, 1994).

$$P_i\,(\theta)= \frac{e^{Da_i(\theta-b_i)}}{1+\, e^{Da_i(\theta-b_i)}} \qquad (3)$$

Birnbaum suggested adding a third parameter to explain the performances of low ability individuals different from zero in multiple-choice items or tests. According to him, the scores different from zero do not result from the possibility of responding correctly. After adding the third parameter, “c”, the equation is formed as in equation 4.

$$P_i\,(\theta)= c_i + (1 - c_i)\,\frac{e^{Da_i(\theta-b_i)}}{1+\, e^{Da_i(\theta-b_i)}} \qquad (4)$$

Despite the fact that equation 4 does not indicate a logistic function anymore, the model is still known as the three-parameter logistic model (van der Linden & Hambleton, 1997). It is important to know that parameter c does not vary as a function of “θ”. For this reason, the probability of responding correctly by guessing is the same for low and high-ability individuals (Baker, 2001).

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (5)$$

The introduction of IRT to the field of language testing essentially came too late. This introduction took place with the Rasch model, and today, it can be seen that this one-parameter model is broadly used in the field (Aryadoust et al., 2020). The function that forms part of the Rasch model (Rasch, 1980) is given in equation 5. This model is less complicated than the two- or three-parameter logistic models and only comprises the difficulty parameter. The utilization of IRT models like the two- or three-parameter logistic models is more limited than the Rasch model. The Rasch – one parameter – model has generally been considered inadequate to indicate the item characteristics in full measure and has been found too simple in the community of educational measurement researchers. However, applied linguistics researchers immediately embraced the Rasch model, whose simplicity is actually quite deliberate (McNamara & Knoch, 2012).

The researchers in language testing generally do not enter the field as graduates of statistics or psychometrics but as graduates of language teaching. For this reason, their background in mathematics and statistics is not so strong, and, accordingly, it can be said that the use of IRT analyses in language testing was a little bit delayed. Moreover, unidimensionality, an IRT assumption, held language researchers back from IRT because the fact that language proficiency is multidimensional in nature and that there are lots of variables intervening in the language learning process gave them the impression that the use of IRT is not appropriate for language studies (McNamara & Knoch, 2012). However, after the introduction of multidimensional IRT models (Reckase, 2009), an approach to the effective use of IRT in language testing was opened (Ockey & Choi, 2015).

Vocabulary Size Test

The Vocabulary Size Test (VST) was developed by Nation and Beglar in 2007 to measure English vocabulary size. The VST was formed by selecting 140 words from the most frequently used 14,000 words according to the British National Corpus (BNC). Firstly, 14,000 words were split into 14 levels (1000 words in each level), and then a sample of 10 words was selected from each level. The words in the BNC are ordered according to the frequency of use in English texts. The frequency of use of a word decreases as its order in the list increases. Thus, among the 14 levels of the VST, the items in the first level are envisaged to be easier than the ones in later levels. The VST items are provided in multiple-choice format. The item stems are kept short so that any variable other than vocabulary knowledge does not affect the examinees' responses. Here is an example item from the first level.

4. FIGURE: Is this the right **figure**?

a. answer

b. place

c. time

d. number

Bilingual versions of the VST have been developed in various languages to date. Nevertheless, other than the original form of the VST, only a few studies have examined its reliability and validity. One such is a Rasch-based study carried out by Beglar (2010). Different versions of the VST, relying on the study by Beglar, have not sought further evidence for the validity of the original VST in their works. Thus, the issues in the original version have not been fully handled, and these issues have also remained in these different versions (Schmitt et al., 2020). Information related to these bilingual versions and Beglar's study is presented in the next section.

Previous Research

There are bilingual versions of the VST in Arabic, Gujarati, Japanese, Korean, Mandarin, Persian, Russian, Tamil, Thai, and Vietnamese. The development and validation studies of some of these bilingual versions have also been published in various journals with a high impact factor (Elgort, 2012; Karami, 2012; Nguyen & Nation, 2011; Zhao & Ji; 2018). Two sample items from these bilingual versions are presented below. As seen in the items, the bilingual versions have the same question stems, but the choices are in the native tongue of the respondents.

Russian Version (Elgort, 2012)	Vietnamese Version (Nguyen & Nation, 2011)
4. FIGURE: Is this the right **figure**?	4. FIGURE: Is this the right **figure**?
a. ответ	a. câu trả lời
b. место	b. địa điểm
c. время	c. thời gian
d. цифра	d. con số

As stated earlier, these bilingual studies searched for evidence of the validity of their bilingual versions, but not for the original version. They mostly relied on Beglar's Rasch-based validity study. In his study, Beglar (2010) carried out a detailed investigation into the validity of the VST, the findings of which demonstrated that most of its items showed acceptable fits to the Rasch model, and that the VST had a high degree of psychometric unidimensionality when item residuals were analyzed. The VST possesses a high degree of measurement invariance, as evidenced by the similar ability parameters produced by different forms of the VST.

Different from previous studies, Zhang (2013) investigated whether the addition of an "I don't know" option affected the scores of the VST. To this end, he applied three different versions of the VST to 150 university students in China. The first version was the original VST. The "I don't know" option was added to the second version. Additionally, in the third version, a penalty for incorrect answers was also added to the scale. The penalty comprised a one-point deduction for each wrong answer. Zhang (2013) found that the number of guesses significantly decreased in the second and third versions of the scale. But this also decreased the number of correct responses given with partial knowledge. Based on the findings, Zhang suggested that the second or the third versions of the scale be adopted, rather than the original version, to measure the precise word knowledge in order to eliminate the guessing effect.

Purposes of the Study

The purpose of this study is to collect evidence related to the validity of the Vocabulary Size Test (VST) developed by Nation and Beglar (2007) by comparing different item response theory (IRT) models. There is a particular need for a three-parameter logistic model (3PLM) validation study for the VST to examine the guessing effect, which might be conducive to overestimation in VST results (Stewart, 2014). In previous studies, the one-parameter logistic model and the Rasch model were used to validate the VST and to analyze the results. In this study, by comparing different IRT models with the three-parameter model which considers the chance factor, related gaps in the literature have been addressed.

Moreover, there is no detailed differential item functioning (DIF) study for the VST in the literature. DIF occurs when the possibility of responding correctly to a particular item differs as a function of a specific group membership. According to Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014), DIF poses a major threat to fairness in testing because it can lead to biased ability estimations. The first step in detecting item bias is detecting potentially biased items by making DIF analyses (Çepni & Kelecioğlu, 2021; Uysal et al., 2019). Historically, most DIF studies have focused on group differences based on gender or race (Kıbrıslıoğlu Uysal & Atalay Kabasakal, 2017; Zumbo, 2007). In our study, we also investigated gender-related DIF, and by detecting potentially biased items of this significant test, suggestions have been made for further studies and to improve the quality of the test.

Method

Study Group

At the beginning of the study, the intention was to collect data with a paper-pencil format VST. However, due to the pandemic, paper-pencil format data collection was not possible. For this reason, the VST data were collected in an online form by sharing a link via e-mail. The link was shared with 4500 university students from seven different universities (four state – three private). Eight hundred and fifty-four students voluntarily responded to the test. Since this number is not enough for our study, research assistants who are students of Master's and Ph.D. programs were also added to the study group. Then, 4000 research assistants were sent e-mails, 781 of whom responded to the VST. In this way, we reached a total number of 1622 voluntary students.

Data Collection

After obtaining the required permissions from the Hacettepe University Ethics committee (Document number: 35853172-300-E.00001113493) for this study, the data were collected via the 140-item version of the VST. This version can be found by following the link below: (https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-14000.pdf). The online version of the VST was generated using Google Forms. The link for the online VST version was sent to undergraduate and graduate students via e-mail. In the online form, some information about the VST was provided prior to the test. It was also stated that they could skip any questions that included items that were unfamiliar to them and that there was no time limitation. Since the participants volunteered to learn their own vocabulary levels, it was assumed that they did not cheat as they took the online test. The data collection tool also included items about the participants' demographics, such as their level of education, gender, and English proficiency test score (TOEFL or the Foreign Language Proficiency Exam).

Data Analysis

We carried out various analyses to collect evidence of the validity of our assessment tool, the VST. Throughout our validation study, we followed the suggestions offered in "Standards for Educational and Psychological Testing" (AERA, APA & NCME, 2014). As stated in the Standards, there are various sources of validity evidence that can be used to evaluate the validity of an intended interpretation of test scores for a specified use. In different settings, varying combinations of these sources might be used. There is no requirement that every single source should be used in all the validation processes. The sources of evidence listed by the Standards are content, relation to other variables (convergent validity), internal structure, response processes, and consequences of testing. In this study, we gathered evidence for all of these sources, with the exception of the consequences of testing.

Evidence Regarding Internal Structure Validity

With regard to internal structure validity, we carried out 3PLM-based IRT and DIF analyses. Internal structure validity refers to construct validity evidence. By analyzing the internal structure of a test, we examine the relationships between the test items which conform to a construct. In this study, the measured construct is the vocabulary size. By finding the best-fitting IRT model, we confirm that the items in the VST measure the vocabulary size construct.

For internal structure, IRT analyses were conducted using the R software with the "mirt" package (Chalmers, 2012). The model that the data fitted the best was tested with the ANOVA function in the same package. The unidimensionality and local independence assumptions of IRT were checked prior to IRT analyses of the VST.

In the literature, there are three commonly used methods for determining dimensionality, namely the Kaiser rule (Kaiser, 1960), parallel analysis (Horn, 1965), and scree plot (Cattel, 1966; Cho et al., 2009). Weng and Cheng (2005) showed that parallel analysis produced good estimates in the dimensionality analyses with dichotomous items, although there was a risk of obtaining meaningless dimensions. However, Tran and Formann (2009) found the reliability of parallel analysis to be too low when they worked with dichotomous items and Pearson correlation. Moreover, no improvement was observed when tetrachoric correlation was used. For this reason, for the dimensionality analysis of VST, parallel analysis was not preferred. Instead, the number of dimensions was decided by examining the scree plot and the associated eigenvalues. It was investigated whether there was a dominant dimension.

An explanatory factor analysis (EFA) was carried out and weighted least square mean and variance adjusted (WLSMV) was selected as the estimator. WLSMV utilizes tetrachoric correlation for factor extraction. When the factor analysis is carried out with continuous variables, and the data meet the assumption of univariate and multivariate normality, maximum likelihood (ML) estimation methods should be used, and when it is conducted with categorical variables, the least squares methods are recommended (Koyuncu & Kılıç, 2019). It has been found that, when compared to ML methods, WLSMV is better with large models that include categorical or binary data in terms of statistical performance and duration of the analysis (Muthen et al., 1997), and indeed that it can make less biased estimations (Li, 2016).

Yen's (1993) Q3 statistics between item pairs were calculated to test the local independence assumption. De Ayala's (2009) suggestions were followed to determine a cutoff value for the Q3 statistic. A 140x140 matrix was examined to detect potentially dependent item pairs.

Differential item functioning (DIF) analyses were run through the "difR" package (Magis et al., 2010) of R with the Logistic regression method, Lord's chi-square test, and the Mantel-Haenszel method. The difLogistic, difLord, and difMH functions were employed, and then the dichoDif function was run to make comparisons to determine the items that are flagged as DIF items by all of the three methods. The DIF statistics of the items showing large DIF are given in a table. Moreover, item characteristic curves (ICC) of these large DIF items were drawn using the R software.

Evidence Regarding Content Validity

The VST items were written as representative as possible of the English vocabulary corpus by the developers of the original version, and by this way, they provided content-related validity evidence in their work. We also investigated other sources of content validity and generated a person-item map (Wright map) in R to check whether there is a sufficient number of items in the VST and whether they spread fairly on the ability scale of the IRT model.

Evidence Regarding Convergent Validity

For evidence considering relations to other variables (convergent validity), correlations between the VST scores and the scores from two English proficiency exams, namely TOEFL and the Foreign Language Proficiency Exam (FLPE), were examined. The FLPE is a national English proficiency exam applied in Turkey. Evidence regarding response processes examines whether participants answer the questions the way the test developers intended. Although this requires collecting evidence through think-aloud processes, in this study, we indirectly collected evidence to probe the impact of responding correctly by chance by including the guessing effect in the IRT model.

Descriptive Statistics

Before presenting the findings of the study, it will be of value to review some descriptive statistics briefly. Since 165 of the participants completed the test after responding to just the first few questions, the data related to those 165 respondents were removed, and the descriptive statistics of the remaining 1457 students are presented in Table 1, Table 2, and Figure 1.

According to Table 1, 823 of the respondents are female, and 634 of them are male students. In the study group, there are 49 preparatory, 690 undergraduate, 181 master's degree, and 537 Ph.D. students. Descriptive statistics and box-plot for participants' VST scores are shown below.

Table 1

The Students' Educational Levels by Gender

	Preparatory	Undergraduate	Master	PhD	Total
Female	27	395	93	308	823
Male	22	295	88	229	634
Total	49	690	181	537	1457

According to Table 2, the mean score of females is 68.51, and it is 70.5 for males. The mean score of the entire group is calculated as 69.38. When we examine the values of skewness and kurtosis, we can see that the test score distribution does not depart from the normal distribution too much.

Table 2

Scores by Gender

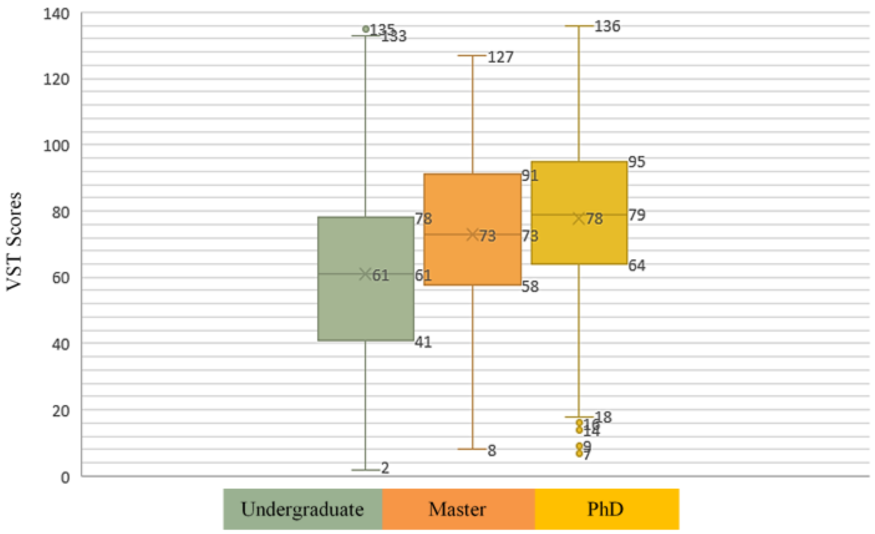
	Minimum	Maximum	Mean	Standard Error	Skewness	Kurtosis
Female	2	135	68.51	26.65	- 0.08	- 0.42
Male	5	136	70.50	27.97	- 0.09	- 0.63
Total	2	136	69.38	27.24	- 0.08	- 0.52

In Figure 1, the box and whisker plot of the VST scores of the respondents is presented. For this plot, the preparatory students are included in the undergraduate group. In the figure, the upper part of a box is the third quartile, the lower part is the first quartile, the number next to the x is the mean score, and the line represents the median.

As may be seen from the plot, all scores (quartiles, means, and medians) increase based on the educational levels of the participants. The means for undergraduate, master and Ph.D. students were found at 61, 73, and 78, respectively. This finding can be considered evidence for the fact that the VST distinguishes students from different education levels. Education level reflects the students' English proficiency to some extent because to become a research assistant and to study in graduate programs, different levels of English proficiency are required in Turkey. The proficiency level demanded for PhD programs is higher than that of master's programs, and for undergraduate programs it is too much lower. Based on these results, it is clear that the VST is also able to distinguish students at different English proficiency levels.

Figure 1

Box and Whisker Plot of the Scores and Education Levels of the Students



Data screening and cleaning were carried out prior to validation analyses. First of all, missing data and out-of-range values were checked. The questions skipped were regarded as incorrect answers because skipping a question means that the respondent does not know the meaning of the word used in the item. Since the data were collected via online forms, no missing values and no out-of-range values were found, and accordingly, there were no univariate outliers either. The Mahalanobis distance was calculated for each respondent to detect multivariate outliers. We also calculated *p*-values for every distance to see if any were statistically significant. It was found that 170 observations had p-values less than .001, and they were considered to be a multivariate outliers. They were excluded from the data, and the remaining analyses were carried out with the data of 1287 respondents.

Findings

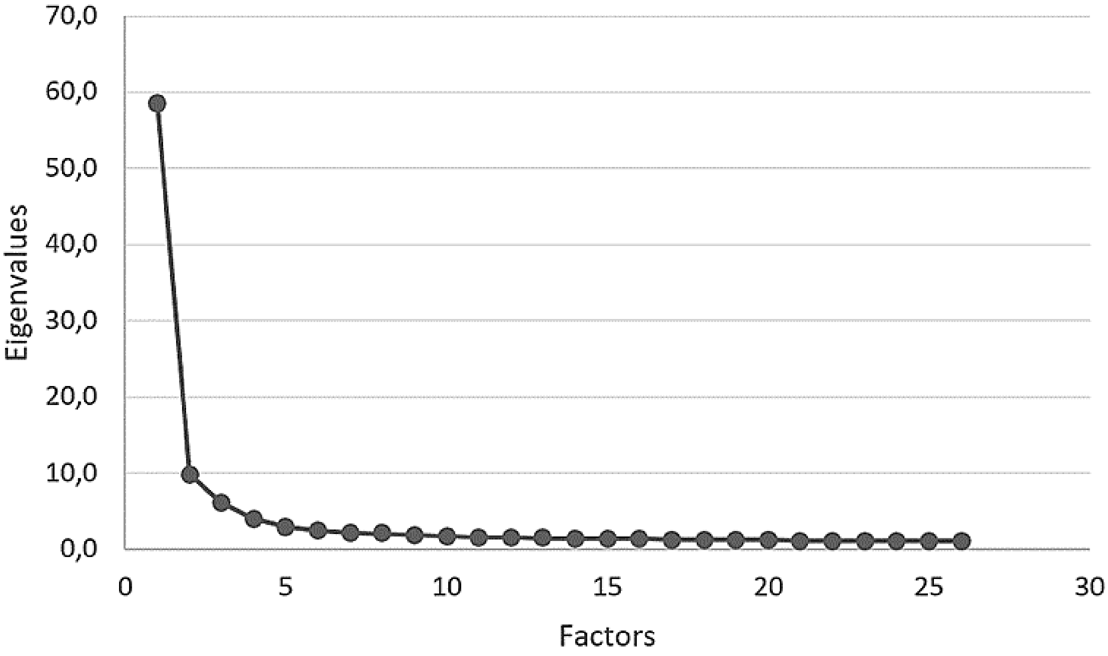
Findings of Evidence Regarding Internal Structure Validity – IRT

Before conducting the IRT analyses of the VST, we tested unidimensionality and local independence, which are two main IRT assumptions.

After the factor analysis was performed to investigate the dimensionality of the VST, it was observed that there was a dominant dimension. A dominant dimension has been considered sufficient for meeting the unidimensionality assumption in IRT analyses (Hambleton & Swaminathan, 1985). In Figure 2, the scree plot showing the eigenvalues of the factor analysis is shown. It is seen that the eigenvalue of the first dimension is almost six times bigger than the eigenvalue of the second one.

Figure 2

The Scree Plot of Eigenvalues and Factors



Moreover, the model fit indices in Table 3 illustrate that the unidimensional model fits the data well. Although the two, three, and four-dimension models had better fit indices than the unidimensional model, the VST was reckoned as unidimensional.

Table 3

Exploratory Factor Analysis Model Fit statistics

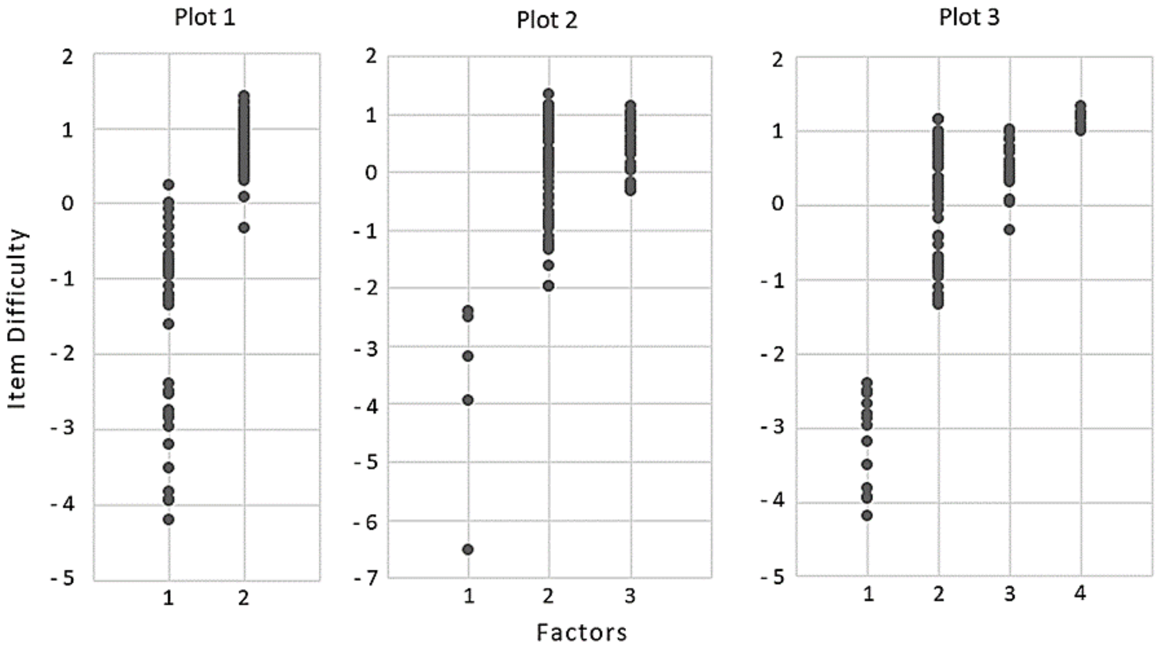
MSodel	X^2	df	X^2/df	RMSEA	CFI	TLI	SRMR
1 Factor	19167.1	9590	1.99	0.028	0.934	0.933	0.101
2 Factor	14514.7	9451	1.53	0.020	0.965	0.964	0.074
3 Factor	11842.6	9313	1.27	0.015	0.983	0.982	0.060
4 Factor	11118.9	9176	1.21	0.013	0.987	0.986	0.055

The reason behind the multidimensional findings is difficulty factors. The history of the problem of “difficulty factors” dates back to almost a century ago (Spearman, 1927; Hertzman, 1936), and it is encountered frequently in factor analyses of binary-scored items (see Hattie, 1985, for more detail). It is known that when the items of a test vary in difficulty parameter to a large extent, “spurious” factors are extracted according to item difficulty regardless of item content (McDonald & Ahlawat, 1974; Yang & Xia, 2015). This problem is generally named “spurious/artificial factors” or “difficulty factors”, and it sometimes causes simple constructs like vocabulary knowledge to seem multidimensional (Reckase et al., 1988).

In Figure 3, there are scatter plots that show the relations between each of the two, three, and four-dimensional models and the difficulty parameters of the items in related dimensions. From plot 1, it is seen that the difficulty parameters of the items in the first dimension are mostly between -4 and 0, while the ones in the second dimension are between 0 and 1.5. In the two-dimensional model, the mean of the item difficulties for the first dimension is -1.57, and it is 0.76 for the second dimension. When Plot 2 and Plot 3 were examined, it was seen that the intervals of the difficulty parameters belonging to the first dimension have lower values than those belonging to other dimensions and that the values of the intervals increase respectively for other dimensions.

Figure 3

The Scatter Plots of Item Difficulties and Factors



In the three-dimensional model, for the first dimension, the mean of item difficulties was calculated as -3.70. It is 0 for the second dimension and 0.45 for the third dimension. In the 4-dimensional model, -3.18, 0, 0.54, and 1.17 are the mean of item difficulties for the first, second, third, and fourth dimensions, respectively. Briefly, when we examine both the outputs related to dimensionality and the item contents, the reason for multidimensionality can be explained as difficulty factors.

Table 4

Locally Dependent Item Pairs

Item	Items	Item	Items	Item	Items	Item	Items
1	2,5,6,8	17	19,21	38	35		
2	1,5,6,8,19,21	19	2,17,21	42	34,43	103	26,30,70,74, 94,105,107
5	1,2,6,8	21	2,6,17,19,22	43	34,42		
6	1,2,5,8,21	22	21	62	65	105	70,94,103, 107,117
7	10	26	103	65	62	107	70,103,105
8	1,2,5,6	30	103	70	103,105,107	116	140
10	7,11	34	42,43	74	103	117	105
11	10	35	38	94	103,105	140	116

Q3 statistics, which show the relations between item residuals, were estimated to investigate the assumption of local independence. Since Q3 is a correlational statistic, its value ranges between -1 and +1, and a high absolute value of Q3 indicates a significant violation of local independence (Paek & Cole, 2020). As a cutoff value for the Q3 statistic, de Ayala (2009) suggested $|Q3|\geq \sqrt{0.5}=.2236$. Since the VST has 140 items, a 140x140 matrix (16900 cells) was examined for detecting potentially dependent item pairs. The item pairs which have Q3 statistics above .2236 were flagged as locally dependent items. It was found that among the 16900 cells, 74 of them have Q3 values higher than the cutoff value and that the 74 cells belong to 30 different items. In Table 4, these 30 items are shown in the "Item" columns, and the items which have high Q3 values with those 30 items are in the "Items" columns. We also reviewed item pairs that are potentially dependent, but couldn't see any cause of dependency. Incorrect or correct replies to a VST item should not result in incorrect or correct replies to another VST item. This is because VST items are vocabulary items, each of which asks for a different vocabulary. The item stems are very short, and there isn't any item pair which has the same item stem. Having a common passage or item stem is not the only source of dependency. According to Ackerman (1987), item parameters (i.e., discrimination and difficulty) and the order of the items (e.g., easy to hard or hard to easy) can also lead to local independence. In our case, the VST items are ordered from easy to hard, and it can be seen from Table 4 that the locally dependent item pairs are mostly neighboring items. If one wants to examine the contents of the item pairs which violate local independence, s/he can reach VST by clicking the link in the section "Assessment Tool".

Table 5

The Comparison of the 1PLM with the 2PLM

	AIC	AICc	SABIC	HQ	BIC	logLik	X^2	Df	p
1PLM	155622.6	155657.6	155902.3	155895.7	156350.2	-77670.29	NaN	NaN	NaN
2PLM	152677.1	152833.5	153232.5	153219.5	154121.9	-76058.56	3223.472	139	0

After testing the assumptions of unidimensionality and local independence, IRT analyses were carried out to determine which IRT model fits the data best. Firstly, the estimations were made with the one- and the two-parameter logistic models. Then two models were compared by conducting a likelihood ratio test and by examining AIC, AICc, SABIC, and BIC model fit indices with "ANOVA". Table 5 displays the results of the likelihood ratio test. In the two-parameter model, there are decreases in the values of AIC, SABIC, and BIC. Besides, a smaller logLik value was calculated. The *p*-value of the likelihood ratio test was estimated as zero, and this means that the 2PLM fits the data better than the 1PLM.

Table 6

The Comparison of the 2PLM with the 3PLM

	AIC	AICc	SABIC	HQ	BIC	logLik	X^2	Df	p
2PLM	152677.1	152833.5	153232.5	153219.5	154121.9	-76058.56	NaN	NaN	NaN
3PLM	152254.2	152662.5	153087.3	153067.7	154421.4	-75707.08	702.946	140	0

After it was found that the two-parameter model had a better fit than the one-parameter model, the same test was carried out to compare the 2PLM with the 3PLM. Table 6 exhibits the results of this comparison. Although there are decreases in the values of AIC, SABIC, and BIC model indices again, these decreases are not as large as in the comparison of the 1PLM and the 2PLM. Moreover, the increase in the value of logLik is not too much, but the *p*-value of the likelihood test is significant, and this indicates that the 3PLM fits the data better than the 2PLM.

Findings of Evidence Regarding Internal Structure Validity – DIF

Differential item functioning analyses were carried out with Logistic regression, Lord's chi-square test, and Mantel-Haenszel methods. There are 34 items that were flagged as DIF items by all three methods. These items are listed in Table 7, and the results of the three DIF methods are visualized in the plots given in Appendix A.

According to the results of the Logistic regression and Lord's chi-square methods, there isn't any item showing a large DIF. However, the Mantel-Haenszel results indicate that, among the 34 items, 24 items show negligible or moderate DIF, and 10 items show large DIF. If the absolute value of the Δ MH for a particular item is higher than 1.50, the item is considered to exhibit a large DIF (Magis et al., 2010). The large DIF items are items of 3, 7, 17, 20, 63, 72, 74, 98, 104, and 138 as displayed in bold in Table 7. It also shows the DIF statistics for the 34 items. The LRT statistics of the DIF items, the *p*-value related to that statistic and Nagelkerke's R^2 (Nagelkerke, 1991) are given in the results of the Logistic regression method. In the results of Lord's chi-square method, Lord's χ^2 statistic, and the *p*-value of that statistic are provided. In the results of Mantel-Haenszel, on the other hand, besides chi-square and *p*-values, α MH and Δ MH values are also presented. As shown in Table 7, the *p* values calculated in all three methods of items showing DIF are smaller than .05. We can conclude the items that show DIF in favor of males and females by examining the deltaMH values (Magis et al., 2010). When the deltaMH (Δ MH) value is negative, it indicates DIF in favor of the reference group, and when it is positive, DIF is in favor of the focal group (Holland & Thayer, 1988). Females were predetermined as the reference group in the codes written for the DIF analysis. It is seen that, among the items which exhibit large DIF, the items which have negative Δ MH values are the items of 72, 74, and 138. These items exhibit DIF in favor of females, and the vocabulary included in these items are *palette, kindergarten, and erythrocyte*, respectively. Moreover, the ICCs of these DIF items are shown in Appendix B. When the ICCs are examined, it is seen that, for females, the possibility of responding correctly to these items is higher on almost every level of the ability scale. Items 3, 7, 17, 20, 63, 98, and 104, which have positive Δ MH values, show large DIF in favor of males. The vocabulary asked in these items are *period, jump, pub,* *pro, stealth, crowbar, and counterclaim*, respectively, and the ICCs of these items are presented in Appendix C.

Table 7

Items Showing DIF and their DIF Statistics

	Logistic Regression			Lord's chi-square		Mantel-Haenszel			
ITEM	LRT Statistic	p-value	R^2	Lord's χ^2	p-value	MH χ^2	p-value	α MH	Δ MH
ITEM 3	**20.1307**	**0.0000**	**0.0305**	**14.3176**	**0.0008**	**15.5451**	**0.0001**	**0.3939**	**2.1892**
ITEM 7	**22.9162**	**0.0000**	**0.0288**	**10.7711**	**0.0046**	**15.7926**	**0.0001**	**0.4775**	**1.7370**
ITEM 14	13.8811	0.0010	0.0096	8.9157	0.0116	9.9325	0.0016	0.6513	1.0078
ITEM 16	10.7591	0.0046	0.0075	7.9392	0.0189	10.0675	0.0015	0.6629	0.9663
ITEM 17	**14.9093**	**0.0006**	**0.0297**	**9.6579**	**0.0080**	**9.9347**	**0.0016**	**0.2531**	**3.2290**
ITEM 20	**56.2356**	**0.0000**	**0.0381**	**33.9760**	**0.0000**	**49.5375**	**0.0000**	**0.4083**	**2.1048**
ITEM 25	10.8871	0.0043	0.0139	9.1278	0.0104	9.7025	0.0018	0.5439	1.4310
ITEM 31	12.1880	0.0041	0.0137	13.2703	0.0033	7.6450	0.0087	0.5606	1.3603
ITEM 32	9.3430	0.0094	0.0060	10.8923	0.0043	8.2690	0.0040	1.4813	-0.9234
ITEM 34	19.5076	0.0001	0.0107	21.1253	0.0000	15.2666	0.0001	1.7712	-1.3434
ITEM 55	12.6630	0.0018	0.0074	9.8334	0.0073	7.9740	0.0047	0.6683	0.9472
ITEM 59	14.4812	0.0007	0.0077	20.3274	0.0000	13.8536	0.0002	1.7509	-1.3163
ITEM 62	11.0667	0.0040	0.0098	12.7461	0.0017	11.9657	0.0005	1.8575	-1.4552
ITEM 63	**59.7718**	**0.0000**	**0.0313**	**42.5657**	**0.0000**	**48.6764**	**0.0000**	**0.3481**	**2.4799**
ITEM 69	8.2869	0.0159	0.0047	10.1917	0.0061	6.0796	0.0137	1.4578	-0.8857
ITEM 72	**39.5102**	**0.0000**	**0.0323**	**49.7523**	**0.0000**	**24.5789**	**0.0000**	**2.3489**	**-2.0068**
ITEM 74	**19.3315**	**0.0001**	**0.0149**	**31.8527**	**0.0000**	**15.2801**	**0.0001**	**2.5784**	**-2.2258**
ITEM 82	7.8601	0.0196	0.0037	13.1187	0.0014	5.7513	0.0165	1.4983	-0.9501
ITEM 90	6.1641	0.0459	0.0056	7.7200	0.0211	6.1266	0.0133	1.4817	-0.9241

Table 7

Items Showing DIF and their DIF Statistics (Continued)

ITEM 92	13.3572	0.0013	0.0079	10.6864	0.0048	13.1646	0.0003	0.5860	1.2560
ITEM 93	9.6763	0.0079	0.0053	12.1010	0.0024	6.4793	0.0109	1.4568	-0.8842
ITEM 98	**41.6161**	**0.0000**	**0.0234**	**30.4000**	**0.0000**	**38.0141**	**0.0000**	**0.3789**	**2.2806**
ITEM 104	**31.2653**	**0.0000**	**0.0153**	**20.8562**	**0.0000**	**33.5489**	**0.0000**	**0.4063**	**2.1163**
ITEM 109	17.3575	0.0002	0.0089	11.0125	0.0041	16.6053	0.0000	0.5471	1.4175
ITEM 114	16.1258	0.0003	0.0080	8.8901	0.0117	14.5232	0.0001	0.5423	1.4380
ITEM 115	9.4762	0.0088	0.0049	13.0773	0.0014	6.6750	0.0098	1.4819	-0.9243
ITEM 116	11.5260	0.0031	0.0113	9.2755	0.0097	7.4143	0.0065	1.6128	-1.1232
ITEM 122	9.6359	0.0081	0.0056	7.7227	0.0210	8.7972	0.0030	0.6240	1.1084
ITEM 123	9.7038	0.0078	0.0050	5.9939	0.0499	7.7186	0.0055	0.6616	0.9709
ITEM 124	10.1171	0.0064	0.0058	12.0558	0.0024	8.3024	0.0040	1.5783	-1.0724
ITEM 128	11.6936	0.0029	0.0061	16.1651	0.0003	7.1759	0.0074	1.5005	-0.9537
ITEM 130	6.0630	0.0482	0.0037	9.7588	0.0076	3.9663	0.0464	1.3197	-0.6519
ITEM 138	**24.9563**	**0.0000**	**0.0156**	**23.2532**	**0.0000**	**20.7391**	**0.0000**	**1.9070**	**-1.5170**
ITEM 140	10.6448	0.0049	0.0116	10.8352	0.0044	11.0913	0.0009	1.8808	-1.4845

Findings of Evidence Regarding Content Validity

After observing the 3PLM fits the data best, item and person parameters were estimated with the three-parameter logistic model to obtain content validity evidence. In Appendix D, the person-item map (wright-map) in which the difficulty parameters of the VST items and ability parameters of the respondents are located on the same scale is given. On this map, it is observed that the VST has a sufficient number of items in every level of ability parameter, meaning that the VST, with its 140 items, is able to measure the vocabulary size of both low and high-proficiency individuals. Among 140 items, the easiest ones are items 6., 2., and 1. Moreover, for these items, the b parameters were estimated as -6.50, -6.23, and -5.45, respectively. The most difficult items are items 96, 58, and 68, and b parameters for these items were found as 3.14, 3.09, and 3.01, respectively. The locations of these items can be seen on the person-item map.

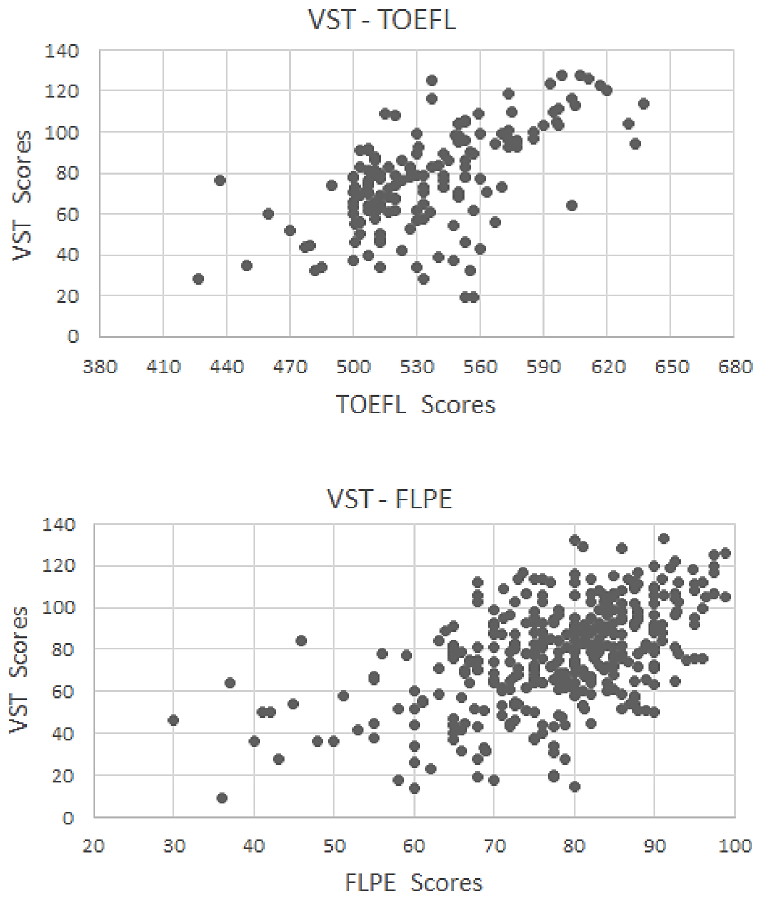
As it has been stated before, when the number of an item increases, the frequency of the word used in this item decreases, and therefore, in theory, the difficulty of the item increases, as well. On the person-item map (Appendix D), we can easily see that this theory is valid to some extent. The first questions are located on the left part of the scale, and when the sequence number of the items increases, they gradually move to the right side. However, there are some exceptions. Firstly, there are some questions which are more difficult than expected. These are items 4, 16, 58, and 68, and the vocabulary used in these items are *figure*, *nil*, *cavalier*, and *azalea*, respectively. These four items have difficulty parameters which are quite higher than the other items in their 10-word group. For instance, the mean of the difficulty parameters of the first 10 items is -3.80; however, the b parameter of the 4th item is 0.55. The departure of item 4 can be clearly seen from the person-item map. Secondly, there are approximately 20 questions which are easier than expected. These are the items of 35, 46, 47, 50, 54, 56, 61, 67, 70, 72, 74, 83, 88, 94, 103, 105, 107, 117, and 126, and the words used in these items are *quiz, cube, miniature, bacterium, accessory, thesis, olive, demography, yoghurt, palette, kindergarten, monologue, octopus, mystique, yoga, puma, aperitif, caffeine,* and *plankton*. Except for *olive* and *kindergarten*, all those easy words are loan words in Turkish, and thus these words were answered correctly quite more than the other items in their group. To illustrate, item 126 included *plankton*, and its parameter b was calculated as -1.30. However, other words in the same group have a mean of 1.70 for the same parameter. Likewise, the departures of these items from their groups can be seen on the person-item map (Appendix D).

Findings of Evidence Regarding Convergent Validity

While collecting data with the VST, we also asked the respondents the last score they obtained from an English proficiency test. Approximately 600 students responded to that question. The responses included scores on two English proficiency tests, namely TOEFL and the FLPE. We examined the relationship between their VST scores and the scores from those two English proficiency tests. One hundred and sixty of the students provided their TOEFL scores.

Figure 4

Scatter Plots of the VST Scores and Two Language Tests



The relation between the VST and TOEFL scores, and the VST and the FLPE scores can be seen in Figure 4, where scatter plots of the VST scores, the TOEFL scores, and the FLPE scores are illustrated. As may be seen from the first plot in Figure 4, there is a positive correlation between the VST and

TOEFL scores. The correlation coefficient for these variables was calculated as 0.60 (95% CI = 0.49, 0.69). Three hundred and sixty-eight of the respondents reported their FLPE results, and the positive correlation between the VST and the FLPE results can be seen in the second plot. The correlation coefficient was found as 0.53 (95% CI = 0.45, 0.60) for these two. Two high correlations (Cohen, 1992) indicate that VST scores relate closely to other measures of English proficiency, and this provides convergent evidence for the validity of the VST.

Discussion and Conclusion

In this study, to validate the VST, we collected validity evidence based on Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014). The sources of evidence for validity that we investigated included content, relation to other variables, internal structure, and response processes.

In order to examine content validity through a person-item map, we checked whether there were a sufficient number of items in the VST and whether they were distributed moderately on the ability scale. The results showed that the VST items possessed a wide variety of difficulty parameters which were located on almost every level of the ability scale, meaning that the VST distinguished high-proficiency students from the low-proficiency ones and had an appropriate number of questions for every level of θ. For internal structure validity, we carried out 3PLM-based IRT and DIF analyses. The internal structure of the VST was found to be unidimensional with the EFA. Moreover, the data and model fit of the VST scores were modelled via the IRT. In the IRT analyses, to determine the best-fitting model, the log-likelihood, its *p*-value, and other model fit indices were considered. The 3PLM IRT model was found to be the best-fitting model. DIF results revealed that there were 10 items which showed large DIF. In addition to the items which showed large gender-related DIF, some questions were identified as potentially problematic because they were easier than their difficulty level. These questions included loan words like yoghurt, microphone, or kindergarten. After appealing to expert opinions, the removal of these questions from the test might be considered to avoid inaccurate estimations of students' vocabulary size and item parameters. For relations to other variables' validity evidence (convergent validity), correlations between the VST scores with the TOEFL and the FLPE scores were examined. Convergent validity analysis revealed that there were high positive correlations between the VST scores and both of these exams. By using the 3PLM model, which also investigated the guessing effect, we indirectly gathered evidence for the response processes.

One of the fundamental properties that a measurement tool should have is validity. By collecting the validity evidence provided above to validate the VST, we contribute to the literature. In his study, Beglar (2010) administered the whole VST to high-proficiency students, but the middle- and low-proficiency groups took different versions of the VST which had fewer items. In our study, we gave the 140-item version of the VST to all participants regardless of their English proficiency levels. In our conditions, the VST was found to represent a valid measurement tool. When the VST is intended to be used in a computer adaptive test, in line with our findings, it is suggested that the 3PLM should be used for the CAT estimations and calculations.

The finding that the 3PLM fitted the VST data better than the one- and two-parameter models also indicates that the guessing effect does exist in answering the VST items, and some precautions suggested in the literature (Stewart, 2014; Zhang, 2013) like increasing the number of distractors, or adding an "I don't know" option, should be considered to decrease this effect.

Declaration

Author Contribution: Mustafa Gökcan: Conceptualization, methodology, investigation, formal analysis, data curation, writing - original draft. Derya Çobanoğlu Aktan: Conceptualization, methodology, investigation, data curation, validation, supervision, writing - review & editing.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: This study was approved by the Ethics Boards and Commissions of Hacettepe University (date: 16.06.2020, document number: 35853172-300-E.00001113493). This paper presents some of the results obtained during the Doctoral Thesis process under the supervision of Asst. Prof. Derya Çobanoğlu Aktan.

References

Ackerman, T. A. (1987, April). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* Paper presented at the annual meeting of the American Educational Research Association. Washington, DC.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum. https://doi.org/10.5040/9781474212151

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Baker, F. (2001). *The basics of Item response theory.* ERIC Clearinghouse on Assessment and Evaluation.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Birnbaum A. (1968) Some Latent Trait Models, In Lord F.M., & Novick M.R. (eds.), *Statistical Theories of Mental Test Scores.* Addison-Wesley.

Camilli, G. (1994). Origin of the Scaling Constant d = 1.7 in Item Response Theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293-295. https://doi.org/10.2307/1165298

Cattell, R. B. (1966). The Scree Test for The Number of Factors. *Multivariate Behavioral Research, 1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the Parallel Analysis Procedure With Polychoric Correlations. *Educational and Psychological Measurement, 69*(5), 748–759. https://doi.org/10.1177/0013164409332229

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037//0033-2909.112.1.155

Çepni, Z. & Kelecioğlu, H. (2021). Detecting Differential Item Functioning Using SIBTEST, MH, LR and IRT Methods. *Journal of Measurement and Evaluation in Education and Psychology, 12*(3), 267-285. https://doi.org/10.21031/epod.988879

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667268

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

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing*, *30*(2), 253–272. https://doi.org/10.1177/0265532212459028

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

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

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139–164. https://doi.org/10.1177/014662168500900204

Hertzman, M. (1936). The effects of the relative difficulty of mental tests on patterns of mental organization. *Archives of Psychology, 197*.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Lawrence Erlbaum.

Horn, J. L. (1965). A Rationale and Test for The Number of Factors in Factor Analysis. *Psychometrika, 30*, 179–185. https://doi.org/10.1007/BF02289447

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal*, *43*(1), 53–67. https://doi.org/10.1177/0033688212439359

Kıbrıslıoğlu Uysal, N., & Atalay Kabasakal, K. (2017). The effect of background variables on gender related differential item functioning. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 373-390. https://doi.org/10.21031/epod.333451

Koyuncu, İ., & Kılıç, A. (2019). The use of exploratory and confirmatory factor analyses: A document analysis. *Education and Science*, *44*(198). http://dx.doi.org/10.15390/EB.2019.7665

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*, 936–949. https://doi.org/10.3758/s13428-015-0619-7

Li, C. H. (2019). Using a Listening Vocabulary Levels Test to explore the effect of vocabulary knowledge on GEPT listening comprehension performance. *Language Assessment Quarterly, 16*(3), 328–344. https://doi.org/10.1080/15434303.2019.1648474

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, *27*(1), 82–99. https://doi.org/10.1111/j.2044-8317.1974.tb00530.x

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching*, 13(3-4), 221-246. https://doi.org/10.1017/S0261444800008879

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Maters. https://doi.org/10.21832/9781847692092

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer. (Eds.), *L2 vocabulary acquisition, knowledge, and use: New perspectives on assessment and corpus analysis* (pp. 57-78). Eurosla Monographs Series. https://www.eurosla.org/monographs/EM02/Milton.pdf

Miralpeix, I. & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching, 56*(1), 1-24. https://doi.org/10.1515/iral-2017-0016

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the word part levels test. *Language Testing, 36*(1), 101–123. https://doi.org/10.1177/0265532217725776

Muthén, B., du Toit, S.H.C. & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished technical report. https://www.statmodel.com/download/Article_075.pdf

Nagelkerke, N. J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692. https://doi.org/10.1093/biomet/78.3.691

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13. https://jalt-publications.org/tlt/issues/2007-07_31.7

Nation, I. S. P. (2013). *Learning vocabulary in another language (2nd ed.).* Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nguyen, L. T. C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. https://doi.org/10.1177/0033688210390264

Noreillie, A. S., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages. *International Journal of Applied Linguistics*, *169*(1), 212-231. https://doi.org/10.1075/itl.00013.nor

Ockey, G. J., & Choi, I. (2015) Item Response Theory. *The Encyclopedia of Applied Linguistics*. 1-8. https://doi.org/10.1002/9781405198431.wbeal1476

Paek, I., & Cole, K. (2020). *Using R for item response theory model applications*. Routledge.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*(3), 193–203. https://doi.org/10.1111/j.1745-3984.1988.tb00302.x

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer. https://doi.org/10.1007/978-0-387-89976-3

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching, 53*(1), 109-120. https://doi.org/10.1017/S0261444819000326

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* MaCmillan.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 271–282. https://doi.org/10.1080/15434303.2014.922977

Tran, U. S., & Formann, A. K. (2009). Performance of Parallel Analysis in Retrieving Unidimensionality in the Presence of Binary Data. *Educational and Psychological Measurement*, *69*(1), 50–61. https://doi.org/10.1177/0013164408318761

Uysal, İ., Ertuna, L., Ertaş, F., G. & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(2), 133-148. https://doi.org/10.21031/epod.534312

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer.

Weng, L.-J., & Cheng, C.-P. (2005). Parallel Analysis with Unidimensional Binary Data. *Educational and Psychological Measurement*, *65(*5), 697–716. https://doi.org/10.1177/0013164404273941

Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods*, *47*(3), 756–772. https://doi.org/10.3758/s13428-014-0499-2

Yen, W.M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zhang, X. (2013). The “i don’t know” option in the vocabulary size test. *TESOL Quarterly*, *47*(4), 790–811. https://doi.org/10.1002/tesq.98

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research,26*(4), 696–725. https://doi.org/10.1177/1362168820913998

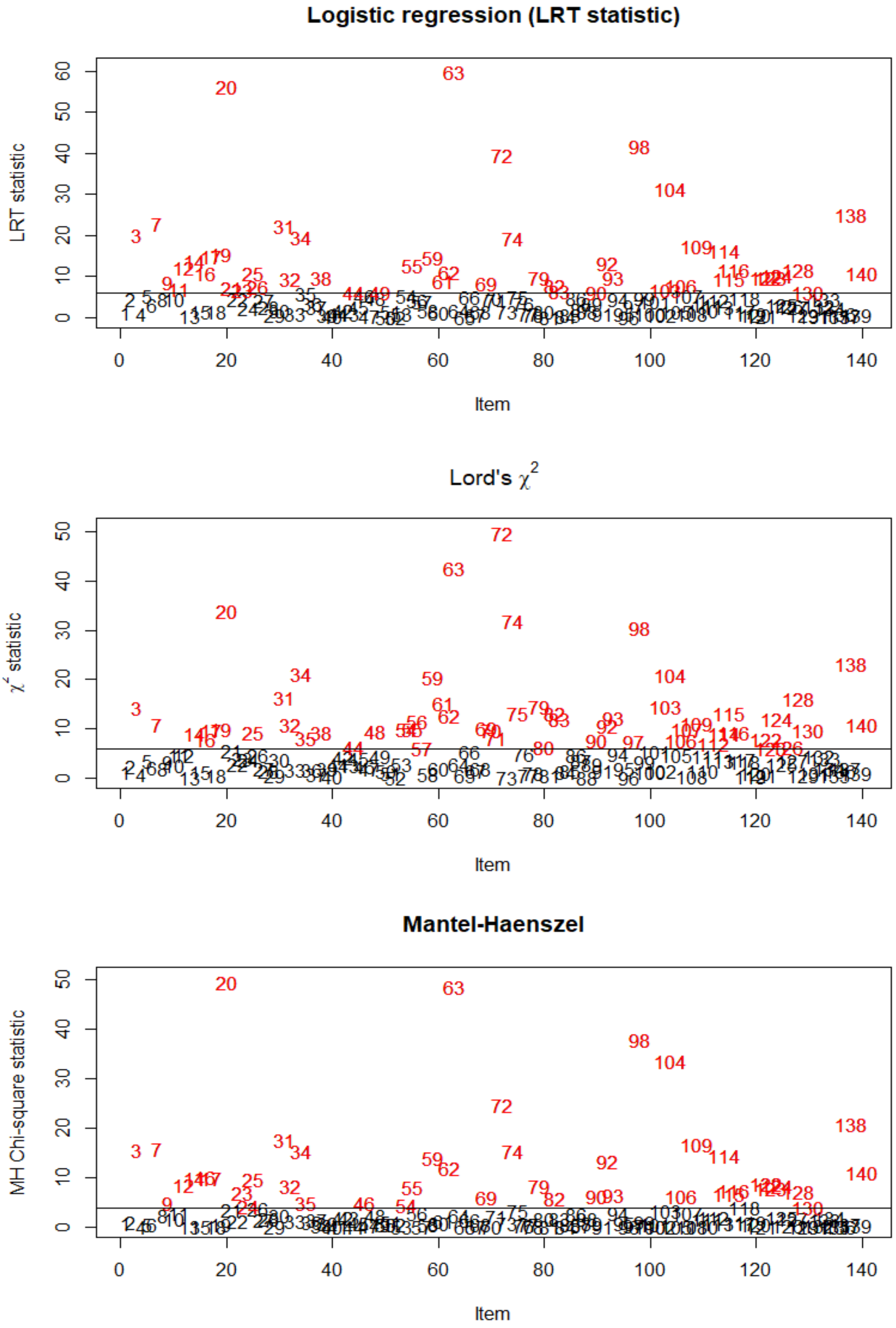
Zhao, P., & Ji, X. (2018). Validation of the Mandarin version of the vocabulary size test. *RELC Journal*, *49*(3), 308–321. https://doi.org/10.1177/0033688216639761

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Appendices

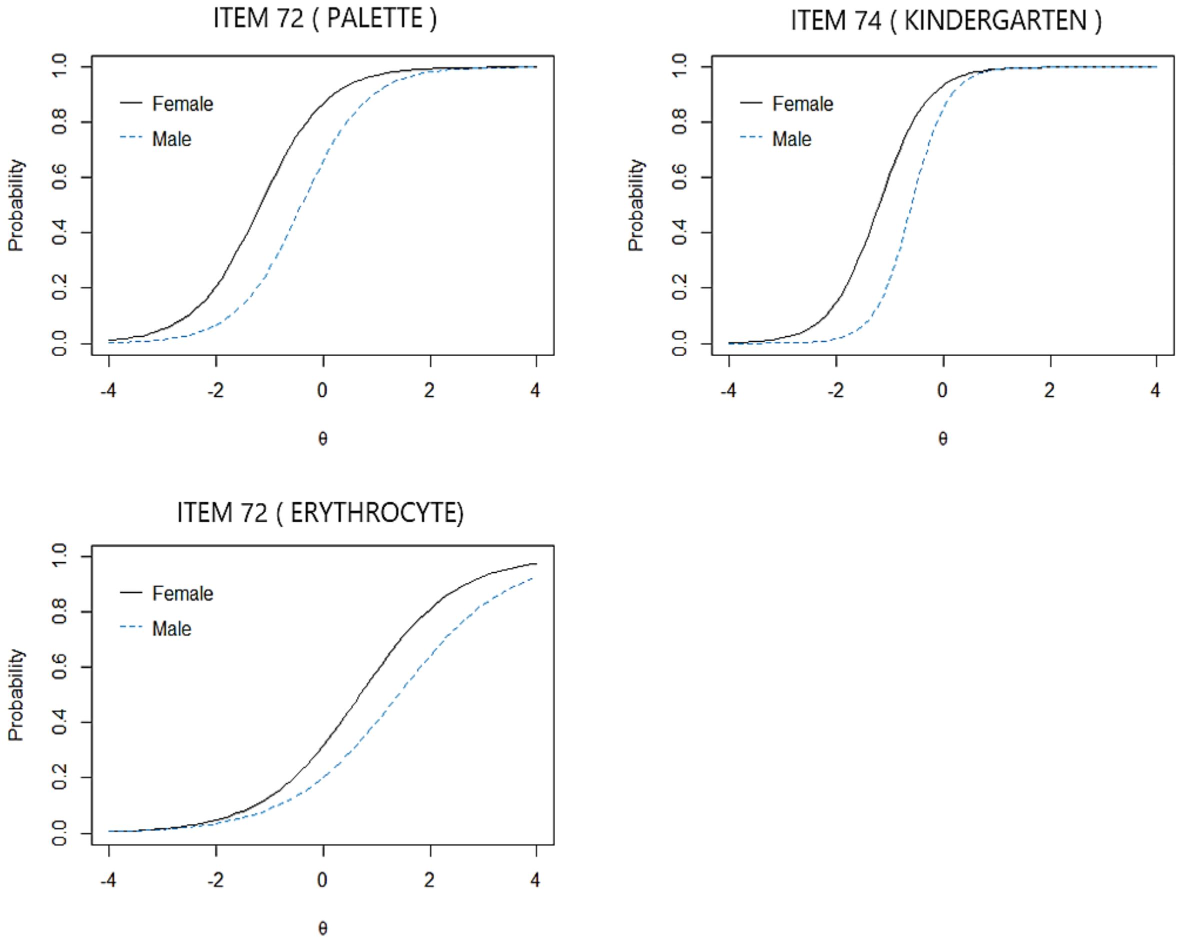
Appendix A

Plots of DIF Results



Appendix B

ICCs of the DIF Items Favoring Females



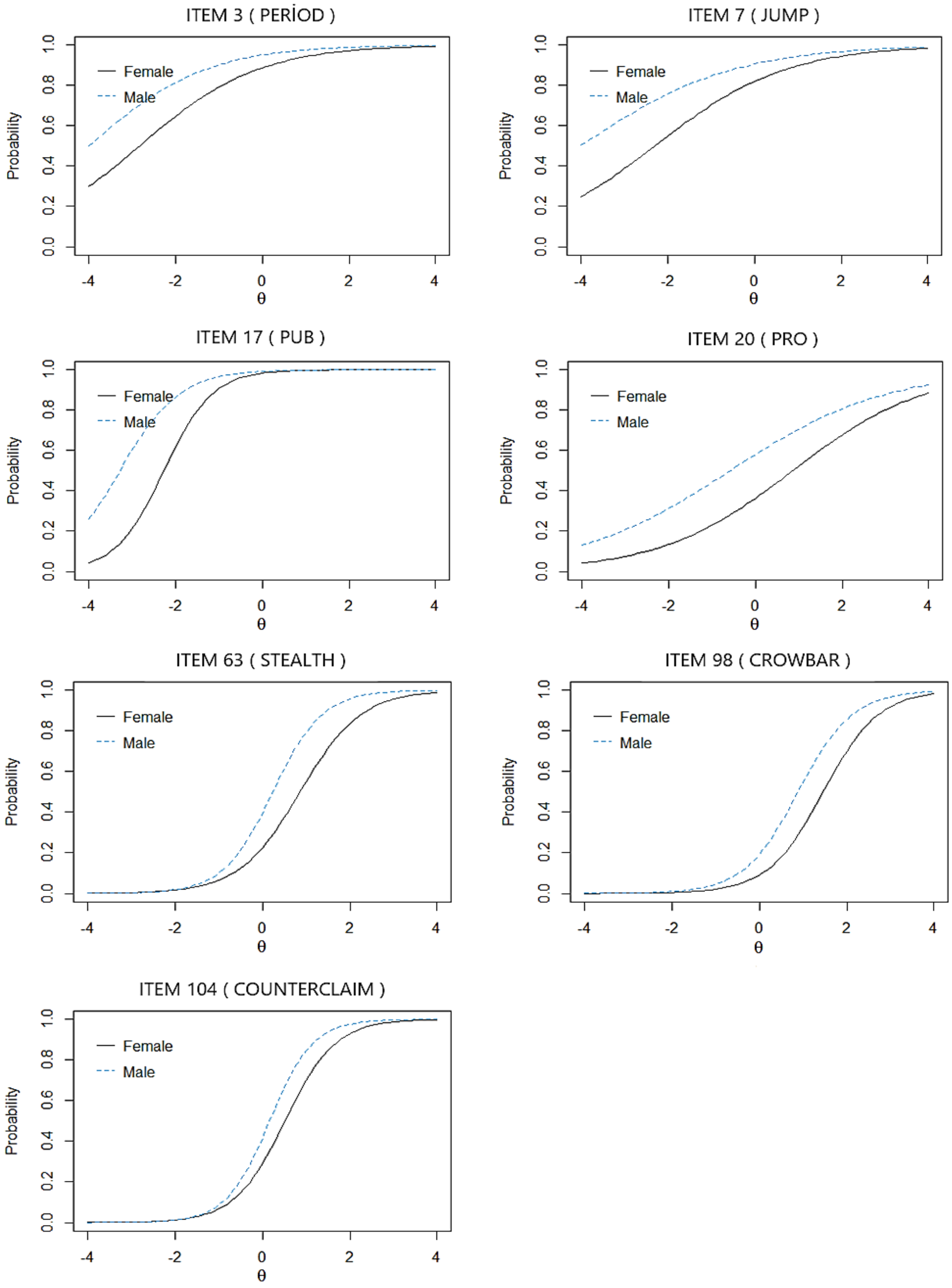
ITEM 72 (PALETTE)

ITEM 74 (KINDERGARTEN)

ITEM 72 (ERYTHROCYTE)

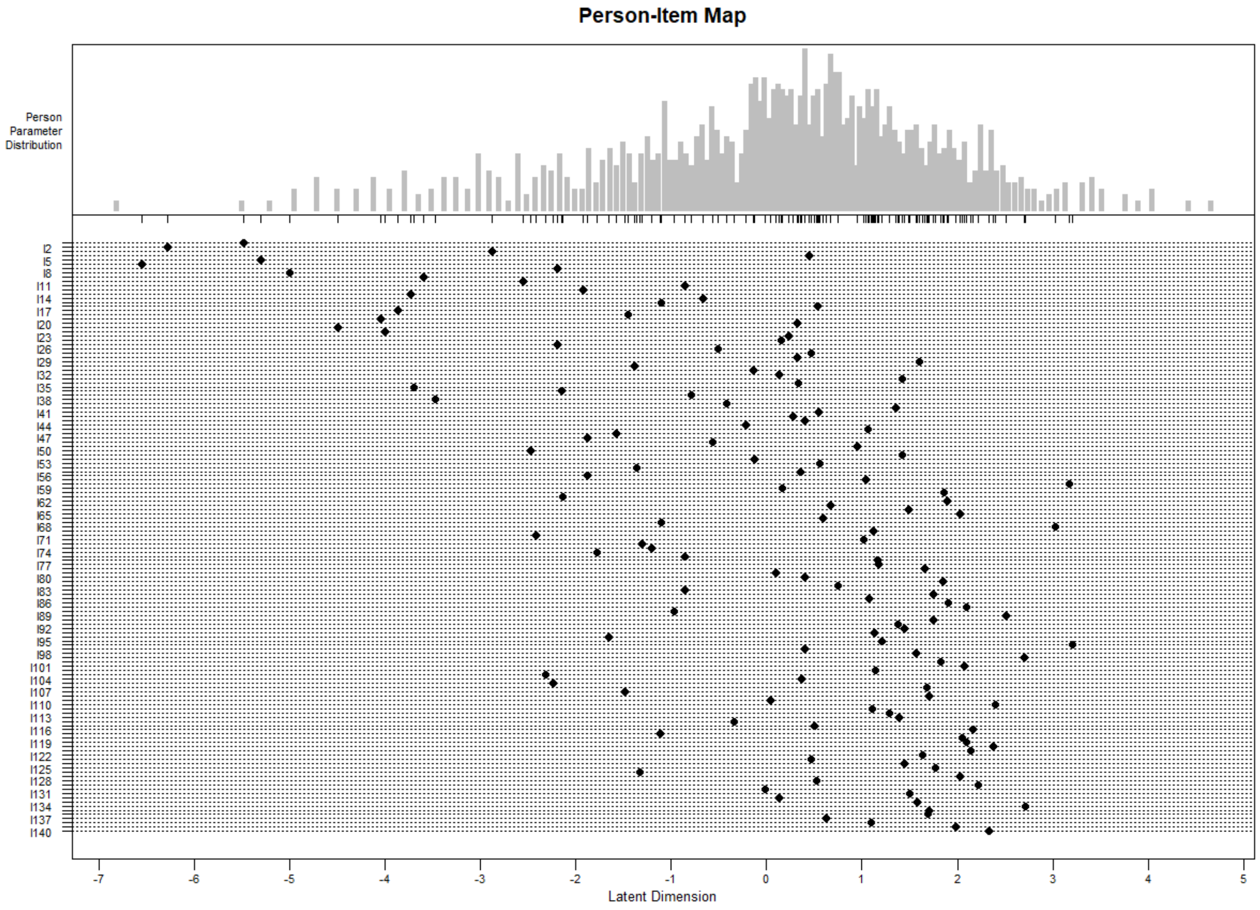
Appendix C

ICCs of the DIF Items Favoring Male



Appendix D

Person-Item Map (Wright Map)



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.

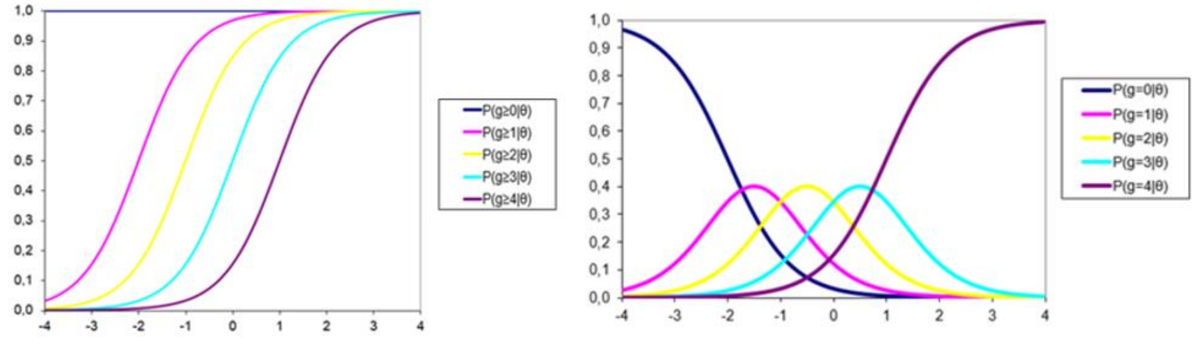
Most of the research on CAT focuses on measuring maximum performance, which mostly consists of dichotomous items (achievement, ability, etc.). However, there are relatively few CAT studies of psychological measurement instruments that require responses to polytomous items (Betz & Turner, 2011; Hol et 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).

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

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

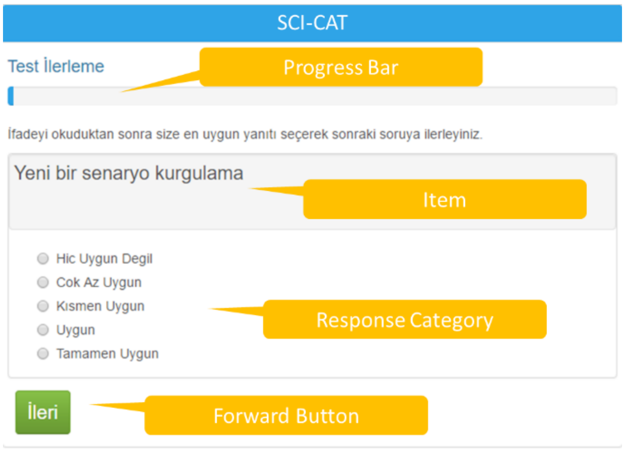
In the research, the vocational interest inventory called SCI, the Turkish version adapted by Şimşek & Tavşancıl (2022), was used to develop the CAT application. The original SCI was developed by Betz et al. (2003) as an updated version of the Strong interest inventory. The SCI paper-pencil version consists 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



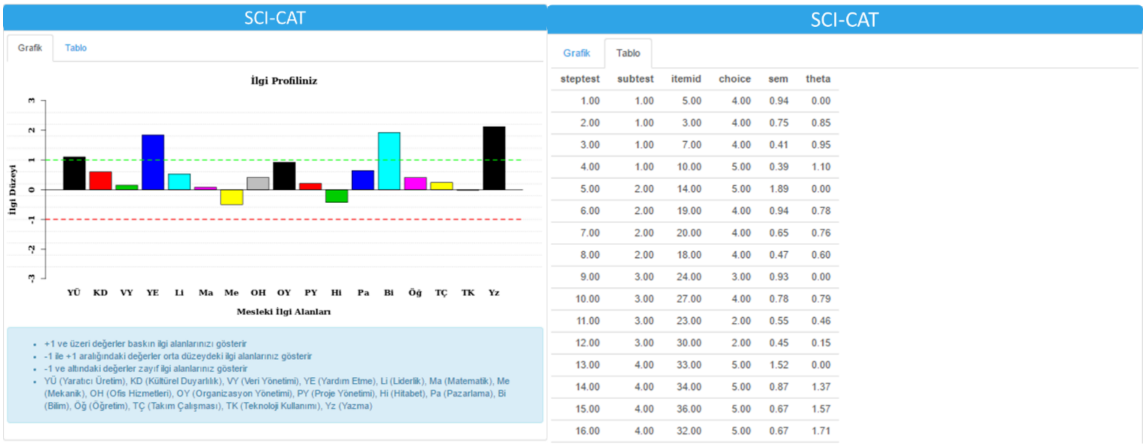
İfadeyi okuduktan sonra size en uygun yanıtı seçerek sonraki soruya ilerleyiniz.

Yeni bir senaryo kurgulama

○ Hic Uygun Degil
○ Cok Az Uygun
○ Kısmen Uygun
○ Uygun
○ Tamamen Uygun

Figure 3

SCI-CAT Result Screen



- +1 ve üzeri değerler baskın ilgi alanlarınızı gösterir
- -1 ile +1 aralığındaki değerler orta düzeydeki ilgi alanlarınız gösterir
- -1 ve altındaki değerler zayıf ilgi alanlarınız gösterir
- YÜ (Yaratıcı Üretim), KD (Kültürel Duyarlılık), VY (Veri Yönetimi), YE (Yardım Etme), Li (Liderlik), Ma (Matematik), Me (Mekanik), OH (Ofis Hizmetleri), OY (Organizasyon Yönetimi), PY (Proje Yönetimi), Hi (Hitabet), Pa (Pazarlama), Bi (Bilim), Öğ (Öğretim), TÇ (Takım Çalışması), TK (Teknoloji Kullanımı), Yz (Yazma)

steptest	subtest	itemid	choice	sem	theta
1.00	1.00	5.00	4.00	0.94	0.00
2.00	1.00	3.00	4.00	0.75	0.85
3.00	1.00	7.00	4.00	0.41	0.95
4.00	1.00	10.00	5.00	0.39	1.10
5.00	2.00	14.00	5.00	1.89	0.00
6.00	2.00	19.00	4.00	0.94	0.78
7.00	2.00	20.00	4.00	0.65	0.76
8.00	2.00	18.00	4.00	0.47	0.60
9.00	3.00	24.00	3.00	0.93	0.00
10.00	3.00	27.00	4.00	0.78	0.79
11.00	3.00	23.00	2.00	0.55	0.46
12.00	3.00	30.00	2.00	0.45	0.15
13.00	4.00	33.00	5.00	1.52	0.00
14.00	4.00	34.00	5.00	0.87	1.37
15.00	4.00	36.00	5.00	0.67	1.57
16.00	4.00	32.00	5.00	0.67	1.71

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

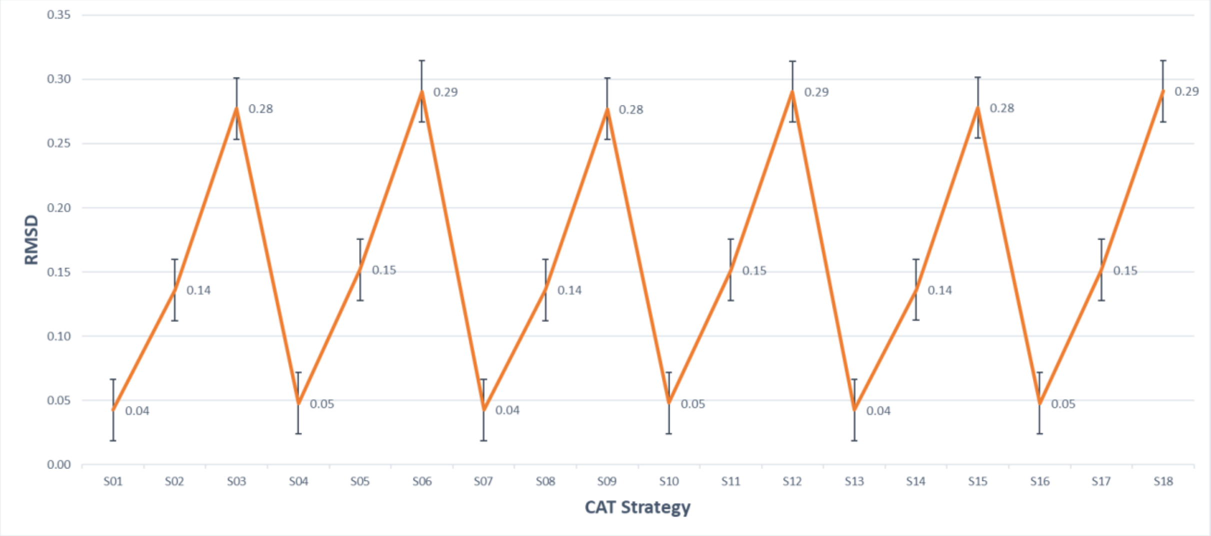


Table 4

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-SE(θ)^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-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