

e-ISSN: 2148-7456

a peer-reviewed
online journal

hosted by **DergiPark**

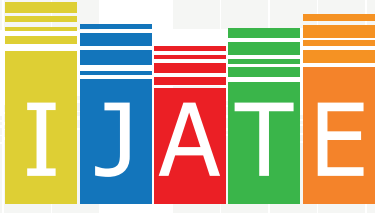
International Journal of Assessment Tools in Education

Volume: 9

Issue: 3

September 2022

<https://dergipark.org.tr/en/pub/ijate>



e-ISSN 2148-7456

<https://dergipark.org.tr/en/pub/ijate>
<http://www.ijate.net>

Volume 9

Issue 3

2022

Editor : Dr. Hakan KOGAR
Address : Akdeniz University, Faculty of Education,
Dumlupinar Bulvari, 07058, Kampus, Antalya, Türkiye
Phone : +90 242 227 4400 Extension: 6079
E-mail : ijate.editor@gmail.com; hakankogar@akdeniz.edu.tr

Publisher Info : Dr. Izzet KARA
Address : Pamukkale University, Faculty of Education,
Kinikli Yerleskesi, 20070, Denizli, Türkiye
Phone : +90 258 296 1036
Fax : +90 258 296 1200
E-mail : ikara@pau.edu.tr

Frequency : 4 issues per year (March, June, September, December)
Online ISSN : 2148-7456
Website : <http://www.ijate.net/>
<http://dergipark.org.tr/en/pub/ijate>

Journal Contact : Dr. Eren Can AYBEK
Address : Department of Educational Sciences, Pamukkale University,
Faculty of Education, Kinikli Yerleskesi, Denizli, 20070, Türkiye
E-mail : erencanaybek@gmail.com
Phone : +90 258 296 1050

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed and academic online journal.

The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).

International Journal of Assessment Tools in Education

International Journal of Assessment Tools in Education (IJATE) accepts original research on the design, analysis and use of evaluation along with assessment to enhance comprehending of the performance and quality of stakeholders in educational settings. IJATE is pleased to receive discriminating theoretical and empirical manuscripts (quantitative or qualitative) which could direct significant national and international argumentations in educational policy and practice.

IJATE as an online journal is hosted by DergiPark [TUBITAK-ULAKBIM (The Scientific and Technological Research Council of Turkey)].

In IJATE, there is no charged under any procedure for submitting or publishing an article.

Indexes and Platforms:

- Emerging Sources Citation Index (ESCI)
- Education Resources Information Center (ERIC)
- TR Index (ULAKBIM),
- EBSCO,
- SOBIAD,
- JournalTOCs,
- MIAR (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib
- Index Copernicus International

Editor

Dr. Hakan KOGAR, *Akdeniz University, Türkiye*

Section Editors

Dr. Safiye BILICAN DEMIR, *Kocaeli University, Türkiye*

Dr. Selma SENEL, *Balikesir University, Türkiye*

Dr. Esin YILMAZ KOGAR, *Nigde Omer Halisdemir University, Türkiye*

Dr. Sumeyra SOYSAL, *Necmettin Erbakan University, Türkiye*

Editorial Board

Dr. Beyza AKSU DUNYA, *Bartın University, Türkiye*

Dr. Stanislav AVSEC, *University of Ljubljana, Slovenia*

Dr. Kelly D. BRADLEY, *University of Kentucky, United States*

Dr. Okan BULUT, *University of Alberta, Canada*

Dr. Javier Fombona CADAVIECO, *University of Oviedo, Spain*

Dr. William W. COBERN, *Western Michigan University, United States*

Dr. R. Nukhet CIKRIKCI, *Istanbul Aydın University, Türkiye*

Dr. Nuri DOGAN, *Hacettepe University, Türkiye*

Dr. Selahattin GELBAL, *Hacettepe University, Türkiye*

Dr. Anne Corinne HUGGINS-MANLEY, *University of Florida, United States*

Dr. Francisco Andres JIMENEZ, *Shadow Health, Inc., United States*

Dr. Nicole KAMINSKI-OZTURK, *The University of Illinois at Chicago, United States*

Dr. Orhan KARAMUSTAFAOGLU, *Amasya University, Türkiye*

Dr. Yasemin KAYA, *Atatürk University, Türkiye*

Dr. Hulya KELECIOGLU, *Hacettepe University, Türkiye*

Dr. Omer KUTLU, *Ankara University, Türkiye*

Dr. Seongyong LEE, *BNU-HKBU United International College, China*

Dr. Sunbok LEE, *University of Houston, United States*

Dr. Froilan D. MOBO, *Ama University, Philippines*

Dr. Hamzeh MORADI, *Sun Yat-sen University, China*

Dr. Nesrin OZTURK, *Izmir Democracy University, Türkiye*

Dr. Turan PAKER, *Pamukkale University, Türkiye*

Dr. Murat Dogan SAHIN, *Anadolu University, Türkiye*

Dr. Hossein SALARIAN, *University of Tehran, Iran*

Dr. Halil İbrahim SARI, *Kilis 7 Aralık University, Türkiye*

Dr. Ragıp TERZİ, *Harran University, Türkiye*

Dr. Turgut TURKDOGAN, *Pamukkale University, Türkiye*

Dr. Ozen YILDIRIM, *Pamukkale University, Türkiye*

English Language Editors

Dr. R. Sahin ARSLAN, *Pamukkale University, Türkiye*

Dr. Hatice ALTUN, *Pamukkale University, Türkiye*

Dr. Arzu KANAT MUTLUOGLU, *Ted University, Türkiye*

Ahmet KUTUK, *Akdeniz University, Türkiye*

Editorial Assistant

Dr. Ebru BALTA, *Agri Ibrahim Cecen University, Türkiye*

PhDc. Neslihan Tuğçe OZYETER, *Kocaeli University, Türkiye*

PhDc. İbrahim Hakkı TEZCİ, *Akdeniz University, Türkiye*

Technical Assistant

Dr. Eren Can AYBEK, *Pamukkale University, Türkiye*

CONTENTS

Research Articles

Investigation of education value perception scale's psychometric properties according to CTT and IRT

Page: 548-564 PDF

Harun DİLEK, Ufuk AKBAŞ

The role of individual differences on epistemic curiosity (EC) and self-regulated learning (SRL) during e-learning: the Turkish context

Page: 565-582 PDF

Ergün AKGÜN, Enisa MEDE, Seda SARAC

Which scale short form development method is better? A Comparison of ACO, TS, and SCOFA

Page: 583-592 PDF

Hakan KOĞAR

Adaptation and psychometric evaluation of the COVID-19 stress scales in Turkish sample

Page: 593-612 PDF

Murat Doğan ŞAHİN, Sedat ŞEN, Deniz GÜLER

Exploring how the use of a simulation technique can affect EFL students' willingness to communicate

Page: 613-630 PDF

Houman BİJANİ, Masoumeh ABBASİ

Differential item functioning across gender with MIMIC modeling: PISA 2018 financial literacy items

Page: 631-653 PDF

Fatıma Münevver SAATÇIOĞLU

The study of the effect of item parameter drift on ability estimation obtained from adaptive testing under different conditions

Page: 654-681 PDF

Merve ŞAHİN KÜRŞAD, Ömay ÇOKLUK-BÖKEOĞLU, Nükhet ÇIKRIKÇI

The Effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests

Page: 682-696 PDF

Başak ERDEM KARA, Nuri DOĞAN

9. A Comparison of type I error and power rates in procedures used determining test dimensionality

Page: 697-712 PDF

Gül GÜLER Nükhet ÇIKRIKÇI

[Evaluation of impact factors of articles in scientific open access journals in Türkiye](#)

Page: 713-727 PDF

Orhan ALAV

[To what extent are item discrimination values realistic? A new index for two-dimensional structures](#)

Page: 728-740 PDF

Abdullah Faruk KILIÇ, İbrahim UYSAL

[Pamukkale critical thinking skill scale: a validity and reliability study](#)

Page: 741-771 PDF

Erdinc DURU, Sevgi OZGUNGOR, Ozen YILDIRIM, Asuman DUATEPE PAKSU, Sibel DURU

[A novel approach for calculating the item discrimination for Likert type of scales](#)

Page: 772-786 PDF



Ümit ÇELEN, Eren Can AYBEK

[A study of reliability, validity and development of the teacher expectation scale](#)

Page: 787-807 PDF

Hasan İĞDE, Levent YAKAR

Investigation of education value perception scale's psychometric properties according to CTT and IRT

Harun Dilek ^{1,*}, Ufuk Akbaş ²

¹Ministry of National Education, İstanbul, Türkiye

²Hasan Kalyoncu University, Faculty of Education, Department of Measurement and Evaluation in Education, Gaziantep, Türkiye

ARTICLE HISTORY

Received: Aug. 24, 2021

Revised: Mar. 11, 2022

Accepted: July 22, 2022

Keywords:

Education value,
Bronfenbrenner's
ecological theory,
Measurement invariance,
Classical test theory,
Item Response Theory.

Abstract: The purpose of this study is to develop Education Value Perception Scale (EVPS) based on Bronfenbrenner's Ecological Theory and to investigate its psychometric properties according to Classical Test Theory (CTT) and Item Response Theory (IRT). The data were collected from 2872 secondary school students by stratified purposeful sampling method. Measurement invariance of EVPS was tested by multigroup confirmatory factor analysis based on gender, and scalar invariance was observed to have been provided. The estimations based on IRT were conducted based on Graded Response Model. While high positive correlations were found between the item discriminations estimated according to different test theories, high negative correlations were identified between item means. McDonald's Omega was calculated to be .79 according to CTT from reliability estimation methods, marginal reliability coefficient was determined to be .77 according to IRT. In the test-retest applications performed at 20-day intervals, the stability coefficient was found to be .81.

1. INTRODUCTION

There are many psychological factors affecting students learning, and these factors influence education and training process (Özbay, 2018). A student should be motivated in order to become successful during education and training process, yet this motivation is not sufficient alone. External factors such as teacher feedback and assignments in line with skill level should also be appropriate (Kelecioğlu, 1992). As it can be understood, there are external factors affecting academic success. Prior studies show that academic success is affected by family attitudes, circle of friends, teachers, school dynamics, social environment (Arıcı, 2007; Sarier, 2016; Sezgin et al., 2016; Tuncer & Bahadır, 2017). Bronfenbrenner (1977) explains such factors affecting academic success and an individual's interaction with environment under Ecological Theory. People interact with the environment where they live during their lives actively and passively, and these environmental factors influence people's development process in active and passive manners. As a result of the interaction of an individual with his/her family, friends and social environment, changes occur in his/her perceptions. These environmental factors are discussed in the Ecological Theory developed by Bronfenbrenner (1977; 1979;

*CONTACT: Harun Dilek ✉ harundilk@gmail.com 📍 Ministry of National Education, İstanbul, Türkiye

1986; 1994). The Ecological Theory argues that there are five systems (microsystem, mesosystem, exosystem, macrosystem and chronosystem) surrounding an individual from close to far (Bronfenbrenner, 1979). While chronosystem was not mentioned in the first period when the theory was coined, its importance was emphasized by Bronfenbrenner (1994) in the following periods (Shelton, 2019). Although it was developed in a relatively old period, the Ecological Theory still draws interest and continues to develop by current studies (Aliyev et al., 2021; Santrock, 2011; Shelton, 2019). The systems explained by the Ecological Theory are as follows:

Microsystem: As the first system of the Ecological Theory, the microsystem refers to anybody in the close circle of an individual and with whom s/he has direct contact. Family, friends, teachers can be given as examples of microsystem elements. As a result of the relationships an individual has with these people, his/her subjective perceptions develop, and these perceptions affect an individual's development (Bronfenbrenner, 1979; Shaffer, 2009).

Mesosystem: Mesosystem, which is made up of microsystems, focuses on the relationships among the elements in the microsystem. This refers to the effect of the interactional relationship of at least two elements in the microsystem on an individual's development (Bronfenbrenner, 1979). This system may be exemplified by the relationship between a family and teachers, the relationship between teachers themselves, the relationship of a family with an individual's friends.

Exosystem: The elements of this system, with whom an individual does not have active relationships, affect an individual and his/her immediate circle. The results of events in this system influence an individual's perceptions indirectly (Santrock, 2011). Working environment of parents and decisions taken by the school administration can be given as examples of this system.

Macrosystem: In this system, the effect of countries, societies, ideologies, belief systems on an individual's development is examined (Bronfenbrenner, 1979). This system covers important decision-makers about the lives of individuals such as those who manage a government or education policies, and the elements guiding large masses such as media organs. The decisions taken by the elements covered in this system result in what individuals will learn, what kind of a life they will live (Shaffer, 2009).

Chronosystem: Indicating the effect of time change on the development, the chronosystem argues that when the properties of an individual change over time, the environment she lives also changes (Bronfenbrenner, 1994). This change occurs in two ways: expected changes and unexpected changes (Bronfenbrenner, 1986). While the transition between levels in school life, entry to business life, marriage, retirement can be given as examples of expected changes, death of a relative, immigration, divorce, diseases can be presented as examples of unexpected changes.

There are environmental elements affecting a student's education in the aforementioned Ecological Theory. These elements cause a student to have a perception regarding his/her education. The perception of education value is the perception of an individual regarding the factors affecting his/her education and the relations between them (Aliyev et al., 2021). Investigating peer bullying, its effects, causes and consequences under the Ecological Theory, Doğan (2010) emphasized the necessity to develop programs that would raise awareness and prevent peer bullying across the country. Hong and Eamon (2012) examined the perceptions of students aged between 10-15 about their insecurity of schools according to the microsystem, mesosystem and exosystem level of the Ecological Theory. The study concluded that the perceptions about the insecurity of schools differed in terms of sociodemographic aspects.

Espelage (2014) discussed aggression, bullying and victimization of young people in accordance with microsystem and mesosystem. He stated the necessity to conduct informative

studies for students, school staff, teachers and adults, emphasized the importance of cooperation and studies should be carried out in other systems of the Ecological Theory. In their study, Özenç and Doğan (2014) developed “Functional Literacy Experience Scale” based upon Ecological Theory for 5th grade students with 3 factors and 32 items. Gençtanırım (2015) discussed the prevention of adolescent suicides in line with the Ecological Theory. He stated the necessity to carry out prevention studies covering each system of the Ecological Theory in order to prevent adolescent suicides. Aslantürk (2018) developed the "School Safety Scale" consisting of 61 items with 12 factors based on Ecological Theory. Zorbaz and Bilge (2019) indicated that the approaches based on the Ecological Theory could be effective on the psychosocial skills of delinquent children. Kopan (2019) argued that nutritional habits of 10th grade students are associated with all systems of the Ecological Theory, and Ecological Theory based studies would urge students to a healthier nutrition. Aliyev et al. (2021) urges that perception of education value is among the predictors of academic resilience.

As seen above, the Ecological Theory is used in many fields, especially in educational psychology. In addition, parents consider education of their children valuable before they start school and make plans for the future (Mapp, 2002). For this reason, it was considered important to measure how children perceive the value given to their education. The reason for developing the measurement instrument for secondary school level is the fact that the students at this stage are between 10-14 age range and are in a developmental threshold. Secondary school students in adolescence are in the period of cognitive and psychosocial development and change (Arı, 2008). Determination of education perceptions of children in this period is of importance for taking required measures and fulfilling responses.

Dated back to the 20th century, Classical Test Theory (CTT) has been used by many researchers and is still used in ability tests, cognitive tests, personality measurements, and psychological measurements. In CTT, allowing to achieve true score based on the observed score by focusing on the whole test, the less the amount of error in the measurement, the closer the true score is. True score is formulated as follows in CTT: $T = X + e$. In this formula, “T” refers to true score, “X” indicates observed score and “e” shows the amount of error in the measurement (Crocker & Algina, 2008). In CTT, individual, test and item parameters depend on the group (Hambleton & Jones, 1993), and this is a basic limitation of the theory (Fan, 1998). This limitation is overcome by estimations based on Item Response Theory (IRT) (Ostini & Nering, 2006).

In IRT, parameter estimations are carried out based on the responses given to each item instead of whole test (Baker, 2001). In IRT, there are dichotomous and polytomous models according to the way the items are answered. Likert-type scales are classified under polytomous models. In this study, the Graded Response Model (GRM), which is a polytomous IRT model based on the 2-parameter logistic model, in which the responses are categorical and ordered, and the probability of responding to the categories above the category to which the individual reacts is estimated, was used (Embretson & Reise, 2000). The parameters estimated based on different theories were compared.

There are studies conducted based on different theories. Sarı and Karaman (2018) has examined the General Mattering Scale in terms of both CTT and IRT. Yaşar and Aybek (2019) have developed a Resilience Scale according to IRT. Arıcak et al. (2020) have attempted to validate the Cyberbullying Sensibility Scale, which has been developed for high school students, for university students based on IRT. In recent years, it is seen that IRT-based scale studies have been carried out.

2. METHOD

2.1. Study Group

When determining the participant number, it was considered that at least 200 participants should

be included in CTT based studies (Comrey & Lee, 1992), and at least 600 participants should be involved in IRT based studies (De Ayala, 2009). Though the study was planned on two applications, after the first application, schools were closed in Turkey as a result of the global COVID-19 pandemic. Therefore, the planned second application could not be actualized. However, since the data was collected from a large student group with the first application, the group was divided into two as it is accepted in the literature and the study is completed even though the study was initially planned on two groups. In this study, more than 300 students from each grade level were included. 2872 students were reached by the stratified purposeful sampling method, which aims to reveal the characteristics of a group and describe a group (Büyüköztürk et al., 2018). Next, the data set was divided into two groups randomly. While the first group (sample 1) was utilized for the development of Education Value Perception Scale (EVPS), the second group (sample 2) was used for the investigation of the psychometric properties of EVPS. The distribution of the data used by grades and genders is provided in Table 1.

Table 1. *Distribution of students in the study group by grade and gender.*

Grade	Sample 1			Grade	Sample 2		
	Male	Female	Total		Male	Female	Total
5 th grade	160	185	345	5 th grade	170	171	341
6 th grade	202	169	371	6 th grade	191	198	389
7 th grade	161	154	315	7 th grade	156	158	314
8 th grade	186	210	396	8 th grade	194	207	401
Total	709	718	1427	Total	711	734	1445

2.2. Data Collection Instruments

EVPS was developed under this study (see Appendix). The literature was reviewed before preparing items form (Aliyev et al., 2021; Bronfenbrenner, 1986; Bronfenbrenner & Ceci, 1994; Darling, 2007; Leonard, 2011; Tudge et al., 2009; Onwuegbuzie et al., 2013). Later on, a total of 24 secondary school students, studying in different grades, were asked to answer an open-ended question to measure the extent that their surrounding attaches importance to their own education. The answers of these students were examined, and it was seen that their answers progressed towards the macrosystem, while no expression was determined in the chronosystem stage. For example, one of the students expressed as follows: “Textbooks are given, there are smart boards in the classroom, the Ministry of National Education publishes sample questions and I examine these questions. All these are done so that I can get a good education”. As a result of these studies, an item pool of 29 items was achieved. Receiving a high score from the scale indicates that the perception of education value is high. There is no reversed item in the scale.

Next, in order to receive expert opinion, these items were sent to a total of nine experts (7 PhD’s and 2 PhD candidates), three of which were expert in assessment and evaluation in education, two of them in educational psychology, four of them in psychological counseling and guidance, all of them had studies on education value, scale development, developmental psychology. The expert opinions were analyzed by the Lawshe technique that is a method used to identify content validity of the items in items form (Yurdugül, 2005). As a result of the analysis, one item was removed by considering the criteria suggested by Ayre and Scally (2014), and the content validity index for 28 items was calculated as .93. Four items recommended by experts were added to the scale, and the items form of 32 items was provided. Before conducting a pilot study with this items form, a pre-pilot study was administered with 17 secondary school students studying at different grade levels in order to determine the clearness of the items at the student level. In this study, students were requested to explain why they marked the category, they

chose, in each item. These explanations were examined, and 13 items, perceived and interpreted differently by students, were decided to be removed. A pilot study was carried out with the remaining 19 items, which had four response categories as follows: “Not Proper For Me”, “A Little Proper For Me”, “Significantly Proper For Me”, “Completely Proper For Me”.

In order to specify criterion-related validation of the developed measurement instrument, the “family support sub-scale” of the Social Relationship Factors Scale developed by Turner et al. (1983) and adapted into Turkish by Duyan et al. (2013) was used. The answers that may be provided for items consisted of five categories ranging between “Never Applicable for Me” and “Completely Applicable for Me”. A high score from the scale indicates high family support. The reason for using the family support sub-scale of this scale as a criterion is that the family has an important place at the microsystem level of the Ecological Theory.

In addition to family support, examinations regarding scale validity were carried out by Vallerand et al., (1992) Academic Motivation Scale (AMS) developed by Vallerand et al., (1989) and adapted into English. The scale was adapted into Turkish by Yurt and Bozer (2015). Consisting of seven graded items (1- non-compliant, 7- fully-compliant), the AMS involves seven factors, each of which has four items, as follows: Intrinsic Motivation to Know (IMTK), Intrinsic Motivation Toward Accomplishments (IMTS), Intrinsic Motivation to Experience Stimulation (IMTES), Extrinsic Motivation-Introjected Regulation (EMIR), Extrinsic Motivation- External Regulation (EMER), Identified Regulation (IR) and Amotivation (A). Receiving a high score in each sub-scale refers that the structure in that sub-scale has a high degree. The reason for using AMS as a criterion validity is that a positive relationship was determined between the perception of education value and academic motivation in the study carried out by Aliyev et al., (2021).

2.3. Data Collection

The items form was administered in the spring term of the 2019 – 2020 school year. EVPS and the family scale were administered to 84 secondary students in the classroom environment. 51 of the students were male and 33 of them were female. 33 of them were in the 5th grade, 20 in the 6th grade, 15 in the 7th grade and 16 in the 8th grade.

EVPS and AMS were administered to 96 secondary schools in the classroom environment. 51 of the students were male and 44 of them were female. 27 of them were in the 5th grade, 30 in the 6th grade, 21 in the 7th grade and 18 in the 8th grade.

In order to determine the stability of the developed EVPS, a test-retest administration was carried out. EVPS was administered to a total of 22 students (8th grade), 12 males and 10 females, with an interval of 20 days.

2.4. Data Analysis

It was seen that there was a 3.02% missing data. The distribution of the missing data per substance, category and gender, their probability of occurring together were examined and no systematic pattern has been identified. The missing with the Missing Completely at Random mechanism were removed from the data set with the listwise method. In the remaining data, it was examined whether there were multivariate outliers. The results before and after the deduction of outliers were examined; it was observed that they were similar and that it is appropriate for them to be included in the data set in order to prevent the sample from getting smaller. Thus, it was decided to keep some of the determined outliers in the data set by assuming that they could be considered reasonable for samples of this size (Akbaş & Koğar, 2020). In order to specify the construct validity of EVPS, exploratory factor analysis (EFA) with R program and confirmatory factor analysis (CFA) were performed according to gender and grade level. In EFA, principal axis method, which allows factor extraction by analyzing the common variance, was used (Tabachnick & Fidell, 2013). The EFA and CFA estimates have been

calculated with polychoric correlation. The Unweighted Least Squares maximum likelihood method was utilized for the CFA estimates. CFA estimates have been conducted by using The Unweighted Least Squares method (Katsikatsou et al., 2012; Koğar & Yılmaz Koğar, 2015) with lavaan package (Rosseel, 2012). Prior to the analysis, the normal distribution hypothesis for CTT and one-dimensionality, local independence hypotheses for the IRT were tested. Subsequently, the model data compliance for the IRT estimations were designated by pairwise comparison IRT analyzes were performed with the mirt package (Chalmers, 2012) of the R (R Core Team, 2020) program. While the interactions between hypotheses were being examined, when consistency and normality were ensued and the Pearson Moments correlation coefficient is not present, the Spearman's rank correlation coefficient was utilized

3. RESULT

Kaiser-Meyer-Olkin (KMO) value and Bartlett sphericity test analysis, which are preconditions for performing EFA on the collected data, were administered firstly according to whole data set, then only female and male students and finally grade levels. The EFA analysis were conducted by utilizing polychoric correlation matrix. The KMO value is .83 or higher for all data sets and the Bartlett's tests for sphericity are meaningful.

For various categories and sexes, it was observed that the determinant is positive, the VIF values are lower than 2, the tolerance values are higher than .5 and no multicollinearity problem is present. The correlation between the substances were examined with dispersion diagrams and it was observed that there is a linear correlation (Tabachnik & Fidell, 2013).

According to analysis results, it was identified that the eigenvalue of the 1st factor was 3 times the eigenvalue of the 2nd factor in all groups, when the scree plot was examined in all groups (Erkuş, 2016), there was a high decrease after the 1st factor, and all items had a high factor loading in the 1st factor (Büyüköztürk, 2018). Parallel analysis results also supported one factor result (Watkins, 2000). Based on these data, the scale was determined to have one dimension. As a result of the analysis, 11 items were removed due to cross loading and low factor loading, and it was seen that there were eight items with a factor loading above .32 (Tabachnick & Fidell, 2013). The factor loadings and corrected item-total correlations of the remaining items in the scale are shown in Table 2.

Table 2. Factor loadings (λ) of items by groups and corrected item total correlations (r_{jx}).

	Total		Male		Female		8th grade		7th grade		6th grade		5th grade	
	λ	r_{jx}	λ	r_{jx}	λ	r_{jx}	λ	r_{jx}	λ	r_{jx}	λ	r_{jx}	λ	r_{jx}
I8	.76	.58	.75	.56	.76	.60	.78	.59	.82	.64	.69	.49	.70	.50
I6	.73	.55	.72	.53	.75	.57	.69	.49	.77	.60	.72	.52	.75	.57
I5	.73	.51	.69	.47	.76	.55	.73	.51	.67	.47	.70	.45	.78	.53
I9	.67	.49	.66	.46	.68	.51	.66	.47	.65	.47	.61	.41	.70	.51
I14	.65	.47	.62	.43	.68	.51	.61	.43	.67	.50	.66	.45	.63	.43
I12	.63	.46	.57	.44	.68	.49	.57	.43	.64	.46	.51	.40	.65	.41
I15	.66	.45	.63	.39	.68	.51	.62	.40	.64	.48	.61	.33	.63	.41
I2	.61	.42	.58	.39	.63	.44	.60	.40	.57	.39	.59	.38	.65	.44
	$\omega = .82$		$\omega = .82$		$\omega = .83$		$\omega = .80$		$\omega = .83$		$\omega = .80$		$\omega = .84$	
	EV ¹ = %46.3		EV= %43.0		EV= %49.6		EV= %43.4		EV= %46.4		EV= %41.0		EV= %47.4	

¹EV, extracted variance

When examining [Table 2](#), the factor loadings were between .57 and .82 for all groups. Item discrimination index varies between .33 and .64. Item discrimination index should be .30 and above (Crocker & Algina, 2008). The extracted variance ratio are between 41% and 49%. For a one factor scale, it is sufficient that the extracted variance rate is 30% or more (Büyüköztürk, 2018). Based on these data, it was observed that extracted variance ratio, factor loadings and item discrimination indexes were acceptable. The fit of the one factor model identified by EFA was tested CFA by using the data obtained from the second sample. For various categories and sexes, it is observed that the determinant is positive, the VIF values are lower than 2, the tolerance values are greater than .5 and there is no multicollinearity problem. The correlation between the substances were examined with dispersion diagram and that there is a linear correlation (Tabachnick & Fidell, 2013). The fit indices obtained as a result of CFA are provided in [Table 3](#).

Table 3. *Confirmatory Factor Analysis results.*

		total	male	female	8 th grade	7 th grade	6 th grade	5 th grade
N		1445	711	734	401	314	389	341
χ^2		115.18	72.52	47.68	67.68	45.60	25.32	17.54
df		20	20	20	20	20	20	20
CFI	.90 ≤ good fit ≤ .95 ≤ perfect fit	.98	.97	.99	.96	.97	.99	1.00
TLI	.90 ≤ good fit ≤ .95 ≤ perfect fit	.97	.96	.98	.94	.96	.99	1.00
SRMR	perfect fit ≤ .05 ≤ good fit ≤ .08	.04	.05	.04	.06	.06	.04	.04
RMSEA	perfect fit ≤ .05 ≤ good fit ≤ .08 ≤ poor fit ≤ .10	.06	.06	.07	.07	.06	.03	.00

When the values shown in [Table 3](#) are compared with the limit values recommended in the literature (Browne & Cudeck, 1993; Moosburger & Müller, 2003; Schumacker & Lomax, 2004; Kline, 2016), it is seen that the CFI, TLI, SRMR and RMSEA values are in ranges of the limit values recommended literature.

The Composite Reliability (CR) and Average Variance Extracted (AVE) values calculated to determine the convergent validity of the structure confirmed by CFA analysis were given in [Table 4](#).

Table 4. *Composite Reliability and Average Variance Extracted values.*

	CR	AVE
Total	.91	.56
male	.93	.62
female	.89	.51
8 th grade	.91	.60
7 th grade	.91	.63
6 th grade	.90	.55
5 th grade	.88	.45

When [Table 4](#) is examined, it is seen that the CR values are greater than .70 and the AVE values are greater than .50, except for the 6th grades. These values show that convergent validity has been achieved (Hair et al., 2014).

3.1. Measurement Invariance

It was examined by multigroup confirmatory factor analysis (MCFA) whether the developed EVPS resulted in the same structure according to gender. The measurement invariance has four stages: configural invariance, metric invariance, scalar invariance and strict invariance (Meredith, 1993; Steenkamp & Baumgartner, 1998; Kline, 2016). If the CFI index obtained in different stages of MCFA is lower than $|.01|$, this shows that invariance is provided between stages (Cheung & Rensvold, 2002; Ho, 2006). Another value accepted as a criterion for measurement invariance is the SRMR value. The SRMR values being less than $|.01|$ at each stage shows that measurement invariance has been achieved (Chen, 2007). MCFA results by gender are given in Table 5.

Table 5. Multigroup Confirmatory Factor Analysis results by gender.

Stages	χ^2	df	CFI	GFI	RMSEA	SRMR	Δ SRMR	Δ CFI
Configural Invariance	198.66	40	.921	.965	.052	.041	-	-
Metric Invariance	207.69	47	.920	.963	.049	.045	.004	.002
Scalar Invariance	207.82	48	.920	.963	.048	.047	.006	.002
Strict Invariance	221.82	56	.917	.960	.045	.050	.009	.005

It is shown in Table 5 that strict invariance was provided according to gender, measurements referred to the same structure for female and male students.

As a result of the CFA, the measurement invariance of the EVPS which was confirmed to be unidimensional, were tested in accordance with the grade levels. According to the invariance stages, it was seen that only the configural invariance was ensured. It was seen that Δ CFI and Δ SRMR levels are greater than .01 in the other stages.

3.2. Criterion-Related Validity and Stability

In order to examine the criterion-related validation of EVPS, the correlation calculated using the family support sub-scale of the Social Support Scale, was found to be highly positive relationship between scales ($r = .43, p < .01$). The Pearson Product-Moment Correlation Coefficients, calculated by the data obtained with the AMS used in the other examination for the criterion-related validity of the EVPS are given in Table 6.

Table 6. Correlations between Education Value Perception Scale and Academic Motivation Scale.

	\bar{x}	df	IMTK	IMTS	IMTES	EMIR	EMER	IR	A	EVPS
IMTK	23.30	4.74		.60	.62	.45	.45	.60	-.29	.49
IMTS	21.97	5.09	.60		.67	.61	.54	.62	-.28	.46
IMTES	20.48	5.21	.62	.67		.51	.51	.62	-.27	.38
EMIR	19.46	6.06	.45	.61	.51		.58	.36	.00	.21
EMER	21.52	5.19	.45	.54	.51	.58		.49	-.81	.32
IR	24.67	4.57	.60	.62	.62	.36	.49		-.39	.42
A	6.43	4.12	-.29	-.28	-.27	.00	-.81	-.39		-.41
EVPS	24.89	5.02	.49	.46	.38	.21	.32	.42	-.41	

IMTK: Intrinsic Motivation to Know

IMTS: Intrinsic Motivation Toward Accomplishments

IMTES: Intrinsic Motivation to Experience Stimulation

EMIR: Extrinsic Motivation-Introjected Regulation

EMER: Extrinsic Motivation- External Regulation

IR: Identified Regulation

A: Amotivation

As it is seen in [Table 6](#), there were positive correlations among EVPS with IMTK, IMTS, IMTES, EMIR, EMER, IR scores as expected, while there were negative correlations between EVPS and A scores as expected.

A test-retest administration was carried out for examining the stability of the data obtained by EVPS. It was observed that the correlation coefficient calculated by the data obtained from 22 students with an interval of 20 days was $\rho = .81$ ($p < .05$).

3.3. Model-Data Fit in IRT

For analyses regarding the IRT, in order to determine with which categorical model that the scale is in compliance, the model-data fit analysis was conducted with pairwise comparisons. The comparisons made between Graded Response Model (GRM) and Partial Credit Model (PCM) and Generalized Partial Credit Model are presented in [Table 7](#).

Table 7. Results of the Item Response Theory model comparisons.

	AIC	AICc	SABIC	HQ	BIC	LL	χ^2	df	p
PCM	27228.44	27229.36	27280.92	27277.67	27360.34	-13589.22			
GRM	27098.97	27100.43	27166.14	27161.98	27267.79	-13517.48	72.603	0	0

Upon examining [Table 7](#), it is observed that the AIC, BIC and logLikelihood values are lower in GRM and that the p value of χ statistic is meaningful signifies that the GRM model is more appropriate.

Table 8. Results of the Item Response Theory model comparisons.

	AIC	AICc	SABIC	HQ	BIC	LL	χ^2	df	p
GPCM	27171.57	27173.07	27238.74	27234.58	27340.40	-13553.78			
GRM	27098.97	27100.43	27166.14	27161.98	27267.79	-13517.48	72.603	0	0

Upon examining [Table 8](#), it is observed that the AIC, BIC and logLikelihood values are lower in GRM and that the p value of χ statistic is meaningful signifies that the GRM model is more appropriate. As a result of pairwise comparisons, it was observed that the model best suitable for the data is the GRM.

In the fit of the data obtained by sample 2 in the model, firstly, the fit of each item in GRM was examined, which was followed by the examination of the fit of whole scale in GRM. In order to ensure model-data fit according to the items, RMSEA value should be less than .08, p significance value should be greater than .05 and the χ^2/df value should be below 3. The results obtained for GRM are provided in [Table 9](#).

Table 9. The fit of items in Graded Response Model.

Statistics	I1	I2	I3	I4	I5	I6	I7	I8
χ^2	47.91	52.14	51.32	57.36	57.18	52.88	59.24	51.80
df	49	46	43	42	45	47	46	48
RMSEA	.00	.01	.01	.02	.01	.01	.01	.01
p	.52	.25	.18	.06	.11	.26	.10	.33

When [Table 9](#) is examined, it is seen that all items fitted in the model ($p > .05$), RMSEA values were lower than .08 and χ^2/df ratio was lower than 3. When examining the model-fit of the whole scale, it is observed that CFI value was .95, NNFI value was .93, RMSEA value was .08, SRMSR value was .05. These findings show that model fit was provided.

The invariance of item parameters was tested by randomly dividing 1445 students into two groups to determine parameter invariance, and the invariance of ability parameters was tested by dividing 8 items into 2 groups. The correlation between the invariance of the b_1, b_2, b_3 item parameters which estimated according to the two groups was found to be positive excellent ($\rho = 1.00, p < .05$), and the correlation between a parameters was found to be positive high ($\rho = .92, p < .01$). The relationship between abilities (θ) was found to be positive moderate level ($r = .60, p < .01$).

The item information functions obtained according to GRM are shown in Figure 1, and the test information function for the whole test is demonstrated in Figure 2. In Figure 2, it is seen that most information was provided between ability -2 and +1.5 in the scale.

Figure 1. Item information functions.

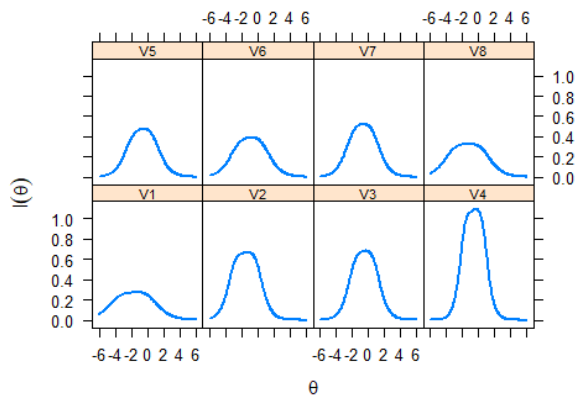
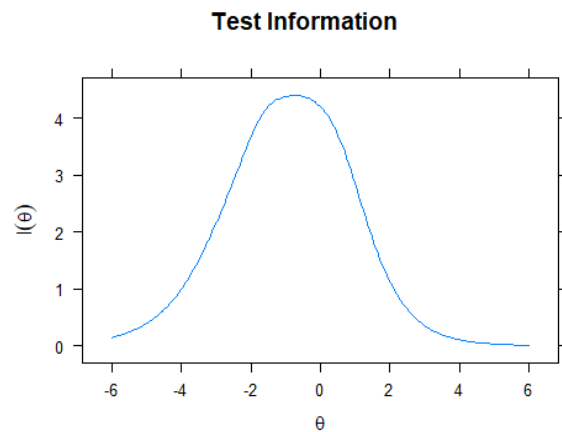


Figure 2. Test information functions.



3.3. Examination of Measurements According to Different Test Theories

The relationship between the item discriminations estimated by the CTT and IRT of the EVPS was examined. Corrected item-total correlation coefficient according to CTT, and a parameter according to IRT were estimated. The findings related to estimation are given in Table 10.

Table 10. Item discrimination indices.

Items	CTT	IRT (a parameter)	SEa
1. Ailem, evde ders çalışmam için uygun bir ortam hazırlar. (My family prepares an appropriate environment for me to study at home.*)	.36	.94	.08
2. Öğretmenlerim, eğitimimle ilgili beni yönlendirir. (My teachers guide me about my education.*)	.45	1.49	.10
3. Ailem ve öğretmenlerim eğitimimi iyileştirmek için iş birliği yapar. (My family and my teachers collaborate to better my education.*)	.50	1.49	.10
4. Derslerime giren öğretmenlerim, eğitimim için iş birliği yaparlar. (My teachers in my classroom collaborate for my education.*)	.56	1.91	.12
5. Okul idaresi, ailemi eğitim faaliyetleri hakkında bilgilendirir. (School administration informs my family about educational activities.*)	.46	1.24	.08
6. Okulumda yapılan sosyal etkinlikler, eğitim sürecime katkıda bulunur. (Social activities in my school contribute my education process.*)	.43	1.12	.08
7. Ülkemde, iyi bir eğitim almam için fırsatlar sunulmaktadır. (Opportunities are provided to me to get a good education in my country.*)	.48	1.30	.09
8. Bu eğitim sisteminde başarılı olabilirim. (I can be successful at this education system.*)	.40	1.03	.08

* Unvalidated English translation

When examining [Table 10](#), item discriminations range between .36 (2nd item) and .56 (8th item) according to CTT. According to the CTT, it is sufficient for an item discrimination of .30 and above (Büyüköztürk, 218). It is observed that a parameter changed between .94 (2nd item) and 1.91 (8th item) according to IRT. According to IRT, the distinctiveness of a parameter was specified to be very low between .01-.34, moderate between .35-.64, high between 1.35-1.69 and very high in 1.70 and above (Baker, 2001). The correlation between discriminations which tested according to CTT and IRT was found to be a highly positive significant ($\rho = .90, p < .05$).

The relationship between the item means estimated according to the CTT and IRT was examined. Item means were determined by taking the average of the responses given by participants to categories according to the CTT. b parameter was estimated as one less than the number of categories according to the IRT. Three b-parameters were estimated, as the EVPS had four categories. The findings for the item means are shown in [Table 11](#).

Table 11. *Item means rates.*

Item	CTT	<i>sd</i>	b1	SE	b2	SE	b3	SE
1	3.23	.03	-3.32	.25	-1.42	.12	-.16	.07
2	3.38	.02	-2.67	.16	-1.45	.09	-.35	.05
3	2.78	.03	-1.47	.08	-.40	.05	.57	.06
4	2.89	.03	-1.51	.08	-.50	.05	.39	.05
6	2.84	.03	-1.64	.10	-.50	.06	.46	.06
6	3.00	.03	-2.19	.14	-.89	.08	.32	.06
7	2.89	.03	-1.68	.10	-.62	.06	.41	.06
8	3.15	.03	-2.83	.20	-1.32	.10	.15	.06

When [Table 11](#) is examined, it is seen that participants mostly reacted to high categories. The relationship between item means which estimated according to the both theories, was found to be negative high ($\rho = -.94, p < .05$). In CTT, as the difficulty increases, the item becomes easier. In IRT, on the other hand, it is the opposite. Therefore, the item means showed great similarity according to both theories.

The relationship between students' total score according to CTT and their perceptions of education value estimated by the Expected a Posteriori (EAP) method according to IRT was examined. EAP method enables to estimate θ levels of students, who had full score or the lowest score from the scale (Embretson & Reise, 2000). The relationship between students' perception of education value which estimated according to the both theories was found to be positive highly significant ($r = .98, p < .01$).

The McDonald's Omega level of internal consistency estimated according to the CTT of the scale was calculated as .79, and the marginal reliability coefficient estimated according to the IRT was calculated to be .77. Accordingly, it is seen that the reliability coefficients are considered satisfactory for measurement instruments used in education and psychology (Nunnally & Bernstein, 1994).

4. DISCUSSION and CONCLUSION

The EVPS that has been developed based on the Ecological Theory is a scale of eight items and represents four systems of the theory. Among the substances remaining in the scale, the 1st and the 2nd are related to the factors in the microsystem stage of the Ecological Hypothesis, the 3rd, the 4th and the 5th are related to the factors in the mesosystem stage, the 6th substance is related to the factors in exosystem stage. Therefore, the total score received from the scale is

able to unidimensionally present the perception of the student regarding education. Though they are not at the same educational stage, the scale developed by Aliyev et al. (2021) with the purpose of assessing the educational perspective of university students, is also unidimensional. Thus, the argument claiming that the educational value perspective in the Turkish culture is a unidimensional structure is further supported. Even if the items, which represent chronosystem level were written in the items form, they were not included in the final scale as they were identified to be inadequate as a result of EFA, and students were determined not to have perceived chronosystem level after examining their responses they gave to open ended questions. It was observed that strict invariance was provided in the measurement invariance of the EVPS according to gender. In this case, the scores to be obtained from the scale may be compared between groups. The reason for differences between groups will arise due to perceptions of education value.

When examining the item discriminations of the EVPS, it is seen that they are highly distinctive according to both CTT and IRT. It has been observed that there are high positive correlations between the parameters estimated according to different theories. The findings show similarity with the previous studies in this respect (Ferhan, 2018; Karakılıç, 2009; Köse, 2015; Nartgün, 2002; Uysal, 2015; Yaşar, 2019).

It has been achieved that EVPS can be used in studies not only based on CTT but also IRT. In this regard, ability estimations independent from sample can be made, and standard error can be calculated according to different ability levels. It enables to choose the most proper model to make more accurate estimations. It also allows to have detailed information by focusing on the responses to items. As it is likely to make probability estimations about how individuals will response to any item, it can be benefited from advantages of determining ability levels of individuals more accurately (Crocker & Algina, 2008; Hambleton et al., 1991).

Secondary school students' perceptions of education value can be measured by taking advantage of qualitative methods to eliminate the missing chronosystem level in EVPS. Although attention was paid to reach a heterogeneous study group at the development of the EVPS, city of Gaziantep takes places at lower side of education level. EVPS can be used in studies that are focused on comparing groups which have different educational levels. Changing of students' perceptions of education value can be examined through longitudinal studies.

Acknowledgments

This paper was produced from the first author's master's thesis prepared under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Hasan Kalyoncu University/ Social Science Institution, 13-01-2020/02.

Authorship Contribution Statement

Harun Dilek: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing -original draft. **Ufuk Akbas:** Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, Writing -original draft, Supervision, and Validation.

Orcid

Harun Dilek  <https://orcid.org/0000-0001-5671-6858>

Ufuk Akbas  <https://orcid.org/0000-0002-6122-154X>

REFERENCES

- Akbaş, U., & Koğar, H. (2020). *Nicel araştırmalarda kayıp veriler ve uç değerler: çözüm önerileri ve SPSS uygulamaları [Missing data and outliers in quantitative research; solution suggestions and SPSS applications]*. Pegem Akademi.
- Aliyev, R., Akbaş, U., & Özbay, Y. (2021). Mediating role of internal factors in predicting academic resilience. *International Journal of School & Educational Psychology*, 3(29), 1-16. <https://doi.org/10.1080/21683603.2021.1904068>
- Arı, R. (2008). *Eğitim psikolojisi [Educational psychology]*. Nobel.
- Arıcak, O.T., Avcu, A., Topçu, F., & Tutlu, M.G. (2020). Use of item response theory to validate cyberbullying sensibility scale for university students. *International Journal of Assessment Tools in Educaiton*, 7(1), 18-29. <https://doi.org/10.21449/ijate.629584>
- Arıcı, İ. (2007). *The effective factors on the students in the religious culture and ethics course [Doctoral dissertation]*. Hacettepe University, Ankara. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp>
- Aslantürk, İ. (2018). *Ecological system theory development of a basic school safety measurement [Master's dissertation]* Ahi Evran Universtiy, Kırşehir. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp>
- Ayre, C., & Scally, A.J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, 47(1), 79-86. <https://doi.org/10.1177/0748175613513808>
- Baker, F.B. (2001). *The basic of item response theory*. ERIC.
- Başusta, N.B., & Gelbal, S. (2015). Gruplararası karşılaştırmada ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği [Testing measurement invariance in comparisons between groups: Pisa student survey sample]. *Hacettepe U. Journal of Education*, 30(4), 80-90.
- Bronfenbrenner, U. (1977). Toward experimental ecology of human development. *American Psychologist*, 32(7), 513-531. <https://doi.org/10.1037/0003-066X.32.7.513>
- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and desing*. Harvard Universty Press.
- Bronfenbrenner, U. (1986). Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, 22(6), 723-742. <https://doi.org/10.1037/0012-1649.22.6.723>
- Bronfenbrenner, U. (1994). Ecological models of human development. *International Encyclopedia of Education*, 3(2), 37-43.
- Bronfenbrenner, U., & Ceci, S.J. (2004). Nature-nurture reconceptualized in developmental perspective: A bioecological model. *Psychological Review*, 101(4), 568-586. <https://doi.org/10.1037/0033-295X.101.4.568>
- Browne, M.W., & Cudeck, R. (1993). *Alternative ways of assessing model fit*. K. A. Bolen and J. S. Long. (Ed), Testing Sructural Equation Models. SAGE Puplications.
- Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2018). *Bilimsel araştırma yöntemleri [Scientific research methods]*. Pegem Akademi.
- Büyüköztürk, Ş. (2018). *Sosyal bilimler için veri analiz el kitabı [Data analysis handbook for social sciences]*. Pegem Akademi.
- Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464-504. <https://doi.org/10.1080/10705510701301834>
- Cheung, G.W., & Rensvold, R.B. (2002). Evalutaing goodness-of-fit indexes for testing measurement invariance. *Structual Equation Modelling*, 9(2), 223-255. https://doi.org/10.1207/S15328007SEM0902_5

- Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis*. Lawrence Erlbaum Associates, Inc.
- 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>
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.
- Darling, N. (2007). Ecological systems theory: The person in the center of the circles. *Research in Human Development*, 4(3), 203-217.
- De Ayala, R.J. (2009). *The theory and practice of item response theory*. The Guilford Press.
- Doğan, A. (2010). Ekolojik sistemler kuramı çerçevesinde akran zorbalığı incelemesi [Ecological systems model as a framework for bullying]. *Turk J Child Adolesc Ment Health*, 17(3), 149-162.
- Duyan, V., Gelbal, S., & Var, E.Ç. (2013). Sosyal ilişki unsurları ölçeğinin Türkçeye uyarlama çalışması [The adaptation study of the provision of social relations scale to Turkish]. *Hacettepe University Journal of Education*, 44(44), 159-169.
- Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Ersbaum.
- Erkuş, A. (2016). *Psikolojide ölçme ve ölçek geliştirme-1 [Education and development in psychology-1]*. Pegem Akademi.
- Espelage, D.L. (2014). Ecological theory: Preventing youth bullying, aggression, and victimization. *Theory into Practice*, 53(4), 257-264. <https://doi.org/10.1080/00405841.2014.947216>
- Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357-381. <https://doi.org/10.1177/0013164498058003001>
- Ferhan, M. (2018). *The psychometric characteristics of PISA 2012 mathematics interest scale by classical test theory and item response theory*. [Master's dissertation]. Hasan Kalyoncu University, Gaziantep]. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp>
- Gençtanırım, D. (2015). Ergen intiharlarının önlenmesi: Ekolojik bakış açısı [Prevention Adolescents Suicide: Ecological Perspective]. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi (KEFAD)*, 16(1), 151-164.
- Hair, J.F., Hult, G.T., Ringle, C.M. & Sarstedt M. (2014). *A Primer on partial least squares structural equation modeling (PLS-SEM)*. SAGE Publications.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Ho, R. (2006). *Handbook of univariate and multivariate data analysis and interpretation with SPSS*. Taylor & Francis Group.
- Hong, J.S., & Eamon, M.K. (2012). Students' perceptions of unsafe schools: An ecological systems analysis. *Journal of Child and Family Studies*, 21(428-438).
- Karakılıç, M. (2009). *An investigation of attitude scale measuring students attitudes toward physical education through psychometric theories*. [Doctoral dissertation] Ankara University, Ankara. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp>
- Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Joreskog, K. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis*, 56(12), 4243- 4258. <https://doi.org/10.1016/j.csda.2012.04.010>
- Kelecioğlu, H. (1992). Güdülenme [Motivation]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 7, 175-181.

- Kline, R.B. (2016). *Principles and practice of structural equation modeling*. The Guilford Press.
- Koğar, H., & Yılmaz Koğar, E. (2015). Comparison of different estimation methods for categorical and ordinal data in confirmatory factor analysis. *Journal of Measurement and Evaluation in Education and Psychology*, 6(2), 351-364 <https://doi.org/10.21031/epod.94857>
- Kopan, D. (2019). *Evaluation of nutrition habits the second grade students in Seferihisar region with the ecological framework* [Master's dissertation]. Ege University, İzmir. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp>
- Köse, A. (2015). Aşamalı tepki modeli ve klasik test kuramı altında elde edilen test ve madde parametrelerinin karşılaştırılması [Comparison of test and item parameters under graded response model (IRT) and classical test theory]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 15(2), 184-197. <https://doi.org/10.17240/aibuefd.2015.15.2-5000161319>
- Leonard, J. (2011). Using Bronfenbrenner's ecological theory to understand community partnerships: A historical case study of one urban high school. *Urban Education*, 0042085911400337. <https://doi.org/10.1177/0042085911400337>
- Mapp, K.L. (2002). *Having their say: Parents describe how and why they are involved in their children's education*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April, 1-5.
- Meredith, W. (1993). Measurement invariance factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543.
- Nartgün, Z. (2002). *The investigation of item and scale properties of likert type scale and metric scale measuring the same attitude according to classisical test theory and item response theory* [Doctoral dissertation]. Hacettepe University, Ankara. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp>
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric Theory*. McGraw-Hill.
- Onwuegbuzie, A.J., Collins, K.M.T. & Frels, R.K. (2013). Foreword: Using Bronfenbrenner's Ecological Systems Theory to frame quantitative, qualitative, and mixed reserch. *International Journal of Multiple Research Approaches*, 7(1), 2-8. <https://doi.org/10.5172/mra.2013.7.1.2>
- Özbay, Y. (2018). *Eğitim psikolojisi* [Educational psychology]. Pegem Akademi.
- Özenç, E.G., & Doğan, M.C. (2014). Ekolojik kurama dayalı işlevsel okuryazarlık yaşantısı ölçeğinin geliştirilmesi ve geçerlik güvenirlik çalışması [The development of the functional literacy experience scale based upon ecological theory (FLESBUET) and validity-reliability Study]. *Educational Sciences: Theory & Practice*, 14(6), 2239-2258. <https://doi.org/10.12738/estp.2014.6.1791>
- R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36. <https://doi.org/10.18637/jss.v048.i02>
- Santrock, J.W. (2011). *Life-span development*. The McGraw-Hill.
- Sarı, H.İ., & Karaman, M.A. (2018). Gaining a better understanding of general mattering scale: An application of classical test theory and item response theory. *International Journal of Assessment Tools in Education*, 5(4), 668-681. <https://doi.org/10.21449/ijate.453337>
- Sarıer, Y. (2016). Türkiye'de öğrencilerin akademik başarısını etkileyen faktörler: Bir meta-analiz çalışması [The factors that affects students' academic achievement in Turkey: A meta-analysis study]. *Hacettepe University Journal of Education*, 31(3), 609-627. <https://doi.org/10.16986/HUJE.2016015868>
- Schumacker, R.E., & Lomax, R.G. (2004). *A beginner's guide to structural equation modelling*. Lawrence Erlbaum Associates.

- Sezgin, F., Koşar, D., & Koşar, S. (2016). Liselerde akademik başarısızlık: Nedenleri ve önlenmesine ilişkin öğretmen ve okul yöneticilerinin görüşleri [Teachers' and school administrators' views on reasons and prevention of academic failure in high schools]. *İnönü University Journal of the Faculty of Education*, 17(1), 95-111. <https://doi.org/10.17679/iuefd.17119535>
- Shaffer, D.R. (2009). *Social and personality development*. Wadsworth.
- Shelton, L.G. (2019). *The Bronfenbrenner primer: A guide to develeceology*. Routledge.
- Steenkamp, J., & Baumgartner, H. (1998). Assesing measurement invariance in cross national consumer research. *Journal of Consumer Research*, (25), 78-90. <https://doi.org/10.1086/209528>
- Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics*. Pearson.
- Tudge, J.R.H., Mokrova, İ., Hatfield, B.E., & Karnik, R.B. (2009). Uses and misuses of Bronfenbrenner's bioecological theory of human development. *Journal of Family Theory and Review*, 1, 198– 210. <https://doi.org/10.1111/j.1756-2589.2009.00026.x>
- Tuncer, M., & Bahadır, F. (2017). Ortaokul öğrenci görüşlerine göre başarısızlığın nedenleri [Reasons for underachievement by secondary schools' students opinions]. *Kahramanmaraş Sütcüimam Üniversitesi Eğitim Dergisi*, 1(1), 1-11.
- Turner, R.J., Frankel, B.G., & Levin, D.M. (1983). Social support: conceptualization, measurement, and implications for mental health. In J. R. Greeley (Ed.), *Research in community and mental health* (67-111). JAI Press.
- Uysal, M. (2015). *An investigation of psychometric properties of research self-efficacy scale according to classical test theory and item response theory* [Master's dissertation]. Gazi University, Ankara. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp>
- Vallerand, R.J., Blais, M.R., Brière, N.M., & Pelletier, L.G. (1989). Construction et validation de l'échelle de motivation en education (EME) [Construction and validation of the échelle de motivation en education (EME)]. *Canadian Journal of Behavioral Sciences*, 21, 323-349.
- Vallerand, R.J., Pelletier, L.G., Blais, M.R., Brière, N.M., Senécal C., & Vallières, E.F. (1992). The academic motivation scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement*, 52(4), 1003-1017.
- Watkins, M.W. (2000). *Monte Carlo PCA for parallel analysis* [Computer software]. Ed & Psych Associates.
- Yaşar, M., & Aybek, E.C. (2019). Üniversite öğrencileri için bir yılmazlık ölçeğinin geliştirilmesi: Madde tepki kuramı temelinde geçerlilik ve güvenilirlik çalışması [A resilience scale development for university students: Validity and reliability study based on item response theory]. *Elementary Education Online*, 18(4), 1687-1699. [10.17051/ilkonline.2019.635031](https://doi.org/10.17051/ilkonline.2019.635031)
- Yaşar, M. (2019). Development of a "Perceived Stress Scale" based on classical test theory and graded response model. *International Journal of Assessment Tools in Education*, 6(3), 522-538. <https://doi.org/10.21449/ijate.626053>
- Yurdugül, H. (2005). Ölçek geliştirme çalışmalarında kapsam geçerliği için kapsam geçerlik indekslerinin kullanılması [Using content validity indexes for content validity in scale development studies]. XIV. Ulusal Eğitim Bilimleri Kongresi, Pamukkale Üniversitesi Eğitim Fakültesi, 28-30 Eylül, Denizli.
- Yurt, E., & Bozer, E.N. (2015). Akademik Motivasyon Ölçeğinin Türkçeye Uyarlanması [The adaptation of the academic motivation scale for turkish context]. *Gaziantep University Journal of Social Sciences*, 14(3), 669-685. <https://doi.org/10.21547/jss.256759>
- Zorbaz, O., & Bilge, F. (2019). Suça sürüklenen çocukların değerlendirmesinde ekolojik sistem yaklaşımının kullanımı: Olgu sunumu [The use of the ecological system approach in assessment of juvenile delinquency: A case study]. *Sosyal Politika Çalışmaları Dergisi*, 19(44), 793-813. <https://doi.org/10.21560/spcd.v19i49119.506405>

APPENDIX

Turkish Version of The Education Value Perception Scale

Eđitim Deęeri Algısı Ölçeęi	Bana Hiç Uygun Deęil	Bana Biraz Uygun	Bana Büyük Ölçüde Uygun	Bana Tamamen Uygun
1. Ailem, evde ders çalıřmam için uygun bir ortam hazırlar.				
2. Öğretmenlerim, eğitimimle ilgili beni yönlendirir.				
3. Ailem ve öğretmenlerim eğitimimi iyileřtirmek için iş birlięi yapar.				
4. Derslerime giren öğretmenlerim, eğitimim için iş birlięi yaparlar.				
5. Okul idaresi, ailemi eğitim faaliyetleri hakkında bilgilendirir.				
6. Okulumda yapılan sosyal etkinlikler, eğitim sürecime katkıda bulunur.				
7. Ülkemde, iyi bir eğitim almam için fırsatlar sunulmaktadır.				
8. Bu eğitim sisteminde başarılı olabilirim.				

The role of individual differences on epistemic curiosity (EC) and self-regulated learning (SRL) during e-learning: the Turkish context

Ergun Akgun ^{1,*}, Enisa Mede ², Seda Sarac ³

¹Bahçeşehir University, Faculty of Educational Sciences, Department of Preschool Education, İstanbul, Türkiye

²Bahçeşehir University, Faculty of Educational Sciences, Department of English Language Education, İstanbul, Türkiye

³Bahçeşehir University, Faculty of Educational Sciences, Department of Preschool Education, İstanbul, Türkiye

ARTICLE HISTORY

Received: Mar. 31, 2021

Revised: June 28, 2022

Accepted: July 22, 2022

Keywords:

E-learning,
Epistemic curiosity,
Self-regulated learning,
Data mining,
Association rule.

Abstract: This study aims to examine the relations and associations between gender, epistemic curiosity (EC), self-regulated learning (SRL), and attitudes toward e-learning in higher education students. The participants were 2438 (862 males, 1576 females) undergraduate students enrolled in a Turkish university. The regression analysis findings showed that although the effect size was low, attitudes towards e-learning can be predicted significantly by gender, EC, and SRL. Datasets are further analyzed using data mining. The findings of the association rule mining revealed that gender plays an influential role. Several association rules among EC, SRL, and attitudes towards e-learning were detected for female students. The results provide recommendations about using data mining as a statistical method in educational and psychological research.

1. INTRODUCTION

With the increasing prevalence of Internet-based courses, attention has been placed on e-learning in educational institutions due to its numerous benefits including the absence of physical and temporal limits, the ease of accessing the material, and the cost-effectiveness (Altun et al., 2021; Howland & Moore, 2002). Specifically, the constructivist approach has had an impact on e-learning which resulted in the design of “constructivist e-learning environments” (CEEs) such as WebQuests, online courses, courses with simulations via computer management games and simulations (Martens et al., 2007, p.82). More specifically, the CEEs are based on constructivist principles which aim to provide challenging, authentic, and meaningful context. In this way, the learners can become intrinsically motivated during their learning process (Bastiaens & Martens, 2000).

As for the field of education, the e-learning environments accompanied by the widespread use and availability of computers and smartphones led to a shift in the process of teaching and learning (Erarslan & Topkaya, 2017). E-learning has started to offer platforms that are learner-centered, convenient for the learners’ own pace of learning, motivating, and available in various

*CONTACT: Ergün AKGÜN ✉ ea.ergunakgun@gmail.com 📍 Bahçeşehir University, Faculty of Educational Sciences, Department of Preschool Education, İstanbul, Türkiye

forms of sources to practice and interact with others through web-based tools (Mohammadi et al., 2011). Recent research indicated that the adoption of e-learning has been widely affected by student-related factors (Bhuasiri et al., 2012). Student attitudes toward e-learning have been crucial in various learning environments. As highlighted by Maio et al. (2018), strong attitudes can guide behavior and positive attitudes toward learning which contributes to the effective use of learning strategies. Therefore, possessing positive attitudes and behaviors regarding e-learning has been considered crucial for the acceptance, easiness, usability, and adoption of online learning (Aixia & Wang, 2011; Martins & Kellermanns, 2004; Selim, 2007).

The current COVID-19 pandemic led to a sudden shift to e-learning in higher education. This sudden transition to e-learning took place beyond the preferences of the students. To put it another way, with the emergency action plan put into effect by the universities, not only the students who deliberately and willingly preferred distance education, but all students had to take all their courses remotely. Under these conditions, it became more important to find out which variables affect students' development of positive or negative attitudes towards e-learning. As Gunnarsson (2001) and Suanpang (2007) revealed in their studies, there is a significant relationship between the students' attitudes and their learning achievement in an online course.

1.1. Gender and Attitudes Towards E-Learning

Gender is considered among the influential factors in students' attitudes toward e-learning. Attitudes toward learning in technology-enhanced environments, such as e-learning, are closely related to how much people are engaged with technology. According to Colley and Comber (2003), males approach computers like toys. They tend to figure out how it works and try to master using them. On the other hand, females regard computers as tools rather than a puzzle to solve. Consistent with these views, several studies showed that men are more interested and more engaged in technology than women, as a result, they are more experienced in using computers (Chen, 1986; Gnambs, 2021; Heo & Toomey, 2020; Temple & Lips, 1989). Due to this prior experience, males were more positive toward computers and computer-related tasks and jobs (Whitley, 1997), which may lead to more positive attitudes toward e-learning as found in several studies (Liaw & Huang, 2011; Ong & Lai, 2006; Wang et al., 2009).

1.2. Self-Regulated Learning and Attitudes Towards E-Learning

Effective learning requires students to self-regulate their motivation, cognition, and behavior (Zimmerman, 1989). Self-regulated learning (SLR) is defined as "the degree to which students are metacognitively, motivationally, and behaviorally active participants in their learning process" (Zimmerman, 2008, p.2). In other words, self-regulated learning involves high motivation and self-direction. According to Zimmerman (2000), self-regulated learning (SRL) comprises three cycles (1) forethought, (2) performance or volitional control, and (3) self-reflection. The forethought phase includes two components namely, task analysis and motivational beliefs. In this stage, students are expected to create an effective learning plan by identifying their learning goals. These goals should be challenging but attainable, proximal, and hierarchically organized with larger overarching goals. Apart from setting goals, students should allocate the appropriate amount of time to complete the learning tasks which should be framed and reframed by the educators to serve basis for future planning. As for the performance phase, self-control and self-observation components are emphasized for students which are expected to use different strategies towards achieving their learning goals as well as to observe the effectiveness of these to complete their learning tasks. Educators can help students at this phase by teaching and modeling various strategies that can be used for completing a learning task. In this stage, educators should equip students with a variety of strategies they can use for completing a task. Finally, the self-reflection phase includes self-judgment and self-reaction

which requires students to self-reflect on their learning outcomes and experiences. This phrase highlights the importance of focusing on what students can learn from their experiences and improve it next time. Simply, self-regulation addresses the self-generated thoughts, feelings, and actions of students which helps them attain the pre-defined goals (Zimmerman, 1994) and aids with the achievement of students in their learning (McCoach, 2002).

Recent research on SRL revealed that many factors are closely related to students' self-regulated learning. To illustrate, in a study conducted by Cazan (2012), self-regulation was found to have a positive relationship with academic adjustment. Similarly, Zimmerman and Kitsantas (2014) emphasized the predictive role of self-regulation in students' grade point average (GPA) and their academic performance. All learning environments, online or not, require learners to attend class, learn the material, submit homework, and do group work (Paul & Jefferson, 2019). However, e-learning environments, unlike face-to-face learning environments, are learner-centered and require autonomy as they present many choices for the learners (Andrade & Bunker, 2011). In addition, e-learning requires them to be digitally skillful to be able to find their way around the learning interface (Hillman et al., 1994). Thus, in e-learning, the control of the process is mostly with the learner and requires the learner to manage his learning and to choose among different options to manage the process. Therefore, success in e-learning is closely related to the self-regulated learning levels experienced by learners (Nikolaki et al., 2017).

1.3. Curiosity and E-learning Attitudes

Apart from the importance of e-learning, the interest in curiosity has gained attention and highlighted the scientific interest in multiple disciplines (Dan et al., 2020). Different disciplinary approaches have proposed various models and reported different to measure curiosity. Initially, epistemic curiosity (EC) is defined as the motive to seek, obtain and make use of new knowledge (Berlyne, 1954; Litman, 2005; Loewenstein, 1994). To put it simply, it is a multifaceted construct consisting of distinctive yet highly correlated dimensions (Nakamura et al., 2021). Berlyne (1966) emphasized two dimensions of EC: diversive and specific. While diversive EC is motivated by feelings of boredom and desire to seek stimulation regardless of source or content and specific EC is motivated by curiosity and initiated a detailed investigation of novel stimuli to acquire new information (p.31). These two dimensions were found to be highly correlated by Litman and Spielberg (2003) who introduced another dimension, the feeling of deprivation. Additionally, Litman (2005) added two more dimensions to EC labeled as Interest-type (I-type) and Deprivation-type (D-type). First, I-type EC is defined as “a desire for new information anticipated to increase pleasurable feelings of situational interest” whereas D-type EC is based on “a motive to reduce unpleasant experiences of feeling deprived of new knowledge” (Lauriola et al., 2015, p. 202). The two dimensions were investigated by distinguished scholars who explored their association with learning and school performance (Eren & Coskun, 2016), acquisition of knowledge (Rotgans & Schmidt, 2014), and self-regulated behavior (Lauriola et al., 2015). Finally, research on individual differences in EC suggests that its I-type and D-type dimensions are related to the variety of underlying processes, information-seeking activities as well as self-directed learning goals (Lauriola et al., 2015). Among the predictors of these differences is the use of different regulations strategies by the learner during the learning process.

Considering the current COVID-19 pandemic which led to a sudden shift to online learning, determining the impact of individual characteristics on students' attitudes towards e-learning is an important research area for educational researchers. Gender, EC, and SRL may be influential factors in students' attitudes toward e-learning. To this end, this study aims to find out the relations and associations among higher education students' gender, SRL, EC, and attitudes towards e-learning.

2. METHOD

2.1. Setting and Participants

The data of the study were collected in the 2020-2021 Fall semester. The sample comprised 2348 (862 males, 1576 females) undergraduate students enrolled in a foundation (non-profit, private) university in Turkey. The participants were studying in various disciplines such as Foreign Languages (N=506), Social Sciences (N=362), Medical Sciences (265), Communication (N=184), Architecture (N=175), Law (N=144), and Other (802) (see Table 1).

Due to the COVID 19 pandemic, all students were taking all their courses online. For this study, they volunteered and filled in the online questionnaires. It was stated to all participants that the questionnaires were anonymous and that they could withdraw at any time. Informed consent was received with yes / no screen questions from all participants before filling out the online questionnaires.

Table 1. Summary of participants' gender, department, and EL, SL, E-Learn Scales Quarters*.

Sex	f	%	Department	f	%	Quarter	Scale		
							EC (f)	SL (f)	E-Learn (f)
Man	862	35%	Foreign Languages	506	21%	First	390	478	489
Woman	1576	65%	Social Sciences	362	15%	Second	834	719	796
			Medical Sciences	265	11%	Third	778	792	691
			Communication	184	8%	Forth	436	449	462
			Architecture	175	7%				
			Law	144	6%				
			(Other)	802	32%				
TOTAL	2438	100%		2438	100%		2438	2438	2438

*Rounded to the nearest decimal.

2.2. Data Collection Tools

2.2.1. The curiosity and exploration inventory-ii

For this study, the Turkish version (Acun et al., 2013) of The Curiosity and Exploration Inventory-II (Acun et al., 2013) developed by Kashdan (2009) was used to measure the epistemic curiosity levels of the students. The self-report scale consists of 10 items with two subscales. The two subscales are the stretching subscale, which is the motivation for seeking information and new experience, and the acceptance of uncertainty and embracing subscale, which reflects the desire to discover the new, uncertain, and unpredictable in daily life. Students responded on a four-point frequency scale where 1=never and 4= always. Higher scores indicate higher epistemic curiosity. The validity and reliability of the original English version of the scale were tested with three different samples and alpha reliability coefficients were reported between .75 and .86 for these samples. The validity and reliability of the Turkish version were tested with two different samples and alpha reliability coefficients for these two samples were calculated as .81 and .82 (Acun et al., 2013). For the current study, the alpha reliability coefficient was calculated as .80.

2.2.2. Self-regulation scale

To measure the self-regulation of the students, the Turkish version (Duru et al., 2009) of the Self-Regulation Scale developed by Tuckman (2002) was used. The scale consists of 9 items –e.g. “I seem to have enough time to complete my work” and “I organize my time”. Students responded on a four-point frequency scale where 1=never and 4= always. Higher scores indicate higher levels of self-regulation. The Alpha reliability coefficient for the original version was

.88 and for the Turkish version was .73. For the current study, the alpha reliability coefficient was calculated as .73.

2.2.3. Attitudes toward the e-learning scale

To measure students' attitudes towards online learning, the Attitude Scale Towards E-Learning Scale developed by Haznedar and Baran (2012) was used. The scale is a five-point Likert scale where 1= definitely disagree and 5= definitely agree. The scale consists of 20 items, e.g. "I like working at my own pace with e-learning" and "E-learning increases the productivity of the learner". Higher scores indicate a positive attitude towards e-learning. The Alpha reliability coefficient of the scale was calculated as .93. For the current study, the alpha reliability coefficient was .97.

2.3. Data Analysis

The data in this study were analyzed in two steps. In the first step, multiple regression analysis was carried out to examine whether gender, EC, and SRL predict attitudes towards e-learning. Before the analysis, the suitability of the dataset for the analysis was tested. There was linearity as assessed by partial regression plots and a plot of studentized residuals against the predicted values. There was the independence of residuals, as assessed by a Durbin-Watson statistic of .086. Homoscedasticity was assessed by visual inspection of a plot of studentized residuals versus unstandardized predicted values and confirmed. For multicollinearity, tolerance values were assessed. All the values were greater than 0.1. No evidence of multicollinearity was detected. There were no studentized deleted residuals greater than ± 3 standard deviations, no leverage values greater than 0.2, and values for Cook's distance above 1. Investigation of the Q-Q Plot confirmed the normality of the data.

In the second step, to further understand the relationships among the variables, the association rule mining was run. With the change in the type and amount of data, it was understood that it would not be possible to obtain meaningful information in the analysis of the available data with existing methods and technologies (Ayık et al., 2007). This limitation prompted researchers to study in-depth for new analysis methods. As a result of these studies, a new data analysis method, data mining has emerged, which enables the analysis of data from different angles and to summarize this data by converting it into useful information (Delavari et al., 2008; Narli et al., 2014). The researchers defined the data analysis method as the process of discovering meaningful information from data stacks using methodologies such as artificial intelligence, statistics, and machine learning (Tan et al., 2006; Aran et al., 2019). The purpose of data mining is to reveal the whole systematic relationships between variables that do not appear to be relational or are assumed to be unrelated (Luan, 2002). Data mining includes different analysis models within itself. Many studies have categorized these models in different classifications (Ayık et al., 2007; Baker & Yacef, 2009; Baradwaj & Pal, 2012; Delavari et al., 2008; Luan, 2002; Narli et al., 2014). The most general definition of the association rule is categorized in the descriptive model which tries to reveal which events can occur simultaneously by examining the relations of the variables in the dataset with each other. The analysis methods used in this study are described below.

2.3.1. Association rules mining

The association rule is aimed at examining the $X \rightarrow Y$ events in the form of cause and effect with each other. Analysis of association rule is performed with sequential or parallel and scattered algorithms depending on the characteristics of the data set. Algorithms such as Apriori, STEM, and AIS are called sequential algorithms and are preferred in cases in which the analyzed data set can be counted (Garcia et al., 2010). Methods such as count distribution, parallel data mining, and common candidate partitioned database are parallel and distributed algorithms and are used for the analysis of large data sets (Agrawal & Srikant, 1994; Inokuchi,

et al., 2000; Zaki et al., 1997). In case the data set has a categorical structure, the apriori algorithm, which is one of the sequential algorithms, is often preferred in the analysis for the association rule (Agrawal & Srikant, 1994). In the scope of this study, the apriori algorithm was used for the association rule.

The Apriori Algorithm developed by Agrawal and Srikant (1994) is an algorithm that is generally used to determine product sales strategy, banking services, and social trends. Findings obtained with this algorithm are presented with support, confidence, lift, and coverage values (Zaki et al., 1997). The support value is the percentage equivalent of the data set of the rule obtained in the whole data set and is calculated with the following formula (Garcia et al., 2010; Merceron et al., 2010; Özçalıcı, 2017).

$$Support = \frac{n(X \cup Y)}{N}$$

In this formula $n(X \cup Y)$ refers to all cases in which X and Y are present together and N refers to the number of all cases in the total data set. In other words, this value shows the ratio of events or clusters in which X takes place to all events or sets for X and Y, which are different from each other (Güngör et al., 2013). The percentage equivalent of how much of the cases in which the X of the examined situation includes Y is the confidence value and is calculated with the following formula.

$$Confidence = \frac{n(X \cup Y)}{n(X)}$$

In this formula, $n(X \cup Y)$ corresponds to the number of cases in which both X and Y, while $n(X)$ only corresponds to the number of cases in which X is presented. The confidence value can only be zero if and only if there is no case in $n(X \cup Y)$ value, that is, X and Y together. Another important value obtained with the apriori algorithm is the lift value. The lift value, which expresses the rate of statistical realization of X and Y independently of each other, is calculated with the following formula.

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(Y)}$$

Lift value, which can take a value between 0 and ∞ according to this formula, is a parameter that helps to interpret how often events occur (Brin et al., 1997). Another important parameter for the apriori algorithm is the coverage value. Coverage values are parameters that show how often the present rule can be applied and it is calculated by the following formula (Garcia et al., 2010; Merceron et al., 2010).

$$Cover = Support(X) = P(X)$$

According to this formula, the coverage value of a situation is equal to the ratio of the cases in which X is located. Therefore, it takes a value between 0 and 1.

For association rules mining, all the variables should be categorical. In this study all the variables, except gender, were continuous. Therefore, EC, SRL, and attitudes towards e-learning variables were divided into 4 groups. For grouping, the students into curiosity, self-regulation, and attitudes towards e-learning groupings, the visual binning procedure was employed using SPSS. Binning was performed by applying cut-points at the mean and ± 1 standard deviation. For each variable, four binned categories were established. (Q1 = low, Q2 = moderately low, Q3 = moderately high, Q4 = high).

For regression we used IBM SPSS 25 and the association rules analyses were carried out using R Studio 1.3.1093 with R version 4.0.3 rules package. For the visualization of findings, we used diagrams.net 14.1.8.

3. RESULT

As previously stated, in the present study we proposed a possible relationship between gender, EC and SRL, and attitudes towards e-learning. The following section examines and reports the obtained results in detail.

3.1. Regression Analysis

To predict attitudes towards online learning from gender, EC, and SRL, a multiple regression analysis was run (see Table 2). Based on the results of regression analysis gender, EC and SRL statistically significantly predicted attitudes towards online learning, $F(3, 2447) = 44.570, p < .001$. All four variables added statistically significantly to the prediction, $p < .05$. R^2 for the overall model was 5% with an adjusted R^2 of 5%. However, the effect size was small according to Cohen (1988).

Table 2. Multiple regression results for attitudes towards online learning.

Online Learning Attitude	B	95% CI for B		SE B	β	R^2	ΔR^2
		Lower Bound	Upper Bound				
Constant	15.292	8.31	22.27	3.559		.05	.05
Gender	2.005	.172	3.84	.935	.043*		
Curiosity	.348	.816	1.23	.090	.079**		
Self-Regulation	1.022	.172	3.84	.105	.198**		

Note: B= unstandardized regression coefficient; CI= confidence interval; SE B= standard error of the coefficient; β = standardized coefficient; R^2 = coefficient of determination; ΔR^2 = adjusted R^2 .

* $p < .05$. ** $p < .01$

3.2. Association Rules Mining

To gain an in-depth analysis of the obtained data, data mining was further employed. We used the association rule mining technique. Association rule mining is generally defined as the process of exploring meaningful knowledge within data sets by making use of such methodology as artificial intelligence, statistics, and machine learning (Tan et al., 2006). To put it simply, association rule data mining (descriptive category) was applied by searching data for frequent if-then patterns and identifying the most important relationships. The following section of this study summarizes the results.

While establishing the association rule, the minimum support value was determined as 0.01 and the confidence value as 0.8. A total of 29 rules were reached that provide these values. The summary information on rule length distribution, support, confidence, coverage, lift, and frequency values regarding all rules were given in Table 3.

Table 3. Summary of quality measures of association rules*.

	Rule length distribution (lhs + rhs)	Support (%)	Confidence (%)	Coverage (%)	Lift	Count (f)
Minimum	3	0.01	0.8	0.01	1.23	26
1st Quarter	3	0.01	0.82	0.01	1.26	29
Median	3	0.01	0.86	0.02	1.33	35
Mean	3.20	0.02	0.87	0.02	1.35	40
3rd Quarter	3	0.02	0.91	0.02	1.4	47
Maximum	4	0.04	1	0.04	1.54	86

*Rounded to the nearest decimal.

When the distribution of the found rules was examined, it was seen that the minimum rule length (lhs + rhs) was 3 (n = 33) and the maximum rule length was 4 (n = 18). The minimum support and coverage value obtained was 0.01, and the highest was 0.04. It was found that the highest Conf value was obtained as 100%. The least repeating rule was n=26, while the most repeating rule was repeated n=86 times. Lastly, the average lift value was found to be 1.36 (min: 1.23, max: 1.50).

The 29 rules within the scope of this research will be presented in two categories. 22 of the rules were composed of different rule sets, including department variables of students, and the remaining 7 rules were composed of only quarters in measurement tools.

3.2.1. Department based findings

A total of 3 rules were found for students who enrolled in EduIns. (n=80, 3,3%), (see Table 3) When these rules were examined, it was revealed that students with 3rdQ (supp: 0.02; conf: 0.95; cov: 0.02; lift: 1.47; f: 39) on the E-Learn scale, 2ndQ (supp: 0.01; conf: 1; cov: 0.01; lift: 1.54; f: 26) on the SRL scale, and 3rdQ (supp: 0.01; conf: 0.87; cov: 0.01; lift: 1.34; f: 27) on the EC scale were female (see Table 4).

Table 4. Association rules and their support, confidence, coverage, and lift values*.

Rule	Mathematical Rule lhs → rhs	Support (%)	Confidence (%)	Coverage (%)	Lift	Count (f)
[R1]	(Dep=EduIns) ∪ (E-Learn=3rdQ) → (Sex=F)	0.02	0.95	0.02	1.47	39
[R2]	(Dep=EduIns) ∪ (SRL=2ndQ) →(Sex=F)	0.01	100	0.01	1.54	26
[R3]	(Dep=EduIns) ∪ (EC=3rdQ) →(Sex=F)	0.01	0.87	0.01	1.34	27

*F: Female, Dep: Department, Q: Quarter

In EduFa (n=122, 5%), it was understood that female participants were in the 2nd and 3rd quarters [R4...R9] in all E-Learn, SRL, and EC scales (see Table 5). In other words, in one or more of these scales, no pattern was found for education faculty students who were in the 1st and 4th quarters.

Table 5. Association rules and their support, confidence, coverage, and lift values*.

Rule	Mathematical Rule lhs → rhs	Support (%)	Confidence (%)	Coverage (%)	Lift	Count (f)
[R4]	$(\text{Dep}=\text{EduFa}) \cup (\text{E-Learn}=3\text{rdQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.92	0.02	1.42	35
[R5]	$(\text{Dep}=\text{EduFa}) \cup (\text{SRL}=2\text{ndQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.82	0.02	1.26	36
[R6]	$(\text{Dep}=\text{EduFa}) \cup (\text{EC}=3\text{rdQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.83	0.01	1.28	29
[R7]	$(\text{Dep}=\text{EduFa}) \cup (\text{SRL}=3\text{rdQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.88	0.02	1.35	35
[R8]	$(\text{Dep}=\text{EduFa}) \cup (\text{E-Learn}=2\text{ndQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.85	0.02	1.31	34
[R9]	$(\text{Dep}=\text{EduFa}) \cup (\text{EC}=2\text{ndQ}) \rightarrow (\text{Sex}=\text{WF})$	0.02	0.90	0.02	1.38	43

*F: Female, Dep: Department, Q: Quarter

A total of 7 rules for MedVoc (n=144, 5.9%) and MedSci (n=265, 10.8%) were obtained (see Table 5). MedVoc students, those in the 2nd [R10] and 4th [R11] quarters in SRL, and those in the 2nd [R12] quarter in EC were identified as female.

Table 6. Association rules and their support, confidence, coverage, and lift values*.

Rule	Mathematical Rule lhs → rhs	Support (%)	Confidence (%)	Coverage (%)	Lift	Count (f)
[R10]	$(\text{Dep}=\text{MedVoc}) \cup (\text{SRL}=2\text{thQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.82	0.01	1.26	27
[R11]	$(\text{Dep}=\text{MedVoc}) \cup (\text{SRL}=4\text{ndQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.90	0.01	1.39	27
[R12]	$(\text{Dep}=\text{MedVoc}) \cup (\text{EC}=2\text{ndQ}) \rightarrow (\text{Sex}=\text{F})$	0.02	0.86	0.02	1.33	50
[R13]	$(\text{Dep}=\text{MedSci}) \cup (\text{SRL}=3\text{rdQ}) \cup (\text{E-Learn}=1\text{stQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.91	0.01	1.4	30
[R14]	$(\text{Dep}=\text{MedSci}) \cup (\text{EC}=2\text{ndQ}) \cup (\text{SRL}=2\text{ndQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	100	0.01	1.54	29
[R15]	$(\text{Dep}=\text{MedSci}) \cup (\text{EC}=2\text{ndQ}) \cup (\text{SRL}=3\text{rdQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.97	0.01	1.5	32
[R16]	$(\text{Dep}=\text{MedSci}) \cup (\text{EC}=2\text{ndQ}) \cup (\text{E-Learn}=2\text{ndQ}) \rightarrow (\text{Sex}=\text{F})$	0.01	0.95	0.02	1.46	36

*F: Female, Dep: Department, Q: Quarter

Furthermore, as shown in the Table 6 above, For MedSci students, those in SRL 3rdQ and E-Learn 1stQ (supp: 0.01; conf: 0.91; cov: 0.01; lift; 1.4; f: 30) were determined to be women, and along with that, both EC 2nd and; [R14] those in SRL 2nd (supp: 0.01; conf: 1; cov: 0.01; lift; 1.54; f: 29) quarter, [R15] those in SL 3rd (supp: 0.01; conf: 0.97; cov: 0.01; lift; 1.5; f: 32) quarter or, [R16] those in E-Learn 2nd (supp: 0.01; conf: 0.95; cov: 0.02; lift; 1.46; f: 36) quarter were obtained as the pattern of female students.

The last section of the department-based rules consists of 6 rules involving Arch students. The findings revealed that the participants from Arch in E-Learn 1st [R18] and 3rd [R19] quarter, in 1st [R17] and 2nd [R22] quarters in EC and SRL 2nd [R20] and 3rd [R21] were female students (see Table 7).

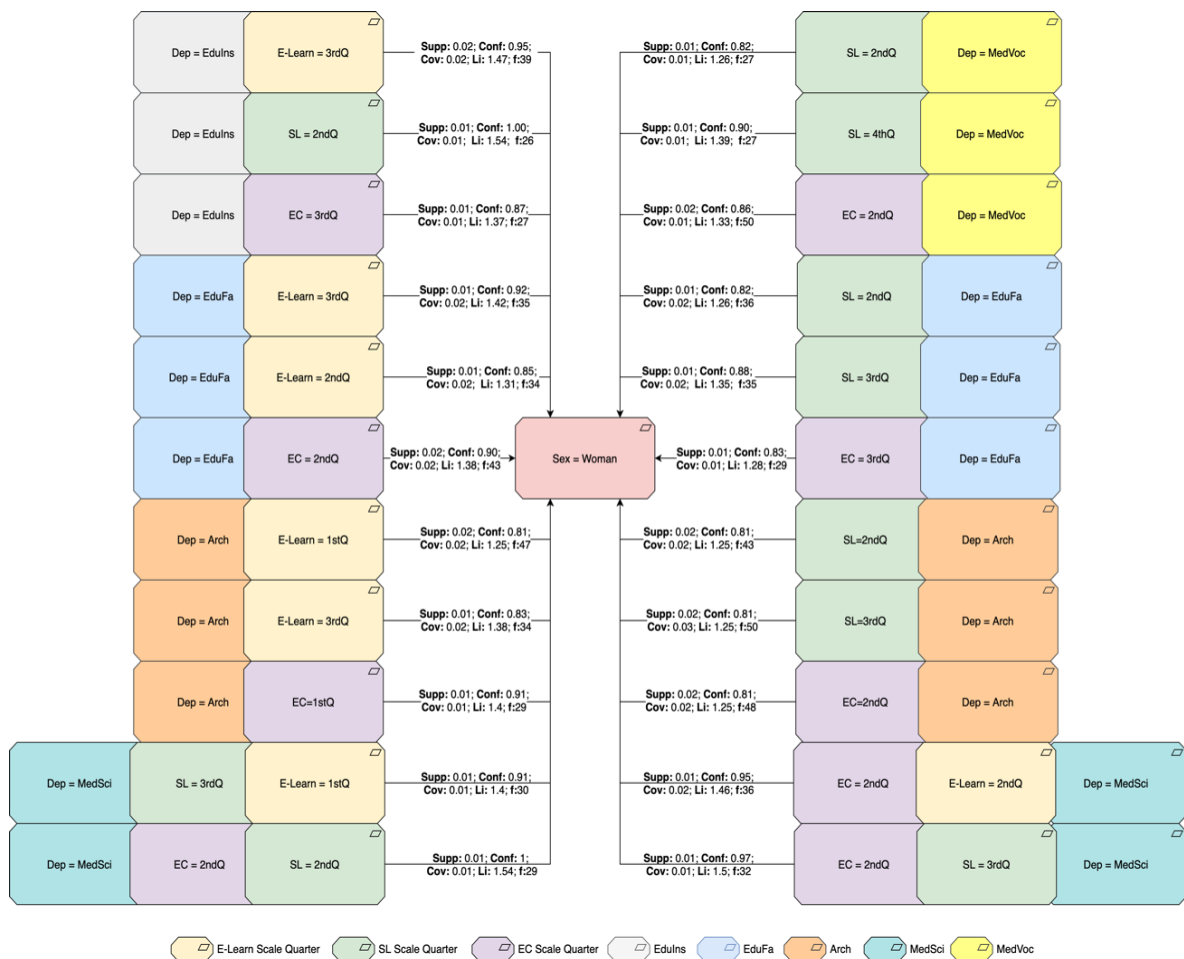
Table 7. Association rules and their support, confidence, coverage, and lift values*.

Rule	Mathematical Rule lhs → rhs	Support (%)	Confidence (%)	Coverage (%)	Lift	Count (f)
[R17]	(Dep=Arch) ∪ (EC=1stQ) →(Sex=F)	0.01	0.91	0.01	1.4	29
[R18]	(Dep=Arch) ∪ (E-Learn=1stQ) →(Sex=F)	0.02	0.81	0.02	1.25	47
[R19]	(Dep=Arch) ∪ (E-Learn=3rdQ) →(Sex=F)	0.01	0.83	0.02	1.28	34
[R20]	(Dep=Arch) ∪ (SRL=2ndQ) →(Sex=F)	0.02	0.81	0.02	1.25	43
[R21]	(Dep=Arch) ∪ (SRL=3rdQ) →(Sex=F)	0.02	0.81	0.03	1.25	50
[R22]	(Dep=Arch) ∪ (EC=2ndQ) →(Sex=F)	0.02	0.81	0.02	1.25	48

*F: Female, Dep: Department, Q: Quarter

Finally, based on the gathered data all department-based rules were given in Figure 1. When all these rules were examined, it was understood that there was a pattern in the data of EduIns, EduFa, MedVoc, MedSci, and Arch departments in this data set, which includes participants from 17 different departments. Therewithal, no pattern was obtained for male participants even though there were both female and male participants. The most unusual finding regarding the department-based rules was that the predictor variable of all rules points to female participants. In other words, the main element of the pattern created in all rules was the gender variable of the female participants.

Figure 1. Departmental rules.



3.2.2. Scales based findings

All rules based on scales were given in the Table 8. When all these rules were examined, it was concluded that those with EC 1stQ and also those in [R23] SL 4th quarter (supp: 0.02; conf: 0,93; cov: 0.02; lift; 1.44; f: 54), [R24] E-learn 3rdQ (supp: 0.03; conf: 0.80; cov: 0.04; lift; 1.23; f: 76), [R24] SRL 3rdQ (supp: 0.04; conf: 0.84; cov: 0.04; lift; 1.3; f: 86) were female.

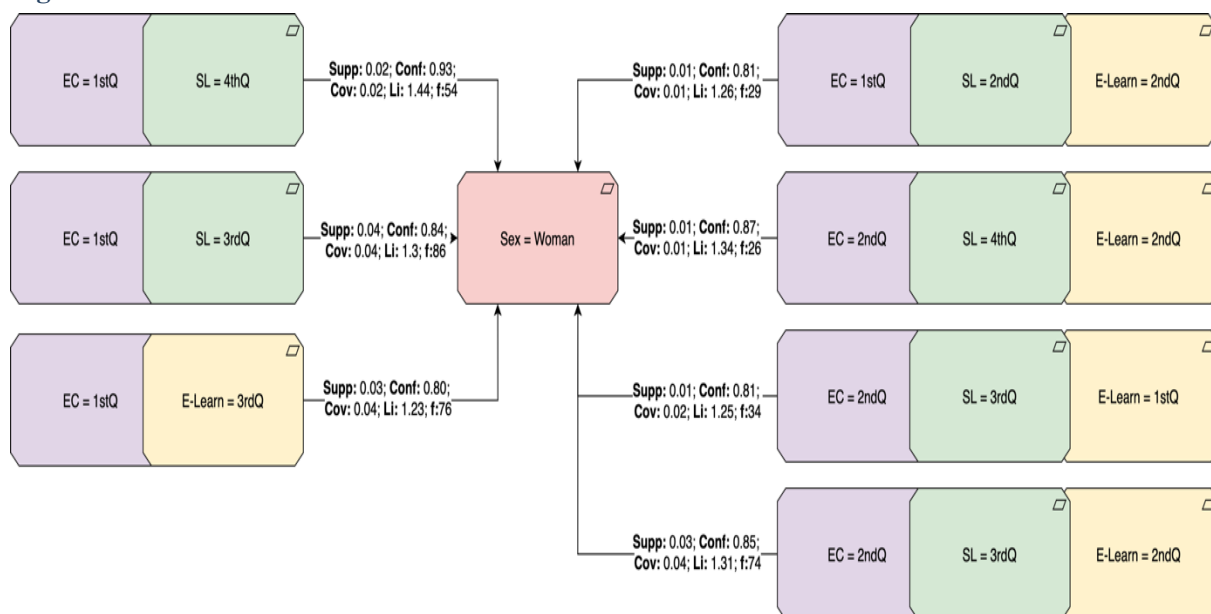
Table 8. Association rules and their support, confidence, coverage, and lift values*.

Rule	Mathematical Rule lhs → rhs	Support (%)	Confidence (%)	Coverage (%)	Lift	Count (f)
[R23]	(EC=1stQ) ∪ (SRL=4thQ) →(Sex=F)	0.02	0.93	0.02	1.44	54
[R24]	(EC=1stQ) ∪ (E-Learn=3rdQ) →(Sex=F)	0.03	0.80	0.04	1.23	76
[R25]	(EC=1stQ) ∪ (SRL=3rdQ) →(Sex=F)	0.04	0.84	0.04	1.3	86
[R26]	(EC=1stQ) ∪ (SRL=2ndQ) ∪ (E-Learn=2ndQ) →(Sex=F)	0.01	0.81	0.01	1.26	29
[R27]	(EC=2ndQ) ∪ (SRL=4thQ) ∪ (E-Learn=2ndQ) →(Sex=F)	0.01	0.87	0.01	1.34	26
[R28]	(EC=2ndQ) ∪ (SRL=3rdQ) ∪ (E-Learn=1stQ) →(Sex=F)	0.01	0.81	0.02	1.25	34
[R29]	(EC=2ndQ) ∪ (SRL=3rdQ) ∪ (E-Learn=2ndQ) →(Sex=F)	0.03	0.85	0.04	1.31	74

*F: Female, Dep: Department, Q: Quarter

The remaining 4 rules on scale-based were 4 rule lengths. Accordingly, all participants who fulfilled the requirements [R26] (EC=1stQ) ∪ (SRL=2ndQ) ∪ (E-Learn=2ndQ) and [R27] (EC=2ndQ) ∪ (SRL=4thQ) ∪ (E-Learn=2ndQ) were women. Finally, along with EC 2ndQ and SRL 3rdQ, all participants in E-Learn, both from 1stQ and 2ndQ were also stated as women. All scale-based rules were given in Figure 2 below.

Figure 2. Scale based rules.



4. DISCUSSION and CONCLUSION

As previously stated, in the present study we proposed that individual differences might be an active and influential on higher education students' attitudes toward e-learning. The statistical analysis of multiple regression revealed that gender, EC and SRL were significant predictors of attitudes towards e-learning. However, the effect size was low. So, to further analyze the relations among the variables, we conducted association rule mining. As expected, we could detect several associations among variables that cannot be detected via regression models. According to the association rule in the descriptive model category, gender was found to have a predictive role in the two behaviors. Specifically, females outperformed males both in SRL and EC during online learning. These findings were contrary to previous studies that revealed no significant gender differences with respect to SRL (Çalışkan & Sezgin-Selcuk, 2010; Hargittai & Shafer, 2006; Yükseltürk & Bulut, 2009). Besides, the findings were opposite to the study conducted by Bashir and Bashir (2016) indicating that males showed higher self-regulation as compared to females. The only partial similarity was reported by Senler and Sungur-Vural (2012) stating that females showed higher self-regulation and effort regulation compared with males.

Considering the statistical analysis methods used in the study, it is important to evaluate the findings revealed by data mining. The association rule used in this study, although it is less used in educational sciences, has wide usage in several areas as Computer Science (Chen et al. 2021), Engineering (Çakır et al., 2021), Decision Sciences (Prathama et al.2021), Mathematics (Li et al., 2020) Business, Management and Accounting (Moodley et al., 2020), Medicine and Dentistry (Tandan et al., 2021), Social Sciences (Cömert & Akgün, 2021), Energy (Odabaşı & Yıldırım, 2019), Environmental Science (Nagata et al., 2014) and Psychology (Elia et al., 2019). Besides, in order to compare the performance of this analysis method, which includes more than one algorithm, many variables such as the distribution, features, and characteristics of the data set should be considered. Therefore, it can be said that which algorithm gives better performance from association rules varies according to the properties of the dataset (Borgelt & Kruse, 2002). With data mining techniques in which appropriate algorithms are selected, it seems possible to reveal detailed characteristic relationships about students and to make predictions for the future (Arora & Badal, 2014).

Based on these overviews, the present study revealed that gender might have a predictive role on SRL and EC among higher education students during e-learning which should be addressed in further studies. Similar to face-to-face education, individual differences have an active and influential role in teaching and learning online as well. Therefore, we propose that future research should examine the role of such personal characteristics in various educational contexts to provide suggestions for more effective pedagogical practices. To gather in-depth information, we also recommend that data mining can be used as a statistical method in educational and psychological research.

Acknowledgments

We would like to thank all students participating in the study.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Bahcesehir University, 10.02.2021 - 2021/02/16.

Authorship Contribution Statement

Ergun Akgun: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Enisa Mede:** Methodology, Supervision, Validation, Formal Analysis, and Writing. **Seda Sarac:** Methodology, Supervision, Validation, Formal Analysis, and Writing.

Orcid

Ergun Akgun  <https://orcid.org/0000-0002-7271-6900>

Enisa Mede  <https://orcid.org/0000-0002-6555-5248>

Seda Sarac  <https://orcid.org/0000-0002-4598-4029>

REFERENCES

- Acun, N., Kapıkıran, Ş., & Kabasakal, Z. (2013). Merak ve keşfetme ölçeği II: Açımlayıcı ve doğrulayıcı faktör analizleri ve güvenilirlik çalışması [Trait Curiosity and Exploration Inventory-II: Exploratory and Confirmatory Factor Analysis and Its Reliability] *Türk Psikoloji Yazıları*, 16(31), 74-85.
- Agrawal, R., & Srikant, R. (1994, September, 487-489). *Fast algorithms for mining association rules*. Proc. of the 20th VLDB Conference, San Francisco, USA.
- Aixia, D., & Wang, D. (2011). Factors influencing learner attitudes toward e-learning and development of e-learning environment based on the integrated e-learning platform. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 1(3), 264-268.
- Altun, T., Akyıldız, S., Gülay, A., & Özdemir, C. (2021). Investigating education faculty students' views about asynchronous distance education practices during Covid-19 isolation period. *Psycho-Educational Research Reviews*, 10(1), 34–45.
- Andrade, M.S., & Bunker, E.L. (2011). The role of SRL and TELEs in distance education: Narrowing the gap. In *Fostering self-regulated learning through ICT* (pp. 105-121). IGI Global. <https://doi.org/10.4018/978-1-61692-901-5.ch007>
- Aran, O., Bozkir, A., Gok, B., & Yagci, E. (2019). Analyzing the views of teachers and prospective teachers on information and communication technology via descriptive data mining. *International Journal of Assessment Tools in Education*, 6(2), 314-329. <https://doi.org/10.21449/ijate.537877>
- Arora, R.K., & Badal, D. (2014). Mining association rules to improve academic performance. *International Journal of Computer Science and Mobile Computing*, 3(1), 428-433.
- Ayık, Y.Z., Özdemir, A., & Yavuz, U. (2007). Lise türü ve lise mezuniyet başarısının, kazanılan fakülte ile ilişkisinin veri madenciliği tekniği ile analizi [Analysis of the relationship of high school type and high school graduation success with the faculty entered by data mining technique] *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 10(2), 441-454.
- Baker, R.S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-17. <https://doi.org/10.5281/zenodo.3554657>
- Baradwaj B.K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417. <https://doi.org/10.48550/arXiv.1201.3417>
- Bashir, H., & Bashir, L. (2016). Investigating the relationship between self-regulation and spiritual intelligence of higher secondary school students. *Indian Journal of Health and Wellbeing*, 7(3), 327.

- Bastiaens, T.J., & Martens, R.L. (2000). Conditions for web-based learning with real events. In *Instructional and cognitive impacts of web-based education* (pp. 1-31). IGI Global. <https://doi.org/10.4018/978-1-878289-59-9.ch001>
- Berlyne, D.E. (1966). Curiosity and exploration. *Science*, 153(3731), 25-33. <https://doi.org/10.1126/science.153.3731.25>
- Berlyne, D.E. (1954). A theory of human curiosity. *British Journal of Psychology*, 45, 180–191.
- Bhuasiri, W., Xaymoungkhoun, O., Zo, H., Rho, J.J., & Ciganek, A.P. (2012). Critical success factors for e-learning in developing countries: A comparative analysis between ICT experts and faculty. *Computers & Education*, 58(2), 843-855. <https://doi.org/10.1016/j.compedu.2011.10.010>
- Borgelt, C., & Kruse, R. (2002). Induction of association rules: Apriori implementation. In *Comstat* (pp. 395-400). Physica-Verlag Heidelberg.
- Brin, S., Motwani, R., Ullman, J.D., & Tsur, S. (1997, June, 255-264). *Dynamic itemset counting and implication rules for market basket data*. Proceedings of the 1997 ACM SIGMOD international conference on Management of data, New York, USA. <https://doi.org/10.1145/253260.253325>
- Cazan, A.M. (2012). Self-regulated learning strategies–predictors of academic adjustment. *Procedia-Social and Behavioral Sciences*, 33, 104-108. <https://doi.org/10.1016/j.sbspro.2012.01.092>
- Chen, M. (1986). Gender and computers: The beneficial effects of experience on attitudes. *Journal of Educational Computing Research*, 2(3), 265-282. <https://doi.org/10.2190%2FWDRY-9K0F-VCP6-JCCD>
- Chen, S., Yuan, Y., Luo, X.R., Jian, J., & Wang, Y. (2021). Discovering group-based transnational cyber fraud actives: A polymethodological view. *Computers & Security*, 102217. <https://doi.org/10.1016/j.cose.2021.102217>
- Colley, A., & Comber, C. (2003). Age and gender differences in computer use and attitudes among secondary school students: what has changed?. *Educational Research*, 45(2), 155-165. <https://doi.org/10.1080/0013188032000103235>
- Cömert, Z., & Akgün, E. (2021). Game preferences of K-12 level students: analysis and prediction using the association rule. *Ilkogretim Online*, 20(1), 435-455. <http://doi.org/10.17051/ilkonline.2021.01.039>
- Çakır, E., Fışkın, R., & Sevgili, C. (2021). Investigation of tugboat accidents severity: An application of association rule mining algorithms. *Reliability Engineering & System Safety*, 209, 107470. <https://doi.org/10.1016/j.ress.2021.107470>
- Çalışkan, S., & Sezgin-Selçuk, G. (2010). Üniversite öğrencilerinin Fizik problemlerinde lullandıkları özdüzenleme stratejileri: Cinsiyet ve üniversite etkileri [Self-regulated strategies used by undergraduate students in physics problems: effects of gender and university]. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi*, 27(1), 50-62.
- Dan, O., Leshkowitz, M., & Hassin, R.R. (2020). On clickbaits and evolution: Curiosity from urge and interest. *Current Opinion in Behavioral Sciences*, 35, 150-156. <https://doi.org/10.1016/j.cobeha.2020.09.009>
- Delavari, N., Phon-Amnuaisuk, S., & Beikzadeh, M.R. (2008). Data mining application in higher learning institutions. *Informatics in Education-International Journal*, 7, 31-54.
- Duru, E., Balkıs, M., Buluş, M., & Duru, S. (2009, October, 57-73). Öğretmen adaylarında akademik erteleme eğiliminin yordanmasında öz düzenleme, akademik başarı ve demografik değişkenlerin rolü [The role of self-regulation, academic achievement and demographic variables in the prediction of academic procrastination in teacher candidates]. 18th Educational Sciences Congress, İzmir, Türkiye.

- Elia, G., Solazzo, G., Lorenzo, G., & Passiante, G. (2019). Assessing learners' satisfaction in collaborative online courses through a big data approach. *Computers in Human Behavior*, 92, 589-599. <https://doi.org/10.1016/j.chb.2018.04.033>
- Erarslan, A., & Topkaya, E.Z. (2017). EFL students attitudes towards e-learning and effect of an online course on students success in English. *The Literacy Trek*, 3(2), 80-101.
- Eren, A., & Coskun, H. (2016). Students' level of boredom, boredom coping strategies, epistemic curiosity, and graded performance. *The Journal of Educational Research*, 109(6), 574-588. <https://doi.org/10.1080/00220671.2014.999364>
- Garcia, E., Romero, C., Ventura, S., Castro, C., & Calders, T. (2010). Association rule mining in learning management systems. In V. Kumar (Ed.). *Handbook of educational data mining*. (pp. 93-106). Taylor & Francis Group.
- Gnambs, T. (2021). The development of gender differences in information and communication technology (ICT) literacy in middle adolescence. *Computers in Human Behavior*, 114, 1-10. <https://doi.org/10.1016/j.chb.2020.106533>
- Gunnarsson, C.L. (2001). Development and assessment of students: Attitudes and achievement in a business statistics course taught online. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 3(2).
- Güngör, E., Yalçın, N., & Yurtay, N. (2013, Kasım, 122-127). *Apriori algoritması ile teknik seçmeli ders seçim analizi [Selection Behavior Analysis of Technical Elective Courses Using Apriori Algorithm]*. Pro. UZEM 2013 Ulusal Uzaktan Eğitim ve Teknolojileri Sempozyumu, Konya, Türkiye.
- Hargittai, E., & Shafer, S. (2006). Differences in actual and perceived online skills: The role of gender. *Social Science Quarterly*, 87(2), 432-448. <https://doi.org/10.1111/j.1540-6237.2006.00389.x>
- Haznedar, Ö., & Baran, B. (2012). Eğitim fakültesi öğrencileri için e-öğrenmeye yönelik genel bir tutum ölçeği geliştirme çalışması [Development of a general attitude scale towards e-learning for faculty of education students]. *Eğitim Teknolojisi Kuram ve Uygulama*, 2(2), 42-59.
- Heo, M., & Toomey, N. (2020). Learning with multimedia: The effects of gender, type of multimedia learning resources, and spatial ability. *Computers & Education*, 146, 103747. <https://doi.org/10.1016/j.compedu.2019.103747>
- Hillman, D.C., Willis, D.J., & Gunawardena, C.N. (1994). Learner-interface interaction in distance education: An extension of contemporary models and strategies for practitioners. *American Journal of Distance Education*, 8(2), 30-42. <https://doi.org/10.1080/08923649409526853>
- Howland, J.L., & Moore, J.L. (2002). Student perceptions as distance learners in Internet-based courses. *Distance Education*, 23(2), 183-195. <https://doi.org/10.1080/015879102000009196>
- Inokuchi, A., Washio, T., & Motoda, H. (2000, September, 13-23). *An apriori-based algorithm for mining frequent substructures from graph data*. Proceedings of the 2000 European symposium on the principle of data mining and knowledge discovery (PKDD'00), Lyon, France.
- Kashdan, T.B. (2009). *Curious? Discover the missing ingredient to a fulfilling life*. William Morrow.
- Lauriola, M., Litman, J.A., Mussel, P., De Santis, R., Crowson, H.M., & Hoffman, R.R. (2015). Epistemic curiosity and self-regulation. *Personality and Individual Differences*, 83, 202-207. <https://doi.org/10.1016/j.paid.2015.04.017>
- Li, H., Wu, Y.J., & Chen, Y. (2020). Time is money: Dynamic-model-based time series data-mining for correlation analysis of commodity sales. *Journal of Computational and Applied Mathematics*, 370, 112659. <https://doi.org/10.1016/j.cam.2019.112659>

- Liaw, S.S., & Huang, H.M. (2011, September, 28-32). *A study of investigating learners' attitudes toward e-learning*. 5th International Conference on Distance Learning and Education, Paris, Fransa.
- Litman, J. (2005). Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition & Emotion*, 19(6), 793-814. <https://doi.org/10.1080/02699930541000101>
- Litman, J.A., & Spielberger, C.D. (2003). Measuring epistemic curiosity and its diversive and specific components. *Journal of Personality Assessment*, 80(1), 75-86. https://doi.org/10.1207/S15327752JPA8001_16
- Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1), 75-98. <https://psycnet.apa.org/doi/10.1037/0033-2909.116.1.75>
- Luan, J. (2002). Data mining and its applications in higher education. *New Directions For Institutional Research*, 2002(113), 17-36.
- Maio, G.R., Haddock, G., & Verplanken, B. (2018). *The psychology of attitudes and attitude change* (3rd ed.). Sage.
- Martens, R., Bastiaens, T., & Kirschner, P.A. (2007). New learning design in distance education: The impact on student perception and motivation. *Distance Education*, 28(1), 81-93. <https://doi.org/10.1080/01587910701305327>
- Martins, L.L., & Kellermanns, F.W. (2004). A model of business school students' acceptance of a web-based course management system. *Academy of Management Learning & Education*, 3(1), 7-26. <https://doi.org/10.5465/amle.2004.12436815>
- McCoach, D.B. (2002). A validation study of the school attitude assessment survey. *Measurement and Evaluation in Counseling and Development*, 35(2), 66. <https://doi.org/10.1080/07481756.2002.12069050>
- Merceron, A., Yacef, K., Romero, C., Ventura, S., & Pechenizkiy, M. (2010). Measuring correlation of strong symmetric association rules in educational data. *Handbook of Educational Data Mining*, 245-256.
- Mohammadi, N., Ghorbani, V., & Hamidi, F. (2011). Effects of e-learning on language learning. *Procedia Computer Science*, 3, 464-468. <https://doi.org/10.1016/j.procs.2010.12.078>
- Moodley, R., Chiclana, F., Caraffini, F., & Carter, J. (2020). A product-centric data mining algorithm for targeted promotions. *Journal of Retailing and Consumer Services*, 54, 101940. <https://doi.org/10.1016/j.jretconser.2019.101940>
- Nagata, K., Washio, T., Kawahara, Y., & Unami, A. (2014). Toxicity prediction from toxicogenomic data based on class association rule mining. *Toxicology Reports*, 1, 1133-1142. <https://doi.org/10.1016/j.toxrep.2014.10.014>
- Nakamura, S., Darasawang, P., & Reinders, H. (2021). The antecedents of boredom in L2 classroom learning. *System*, 98, 102469. <https://doi.org/10.1016/j.system.2021.102469>
- Narli, S., Aksoy, E., & Ercire, Y.E. (2014). Investigation of prospective elementary mathematics teachers' learning styles and relationships between them using data mining. *International Journal of Educational Studies in Mathematics*, 1(1), 37-57.
- Nikolaki, E., Koutsouba, M., Lykesas, G., Venetsanou, F., & Savidou, D. (2017). The support and promotion of self-regulated learning in distance education. *European Journal of Open, Distance and E-learning*, 20(1), 1-11.
- Odabaşı, Ç., & Yıldırım, R. (2019). Performance analysis of perovskite solar cells in 2013–2018 using machine-learning tools. *Nano Energy*, 56, 770-791. <https://doi.org/10.1016/j.nanoen.2018.11.069>
- Ong, C.S., & Lai, J.Y. (2006). Gender differences in perceptions and relationships among dominants of e-learning acceptance. *Computers in Human Behavior*, 22(5), 816-829. <https://doi.org/10.1016/j.chb.2004.03.006>

- Özçalıcı, M. (2017). Veri madenciliğinde birliktelik kuralları ve ikinci el otomobil piyasası üzerine bir uygulama [Association Rules in Data Mining and an Application in Second Hand Car Market]. *Ordu Üniversitesi Sosyal Bilimler Araştırma Dergisi*, 7(1), 45-58.
- Paul, J., & Jefferson, F. (2019). A comparative analysis of student performance in an online vs. face-to-face environmental science course from 2009 to 2016. *Frontiers in Computer Science*, 1,1-9. <https://doi.org/10.3389/fcomp.2019.00007>
- Prathama, F., Senjaya, W.F., Yahya, B.N., & Wu, J.Z. (2021). Personalized recommendation by matrix co-factorization with multiple implicit feedback on the pairwise comparison. *Computers & Industrial Engineering*, 152, 107033. <https://doi.org/10.1016/j.cie.2020.107033>
- Rotgans, J.I., & Schmidt, H.G. (2014). Situational interest and learning: Thirst for knowledge. *Learning and Instruction*, 32, 37-50. <https://doi.org/10.1016/j.learninstruc.2014.01.002>
- Selim, H.M. (2007). Critical success factors for e-learning acceptance: Confirmatory factor models. *Computers & Education*, 49(2), 396-413. <https://doi.org/10.1016/j.compedu.2005.09.004>
- Senler, B., & Sungur-Vural, S. (2012, September, 551-556). *Pre-service science teachers' use of self-regulation strategies related to their academic performance and gender*. The European Conference on Educational Research (ECER), Cadiz, Spain. <https://doi.org/10.1016/j.sbspro.2014.09.242>
- Suanpang, P. (2007). Students experience online learning in Thailand. In S. Hongladarom (Ed.), *Computing and philosophy in Asia*, (pp. 240-.252). Cambridge Scholar Publishing.
- Tan, P.N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Addison Wesley.
- Tandan, M., Acharya, Y., Pokharel, S., & Timilsina, M. (2021). Discovering symptom patterns of COVID-19 patients using association rule mining. *Computers in Biology and Medicine*, 104249. <https://doi.org/10.1016/j.combiomed.2021.104249>
- Temple, L., & Lips, H.M. (1989). Gender differences and similarities in attitudes toward computers. *Computers in Human Behavior*, 5(4), 215-226. [https://doi.org/10.1016/0747-5632\(89\)90001-0](https://doi.org/10.1016/0747-5632(89)90001-0)
- Tuckman, B. (2002, August). Academic procrastinators: Their rationalizations and web-course performance. the Annual Meeting of the American Psychological Association, Chicago, IL.
- Wang, Y.S., Wu, M.C., & Wang, H.Y. (2009). Investigating the determinants and age and gender differences in the acceptance of mobile learning. *British Journal of Educational Technology*, 40(1), 92-118. <https://doi.org/10.1111/j.1467-8535.2007.00809.x>
- Whitley Jr, B.E. (1997). Gender differences in computer-related attitudes and behavior: A meta-analysis. *Computers in Human Behavior*, 13(1), 1-22. [https://doi.org/10.1016/S0747-5632\(96\)00026-X](https://doi.org/10.1016/S0747-5632(96)00026-X)
- Yükseltürk, E., & Bulut, S. (2009). Gender differences in self-regulated online learning environment. *Journal of Educational Technology & Society*, 12(3), 12-22.
- Zaki, M.J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). Parallel algorithms for discovery of association rules. *Data Mining and Knowledge Discovery*, 1(4), 343-373.
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3), 329–339. <https://psycnet.apa.org/doi/10.1037/0022-0663.81.3.329>
- Zimmerman, B.J. (1994). *Dimensions of academic self-regulation: A framework for education. Regulation of learning and performance*. Lawrence Erlbaum.
- Zimmerman, B.J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation* (pp. 13-39). Academic Press.

Zimmerman, B.J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166-183. <https://doi.org/10.3102%2F0002831207312909>

Zimmerman, B., & Kitsantas, A. (2014). Comparing students' self-discipline and self-regulation measures and their prediction of academic achievement. *Contemporary Educational Psychology*, 39(2), 145-155. <https://doi.org/10.1016/j.cedpsych.2014.03.004>

4

Which scale short form development method is better? A Comparison of ACO, TS, and SCOFA

Hakan Kogar ^{1,*}

¹Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya

ARTICLE HISTORY

Received: June 01, 2021

Revised: Mar. 13, 2022

Accepted: July 12, 2022

Keywords:

Short form,
Scale development,
Ant colony optimization,
Tabu search,
Stepwise confirmatory
factor analysis.

Abstract: The purpose of this study is to identify which scale short-form development method produces better findings in different factor structures. A simulation study was designed based on this purpose. Three different factor structures and three simulation conditions were selected. As the findings of this simulation study, the model-data fit and reliability coefficients were reported for each factor structure in each simulation condition. All analyses were conducted under the R environment. According to the findings of this study, the increase in the level of misspecification and the decrease in the sample size can significantly affect the model-data fit. In a situation where the factor structure of the scale is getting more and more complex, model-data fit and Omega coefficients decrease. For scales with a unidimensional factor structure, all of the scale short-form development methods are recommended. For scales with multidimensional factor structure, Ant Colony Optimization, and Stepwise Confirmatory Factor Analysis algorithms and for scales with bifactor factor structure, the ACO algorithm is recommended. When viewed from the framework of metaheuristic algorithms, it has been identified that ACO produces better findings than Tabu Search.

1. INTRODUCTION

The use of short forms of psychological measurement tools has become widespread, especially in the last 20 years. Is it more important than the scale has a well-prepared factor structure but contains many items, or is it more important to obtain sufficient proof of validity by using time correctly with fewer items? This question has accelerated the work on short-form development. The main reason for this situation is to reduce the time and cost required for the application of the test, and the effort and length of the test that the participant would spend on the test items are appropriate (Kleka & Soroko, 2018). Due to these important reasons, academic studies to shorten the long forms of the scales have started to gain an important place in the social science literature. It should be noted that this situation has theoretical reasons as well as practical reasons. According to the Classical Test Theory (CTT), the high number of items makes significant contributions to the reliability of test scores, construct, and content validity. Many items are needed based on CTT to make valid and reliable measurements (Anastasi, 1982; Nunnally, 1978).

*CONTACT: Hakan Koğar ✉ hkogar@gmail.com 📍 Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

Nowadays, many new methods have been developed for scale short-form development (Leite et al., 2008; Olaru et al., 2015). It is possible to talk about dozens of methods within the framework of Metaheuristic Algorithms, Factor Analysis, Item Response Theory, and Rasch analysis. However, it is still known in the literature that classical approaches are frequently used in short-form development. These approaches are mostly based on item statistics or factor analysis approach. Item-total correlation and factor loadings or removing items with a low contribution to internal consistency constitute the basic implementation form of these approaches. This means determining the short form according to the item characteristics (Janssen et al., 2017). However, developing a short form of a scale is a much more complex and comprehensive process.

Classical short-form development approaches potentially have a significant disadvantage when it is desired to reduce the number of items of a previously validated measurement instrument. In these techniques, psychometrically poor items can be detected through item reliability or item-total correlation. However, in this case, depending on the test item excluded, the statistical findings of the remaining items and the findings of the general test will vary. Therefore, a stepwise item selection for the development of a short form will result in different item groups depending on the order of items eliminated (Janssen et al., 2017).

One of the methods developed to overcome many of the disadvantages of these classical approaches is Stepwise Confirmatory Factor Analysis (SCOFA). With this method, latent variable models are used, which provide a comprehensive framework for testing measurement models. This model overcomes one of the disadvantages of classical approaches by focusing directly on dimensionality and factor structure. Reducing the item pool in a measurement process can affect the factor structure of the instrument (Schroeders et al., 2016).

Obtaining a short form with both a strong factor structure and a high validity is quite difficult with traditional short-form development methods. Metaheuristic optimization algorithms have the potential to solve these difficulties because they can optimize and test with multiple validity criteria simultaneously for the developed short forms. Ant Colony Optimization (ACO), the first of the metaheuristic algorithms discussed within the scope of this research, was firstly developed by Colorni et al. (1991) as a metaheuristic to solve a wide range of combination-based problems and can be applied to situations where there are many possible solutions for a problem to be graded based on various criteria. This approach does not require the best solution to existing scale but instead focuses on finding a solution within the set of possible solutions that best meet certain criteria. One of the potential problems that this approach can solve is the development of the short form of the scales, undoubtedly. In this context, any combination of selected items is a possible solution, and these possible solutions will vary according to the previously established degree of competence (Janssen et al., 2017).

Tabu Search (TS), another metaheuristic algorithm, was designed by Marcoulides & Falk (2018) for short-form development.

“TS examines each set of short forms created by changing one item at a time. The main idea behind the TS method is to consistently identify the best short-form currently selected by examining other short forms neighbor the best available short form. If a neighbor examined short form fits better than the existing short form, it is selected as the most suitable new short form. If not, the neighbor short form under study is marked as "taboo." In other words, it is placed in a separate list so that it will not be re-evaluated until certain criteria are met.” (Raborn et al., 2020, p. 5).

Within the scope of this research, ACO and TS algorithms were selected from metaheuristic algorithms with the simulation study conducted by Raborn et al. (2020), taking into account other application-oriented studies and prevalence. In addition, due to the frequent use of classical approaches in the literature, the SCOFA technique, which is thought to represent these approaches and is based on iterations such as metaheuristic algorithms, was chosen. A

simulation study was designed based on the technique of developing the short forms of these three scales. The research questions of this study are as follows:

1. How are the model-data fit and reliability coefficients obtained from different scale short-form development methods according to different factor structures?
2. How are the model-data fits and reliability coefficients obtained from different scale short-form development methods according to different sample sizes, the correlation between factors, and model misspecification?
3. Which scale short-form development method performs best under various conditions?

2. METHOD

2.1. Simulation Conditions

In this study, various factor structures were primarily defined. Unidimensional, multidimensional, and bifactor structures were selected for this study. For each factor structure, research findings from real data sets were used. Simulation-based data sets were produced on various other features, especially the factor loadings of these measurement tools. Instructor Self-Disclosure Scale with 18 items for unidimensional structure (Cyanus & Martin, 2004), the Multidimensional Health Locus of Control Scale (LaNoue et al., 2015) with 3 factors and 18 items for the multidimensional structure, and The Anxiety Sensitivity Index-3 (Ebesutani et al., 2014) with 3 factors and 18 items for the bifactor structure were selected. Factor loading values of the scale vary between 0.28 - 0.70. The correlation coefficients between the factors vary between 0.41 - 0.45. Especially attention has been paid to ensure that each of these studies has an equal number of items. It is aimed to reduce the number of items in each factor structure by half. In this way, it is planned to develop the short form of the Instructor Self-Disclosure Scale with 9 items and a unidimensional scale, and the other scales with a total of 9 items, 3 items in each factor. Confirmatory Factor Analysis (CFA) was repeated on 9 items determined by each scale short form method.

Three different conditions have been manipulated for the purpose of the simulation study based on Raborn et al., (2020). These features are sample size, model misspecification, and correlation between factors.

- *Sample Size*: In CFA, it is stated that a sample size of at least 200 is required for accurate model estimates (Gatignon, 2010; Singh et al., 2016). Considering other similar simulation studies (French & Finch, 2011; Yang & Liang, 2013), two different sample sizes were determined as 200 and 500. Only findings based on sample size are included in the unidimensional structure.

- *Correlation between factors*: This condition were identified to be 0.00, 0.25 and 0.50 in a study by Batley & Boss (1993), 0.20, 0.50 and 0.70 in a study by Jiang et al. (2016), 0.10, 0.40 and 0.70 in a study by Van Abswoude et al. (2004a) and 0.00, 0.20, 0.40, 0.60, 0.80 and 1.00 in a study by Van Abswoude et al. (2004b). In this study, based on the correlations determined by these studies, two different correlation values were selected: 0.30 and 0.70 only for multidimensional factor structure. There is no correlation between factors in a unidimensional structure. In addition, in bifactor models, correlations between dimensions were not included as a simulation condition because “The bifactor model incorporates a general factor, onto which all items load directly, plus a series of orthogonal (i.e., specified as uncorrelated) factors each loading on a sub-set of items.” (Reise, 2012, p. 682).

- *Model Misspecification*: Model Misspecification was applied by ensuring that some of the observed variables were included in the factor that was not loaded. In the multidimensional and bifactor models, 6 items were selected, and it was ensured that these items were loaded in the factors where these items were not loaded in pairs. Three different models of misspecification have been selected: No Misspecification (0.00), 0.30, and 0.60 (Raborn et al., 2020). The

misspecified loadings (loading on the incorrect factor) were not the same as the loadings simulated to be real datasets. Misspecification was not applied to the model in unidimensional structure.

2.2. Data Simulation and Analysis

All data sets were simulated in R v4.0.4 (R Core Team, 2018) using the `simulateData` function on `lavaan` 0.6-8 package (Rosseel, 2012). All factor structures fitted on `lavaan` 0.6-8 package (Rosseel, 2012). First, population models were created with the findings of real data in the production of data belonging to unidimensional, multidimensional, and bifactor structures. These models are then calibrated to the null model. This process is repeated for each simulation condition. 100 iterations have been used per each simulation condition.

The scale short-form selection with ACO and TS was implemented with the `ShortForm` 0.4.6 package (Raborn & Leite, 2018). This package uses the `lavaan` package (Rosseel et al., 2012) to fit unidimensional, multidimensional, and bifactor CFA analysis. Based on previous research (Marcoulides & Falk, 2018; Raborn et al., 2020), some tuning parameters were used for each meta-heuristic algorithm. For ACO, 20 consecutive steps for convergence, 0.9 evaporation, 20 ants, and 50 maximum steps for no improvement were tuned. For TS, 5 tabu sizes for each condition and 50 iterations were specified. Since the iterations made with the `ShortForm` package were very slow, the number of iterations was limited (Raborn et al., 2020).

SCOFA analysis was implemented with `lavaan` 0.6-8 package (Rosseel, 2012). SCOFA algorithm, which iteratively deletes the item with the lowest factor loading from the item pool, is a standard scale short-form development procedure (Krueger et al., 2013). “After estimating a CFA for the original factor structure, the item with the lowest factor loading is removed. The model is then re-estimated with the reduced item set and again the item with the lowest factor loading is removed. This procedure is repeated until the predetermined number of items for the short version is reached.” (Schroeders et al., 2016, p. 8). Weighted Least Squares Mean and Variance adjusted (WLSMV) estimator are used for parameter estimation.

Model-fit was checked using several fit indices, including the Comparative Fit Index (CFI), the Tucker–Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA). CFI and TLI values above 0.90 and 0.95 reflect acceptable and excellent fit, respectively, while RMSEA below or near 0.05 indicates an acceptable fit of data to a model (Hu & Bentler 1999, pp. 24-26).

Omega coefficients were computed from `semTools` package 0.5-4 (Jorgensen et al., 2014). For all factor structures, omega coefficients as composite reliability is computed. Omega hierarchical (ω_H) and omega hierarchical subscale (ω_{HS}) are computed for bifactor structures. Omega hierarchical subscales are computed for multidimensional and omega total coefficients are computed for unidimensional structures.

As the findings of this simulation study, model-data fit and reliability coefficients were reported for each factor structure in each simulation condition.

3. RESULT

3.1. Findings from Unidimensional Factor Structure

In the unidimensional factor structure, findings were reported only according to the changes in the sample size according to the simulation conditions.

Table 1. Model-data fits and omega coefficients from the unidimensional factor structure.

SS	ACO				TS				SCOFA			
	CFI	TLI	RMSEA	RF-g	CFI	TLI	RMSEA	RF-g	CFI	TLI	RMSEA	RF-g
200	.980	.985	.034	.830	.985	.989	.024	.778	.980	.985	.034	.830
500	.998	.998	.011	.812	.972	.979	.036	.794	.999	.999	.009	.812

SS: Sample Size, RF-g: Reliability of general factor

In [Table 1](#), it has been observed that all model-data fits obtained according to the unidimensional factor structure indicate a good fit. When the sample size is 200 TS algorithm, it is seen that the SCOFA method produces the best results when it is 500. When the sample size is 200, the model-data fit and Omega coefficients obtained from ACO and SCOFA techniques are the same. In case the sample size is 500, the model-data fit and Omega coefficients obtained from ACO and SCOFA techniques are very close to each other. The Omega coefficients produced by ACO and SCOFA methods for both sample sizes are the same and higher than the coefficient produced by the short form obtained with TS. As the sample size increases, the model data fit generally increases, while the Omega coefficients decreases except for TS.

3.2. Findings from Multidimensional Factor Structure

In the multidimensional factor structure, findings were reported for all simulation conditions.

Table 2. Model-data fits from the multidimensional factor structure.

MS	SS	CBF	ACO			TS			SCOFA		
			CFI	TLI	RMSEA	CFI	TLI	RMSEA	CFI	TLI	RMSEA
None (0.0)	200	0.3	.998	.998	.006	.994	.996	.010	.991	.994	.013
		0.7	.994	.996	.012	.994	.996	.012	.974	.983	.027
	500	0.3	.984	.989	.017	.983	.989	.015	.972	.981	.022
		0.7	.990	.993	.015	.984	.989	.017	.994	.996	.012
Minor (0.3)	200	0.3	.995	.997	.009	.998	.998	.007	.924	.950	.045
		0.7	.915	.943	.053	.922	.948	.049	.856	.904	.078
	500	0.3	.913	.942	.040	.895	.930	.042	.902	.935	.049
		0.7	.979	.986	.026	.963	.975	.032	.984	.990	.024
Major (0.6)	200	0.3	.936	.957	.054	.892	.928	.057	.934	.956	.054
		0.7	.943	.962	.060	.789	.859	.079	.949	.966	.063
	500	0.3	.937	.958	.058	.890	.926	.051	.937	.958	.058
		0.7	.974	.982	.037	.952	.968	.039	.990	.994	.026

MS: Misspecification, SS: Sample Size, CBF: Correlation Between Factors

According to the findings obtained from the multidimensional factor structure, in [Table 2](#), all model-data fit values of the situation where there was no misspecification showed a good fit. When the sample size was 200 and the correlation between factors was 0.3, the model-data fit values were the highest. Although all three scale short-form development methods produced similar findings, it can be said that the findings of ACO and TS were better. In the case of minor misspecification, when the sample size was 200 and the correlation between factors was 0.3, the sample size was 500 and the correlation between factors was 0.7, with high model-data fits. Some minor and major misspecification conditions findings were obtained with TS and SCOFA, it is observed that sufficient model-data fit was not achieved. In case the major misspecification, when only the sample size was 500 and the correlation between factors was 0.7, all scale short-form development methods showed sufficient model-data fit. Findings

obtained from ACO and SCOFA showed sufficient model-data fit and were similar, specifically in no misspecification conditions; it is seen that the TS algorithm can generally obtain values far from adequate model-data fit.

Table 3. Omega coefficients from the multidimensional factor structure.

MS	SS	CBF	ACO			TS			SCOFA		
			RF-1	RF-2	RF-3	RF-1	RF-2	RF-3	RF-1	RF-2	RF-3
None (0.0)	200	0.3	.583	.606	.664	.437	.575	.585	.586	.595	.635
		0.7	.679	.577	.698	.575	.431	.421	.681	.735	.696
	500	0.3	.639	.648	.665	.552	.492	.498	.737	.758	.458
		0.7	.662	.577	.604	.576	.507	.501	.665	.716	.697
Minor (0.3)	200	0.3	.483	.563	.466	.536	.561	.410	.536	.563	.529
		0.7	.487	.490	.563	.487	.490	.563	.526	.456	.563
	500	0.3	.485	.541	.545	.464	.542	.422	.485	.507	.544
		0.7	.462	.561	.511	.389	.457	.487	.473	.562	.512
Major (0.6)	200	0.3	.426	.529	.392	.550	.440	.600	.559	.528	.615
		0.7	.399	.653	.336	.425	.532	.589	.449	.643	.494
	500	0.3	.508	.438	.525	.511	.409	.519	.504	.592	.525
		0.7	.475	.612	.527	.472	.476	.548	.474	.640	.546

MS: Misspecification, SS: Sample Size, CBF: Correlation Between Factors, RF-1: Reliability of First Factor, RF-2: Reliability of Second Factor, RF-3: Reliability of Third Factor

According to the findings obtained from the multidimensional factor structure, in Table 3, there was a decrease in the Omega coefficient of none to major misspecification, generally. The better Omega coefficient values were the case where the sample size was 500 and the correlation between factors was 0.7 in the no misspecification compared with other short-form techniques. It can be said that the Omega coefficients of ACO and SCOFA are similar. Especially in the no misspecification, the Omega coefficients obtained by the TS algorithm are lower than the other algorithms.

3.3. Findings from Bifactor Factor Structure

In the bifactor factor structure, findings were reported according to the changes in the sample size and the correlation between factors according to the simulation conditions.

Table 4. Model-data fits from the bifactor factor structure.

MS	SS	ACO			TS			SCOFA		
		CFI	TLI	RMSEA	CFI	TLI	RMSEA	CFI	TLI	RMSEA
None (0.0)	200	.983	.992	.028	.981	.990	.027	.981	.990	.030
	500	.994	.997	.016	.991	.996	.020	.976	.988	.033
Minor (0.3)	200	.991	.996	.021	.969	.987	.039	.969	.987	.039
	500	.994	.998	.018	.897	.945	.068	.828	.908	.093
Major (0.6)	200	.881	.941	.072	.869	.934	.077	.892	.946	.076
	500	.914	.957	.068	.918	.959	.057	.955	.977	.045

MS: Misspecification, SS: Sample Size

In Table 4, according to the findings obtained under the bifactor structure, all model-data fits obtained in the no misspecification showed a good fit. However, in the case of minor misspecification, the model-data fit in the case where the sample size obtained from TS and SCOFA algorithms is 500 was not sufficient. Findings obtained with only ACO showed a good

fit. In case of major misspecification and the sample size of 200, no scale short-form development method could show enough model-data fit. In the case of the sample size of 500, although the whole scale short-form development method showed sufficient model-data fit, the better findings were obtained with SCOFA compared with other short-form techniques.

Table 5. *Omega coefficients from the bifactor factor structure.*

MS	SS	ACO				TS				SCOFA			
		RF-1	RF-2	RF-3	RF-g	RF-1	RF-2	RF-3	RF-g	RF-1	RF-2	RF-3	RF-g
None (0.0)	200	.528	.160	.368	.763	.426	.483	.410	.844	.502	.176	.383	.760
	500	.327	.152	.186	.772	.396	.155	.176	.774	.428	.104	.229	.770
Minor (0.3)	200	.444	.507	.437	.403	.485	.284	.378	.713	.424	.456	.537	.344
	500	.365	.447	.543	.413	.684	.202	.518	.675	.336	.419	.308	.560
Major (0.6)	200	.213	.219	.351	.467	.211	.465	.464	.585	.183	.237	.217	.506
	500	.019	.170	.215	.417	.214	.384	.153	.534	.285	.174	.273	.490

MS: Misspecification, SS: Sample Size, RF-1: Reliability of First Factor, RF-2: Reliability of Second Factor, RF-3: Reliability of Third Factor, RF-g: Reliability of general factor

In **Table 5**, according to the findings obtained from the bifactor factor structure, there is a decrease in the Omega coefficient of none to major misspecification. The best Omega coefficients were obtained when there was no misspecification and the sample size was 200. Although the Omega values obtained from the scale short-form development methods are similar, it can be said that the Omega values obtained with the TS algorithm are higher.

4. DISCUSSION and CONCLUSION

It has been determined that all the methods of developing the short form of the scale in the unidimensional factor structure can select the short form with sufficient psychometric properties. Although it is not possible to mention a significant difference between the methods, the Omega coefficients produced by ACO and SCOFA methods are the same and higher than the coefficient produced by the short form obtained with TS. As the sample size increased, the model-data fit generally increased, while a slight decrease was observed in the Omega coefficients. In this case, for scales with unidimensional factor structure, all of the scale short-form development methods used in this study can be recommended.

In the multidimensional factor structure, all model-data fit values for the situation where there is no misspecification shows a good fit. In the case of minor misspecification, it was determined that some findings obtained by TS and SCOFA did not provide sufficient model-data fit. In the case of major misspecification, the findings obtained from ACO and SCOFA show sufficient model-data fit and are similar; it can be said that the TS algorithm can generally obtain values far from adequate model-data fit. It has been determined that the Omega coefficients of ACO and SCOFA are similar. Especially in the absence of misspecification, the Omega coefficients obtained by the TS algorithm are lower than the other algorithms. It is possible to say that the Omega coefficients increased with the increase in the sample size. In this case, for scales with multidimensional factor structure, ACO and SCOFA, which are among the scale short form development methods used in this study, can be recommended.

According to the findings obtained under the bifactor structure, all model-data fits obtained under no misspecification condition show a good fit. In the case of minor misspecification, TS and SCOFA algorithms could not show sufficient model-data fit, but in the case of major misspecification, no scale short form development technique could show sufficient model-data fit. In this case, it is recommended to use ACO, one of the short form development methods used in this study, for scales with a bifactor factor structure.

When viewed from the framework of metaheuristic algorithms, it has been determined that ACO produces better findings than TS. This finding is similar to the study of Raborn et al. (2020). In Raborn et al.'s (2020) study, the simulated annealing (SA) technique showed better performance in terms of fit indices and reliability indices. Next comes the ACO. Since the SA technique was not included in this study, it is possible to say that the findings of both studies are similar.

According to the findings of this study, the increase in the level of misspecification and the decrease in the sample size can significantly affect the model-data fit. In a situation where the factor structure of the scale is getting more and more complex, model-data fit and Omega coefficients decrease. Especially in cases where the factor structure is complex and the sample size is relatively low, it may be recommended to apply multiple scales short-form development methods and to continue studies on the methods that produce the best results.

This study was not carried out on two samples as suggested in the scale short form development studies. Using the required first sample as a "training sample" and choosing the item for the short form with this sample; the second sample should be used as a "testing sample" and the validity of the short form should be ensured with this sample. With such an application, it is ensured that the new short-form is validated in a new sample (Raborn et al., 2020). It is recommended to use these techniques and similar ones.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Hakan Koğar  <https://orcid.org/0000-0001-5749-9824>

REFERENCES

- Anastasi, A. (1982). *Psychological Testing* (5th ed.). Macmillan.
- Batley, R.M., & Boss, M.W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-restricted trait in a multidimensional item response model. *Applied Psychological Measurement*, 17(2), 131-141. <https://doi.org/10.1177/014662169301700203>
- Cayanus, J.L., & Martin, M.M. (2004). An instructor self-disclosure scale. *Communication Research Reports*, 21(3), 252-263. <https://doi.org/10.1080/08824090409359987>
- Colomi, A., Dorigo, M., & Maniezzo, V. (1991). *Distributed optimization by ant colonies*. In: Varela, F. and Bourgine, P., Eds., Proceedings of the European Conference on Artificial Life, ECAL'91, Paris, Elsevier Publishing, Amsterdam, 134-142.
- Ebesutani, C., McLeish, A.C., Luberto, C.M., Young, J., & Maack, D.J. (2014). A bifactor model of anxiety sensitivity: Analysis of the Anxiety Sensitivity Index-3. *Journal of Psychopathology and Behavioral Assessment*, 36(3), 452-464. <https://doi.org/10.1007/s10862-013-9400-3>
- French, B.F., & Finch, W.H. (2011). Model misspecification and invariance testing using confirmatory factor analytic procedures. *The Journal of Experimental Education*, 79(4), 404-428. <https://doi.org/10.1080/00220973.2010.517811>
- Gatignon, H. (2010). *Confirmatory Factor Analysis*. In *Statistical Analysis of Management Data* (pp. 59-122). Springer. https://doi.org/10.1007/978-1-4419-1270-1_4
- Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>

- Janssen, A.B., Schultze, M., & Grötsch, A. (2017). Following the ants: Development of short scales for proactive personality and supervisor support by Ant Colony Optimization. *European Journal of Psychological Assessment*, 33(6), 409. <https://doi.org/10.1027/1015-5759/a000299>
- Jiang, S., Wang, C., & Weiss, D.J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, 7(Article:109), 1-10. <https://doi.org/10.3389/fpsyg.2016.00109>
- Jorgensen, T.D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., ..., & Enders, C. (2016). *semTools: Useful Tools for Structural Equation Modeling. R package version 0.5-4*. Retrieved from <https://cran.r-project.org/web/packages/semTools/index.html>
- Kleka, P., & Soroko, E. (2018). How to avoid the sins of questionnaires abridgement?. *Survey Research Methods*, 12(2), 147-160. <https://doi.org/10.31234/osf.io/8jg9u>
- Kruyen, P.M., Emons, W.H., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, 13(3), 223-248. <https://doi.org/10.1080/15305058.2012.703734>
- LaNoue, M., Harvey, A., Mautner, D., Ku, B., & Scott, K. (2015). Confirmatory factor analysis and invariance testing between Blacks and Whites of the Multidimensional Health Locus of Control scale. *Health Psychology Open*, 2(2), 1-16. <https://doi.org/10.1177/2055102915615045>
- Leite, W.L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an Ant Colony Optimization Algorithm. *Multivariate Behavioral Research*, 43, 411–431. <https://doi.org/10.1080/00273170802285743>
- Marcoulides, K.M., & Falk, C. (2018). Model specification searches in structural equation modeling with R. *Structural Equation Modeling*, 25(3), 484-491. <https://doi.org/10.1080/10705511.2017.1409074>
- Nunnally, J.C. (1978). *Psychometric Theory* (2nd ed.). McGraw-Hill.
- Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale big-five assessments. *Journal of Research in Personality*, 59, 56-68. <https://doi.org/10.1016/j.jrp.2015.09.001>
- Raborn, A.W., & Leite, W.L. (2018). ShortForm: An R package to select scale short forms with the ant colony optimization algorithm. *Applied psychological measurement*, 42(6), 516. <https://doi.org/10.1177/0146621617752993>
- Raborn, A.W., Leite, W.L., & Marcoulides, K.M. (2020). A comparison of metaheuristic optimization algorithms for scale short-form development. *Educational and Psychological Measurement*, 80(5), 910-931. <https://doi.org/10.1177/0013164420906600>
- Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behav. Res.* 47, 667–696. <https://doi.org/10.1080/00273171.2012.715555>
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5-12 (BETA). *Journal of Statistical Software*, 48(2), 1-36. <https://doi.org/10.18637/jss.v048.i02>
- Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLoS One*, 11(11), 1-19. <https://doi.org/10.1371/journal.pone.0167110>
- Singh, K., Junnarkar, M., & Kaur, J. (2016). *Measures of Positive Psychology: Development and Validation*. Springer.
- Van Abswoude, A.A., van der Ark, L.A., & Sijtsma, K. (2004b). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, 28(1), 3-24. <https://doi.org/10.1177/0146621603259277>

- Van Abswoude, A.A., Vermunt, J.K., Hemker, B.T., & van der Ark, L.A. (2004a). Mokken scale analysis using hierarchical clustering procedures. *Applied Psychological Measurement*, 28(5), 332-354. <https://doi.org/10.1177/0146621604265510>
- Yang, Y., & Liang, X. (2013). Confirmatory factor analysis under violations of distributional and structural assumptions. *International Journal of Quantitative Research in Education*, 1(1), 61-84. <https://doi.org/10.1504/ijqre.2013.055642>

Adaptation and psychometric evaluation of the COVID-19 stress scales in Turkish sample

Murat Dogan Sahin^{1,*}, Sedat Sen², Deniz Guler³

¹Anadolu University, Faculty of Education, Department of Educational Sciences, Eskişehir, Türkiye

²Harran University, Faculty of Education, Department of Educational Sciences, Şanlıurfa, Türkiye

³Anadolu University, Faculty of Education, Department of Educational Sciences, Eskişehir, Türkiye

ARTICLE HISTORY

Received: Feb. 03, 2022

Revised: June. 12, 2022

Accepted: June 22, 2022

Keywords:

Covid related stress,
Scale adaptation,
Confirmatory factor
analysis,
Rasch analysis.

Abstract: This study aimed to adapt the COVID-19 Stress Scales (CSS) into Turkish and provide evidence for construct validity. For this purpose, firstly, Confirmatory factor analysis (CFA) was applied for the 5-factor model obtained during the development of CSS and the theoretically expected 6-factor model with total of 546 respondents. The findings revealed that the 6-factor model of CSS had a better fit in the Turkish sample. Factor loadings varied between .62 - .95 and correlations between subscales were between .44 - .76. Cronbach's Alpha and McDonald's ω coefficients for each subscale indicated good-to-excellent internal consistency. To evaluate the criterion-related validity, the Turkish version of The Fear of COVID-19 Scale (FCV-19S) was administered to the participants and the correlation coefficients between this scale and the six subscale of CSS were calculated. We also conducted the Rasch analysis with related items to provide psychometric evidence for their unidimensional structure of each of the six subscales. Lastly, Differential item functioning (DIF) analysis was performed across subgroups by gender, having COVID-19, and being a student. Overall, the results of both CFA and Rasch analyses provided evidence to support the substantive aspect of validity and the appropriateness of the CSS as a measure of COVID-19 stress level in a Turkish sample.

1. INTRODUCTION

World Health Organization (WHO, 2020) announced COVID-19 as a novel coronavirus disease outbreak of international importance in January 2020, shortly after the first case in Wuhan, China, and subsequently declared it a pandemic in March 2020. With this announcement, WHO (2020) made a statement based on the experiences from the previous pandemics, emphasizing that individuals may suffer from stress and mental health problems.

The COVID-19 pandemic has been affecting individuals' mental health since its beginning. While the virus has been spreading worldwide, people's fear of contracting the disease and dying has increased (Valiente et al., 2021). However, getting the disease and dying have not been the only sources of people's fear and anxiety during the pandemic. Various stressors that may affect individuals economically and socially have emerged, including economic problems

*CONTACT: Murat Doğan ŞAHİN ✉ muratdogansahin@gmail.com 📧 Anadolu University, Faculty of Education, Department of Educational Sciences, Eskişehir, Türkiye

due to loss of employment or changes in work life, loved ones' health problems and the possibility of losing them, being socially labeled, and experiencing social exclusion due to catching the disease, restriction of freedom, and staying away from loved ones due to being kept in quarantine or social isolation. The collapse of the health system and scarcity of food and other necessities are among the stress sources of the COVID-19 pandemic that may have long-term effects (Liu et al., 2021; Mertens et al., 2020; Porcelli, 2020; Wang et al., 2021).

Theoretical and empirical studies have confirmed that stress and anxiety due to COVID-19 have a strong association with emotion and behavior problems, such as dis-functionality, depression, health anxiety (Gallagher et al., 2020; Mertens et al., 2020); generalized anxiety disorder and death anxiety (Lee et al., 2020); obsessive-compulsive disorder (Seçer & Ulaş, 2020); post-traumatic stress disorder (Boyras & Legros, 2020); eating disorder (Baenas et al., 2020); panic buying and coping challenges (Taylor et al., 2020a) and sleeplessness and acute stress (Wang et al., 2021). The studies also revealed that COVID-19 related anxiety and stress have a more severe effect on those who already had a mental problem before the pandemic. For example, Asmundson et al. (2020) showed that individuals who had anxiety or mood disorders showed the symptoms of psychological distress, xenophobia, and traumatic stress to a greater extent compared to those who did not have anxiety or mood disorders. This study points out that the stress and anxiety associated with COVID-19 may increase the vulnerability of individuals to various mental health disorders and behavioral problems.

Research has demonstrated a high prevalence of mental health problems worldwide due to COVID-19. For example, a study with about 53,000 participants in China indicated that 35% of the participants experienced psychological distress (Qiu et al., 2020). In Italy, 27% of the participants had high stress levels, 32% had high depression levels, and 18% had high anxiety levels (Mazza et al., 2020). A study in Spain reported that 22.1% of the participants had depression, and 19.6% had anxiety. Later studies reported increases in these percentages compared to former studies (Valiente et al., 2021).

Furthermore, a meta-analysis of twelve large-scale studies showed that the pooled prevalence of depression was 25% during the COVID-19 outbreak (Bueno-Notivol et al., 2021). Finally, another study revealed that 16% of the participants had COVID-19 related stress syndrome at a level requiring mental health services (Taylor et al., 2020a). These results demonstrated the severity of the negative effects of COVID-19 on mental health.

Both mental health professionals and World's governments have essential duties to prevent or minimize the mental health and behavioral problems due to COVID-19 related stress and anxiety (Wang et al., 2020). Therefore, the first step is to examine people's COVID-19 related stress and anxiety levels and evaluate their change over time at specific sessions. In addition, when the COVID-19 outbreak will end is uncertain. Hence, people's hope that the COVID-19 outbreak will end is gradually decreasing while their anxiety and stress levels are increasing (Bernardo & Mendoza, 2020). Moreover, it is predicted that the effects of economic and social problems caused by the outbreak will continue for a long time, even after the pandemic (Gavin et al., 2020; Valiente et al., 2021). Hence, COVID-19 will certainly threaten individuals' physical and mental health for a while. In this sense, while the outbreak continues to develop and progress, the potential effects of factors specific to COVID-19 on mental health should be constantly monitored (Gallagher et al., 2020). As a result, we need valid and reliable measurement instruments to assess COVID-19 related stress and anxiety.

Various instruments assessing COVID-19 related stress and anxiety have been developed shortly after the emergence of the outbreak. One of these was The Fear of COVID-19 Scale (FCV-19S), developed by Ahorsu et al. (2020). The scale has seven items and a single factor. The FCV-19S was originally developed in Iran, and it was adapted to Turkish by Satici et al. (2020). Another scale was Coronavirus Anxiety Scale (CAS), containing five items measured

by a single factor. Developed by Lee (2020), the scale assesses individuals' dysfunctional anxiety levels, and two different research teams adapted it to Turkish (see, Biçer et al., 2020; Evren et al., 2020). Finally, COVID-19 Stress Scales (CSS) was developed by Taylor et al. (2020b) using Canadian and US samples to better understand COVID-19 related distress. CSS comprises 36 items and five factors. It was previously adapted for use with Persian (Khosravani et al., 2021) and Arabic (Abbady et al., 2021) populations.

Selecting the most appropriate instrument that researchers and mental health practitioners can use to identify the COVID-19 related stress and anxiety levels and support the aims of their studies is of particular importance in research. Pakpour et al. (2020) stated that FCV-19S is more advantageous when data need to be collected in a short time since it is a single factor instrument with few items. However, CSS would be more appropriate for assessing individuals' anxiety and stress levels more comprehensively. Taylor et al. (2020b) mentioned that they developed CSS because no other comprehensive instrument existed to measure COVID-19 related distress symptoms, such as fear of being infected, fear of contacting things and surfaces, xenophobia towards people who may have an infection, fear due to socio-economic results of the outbreak, traumatic stress symptoms, and compulsive checking and reassurance-seeking about possible threats related to the pandemic.

In conclusion, various scales are available to measure COVID-19 related stress and anxiety symptoms in Turkish culture; however, none of them comprehensively assess individuals' COVID-19 related stress and anxiety levels. Most of the previous scale validation studies on COVID-19 related stress scale validation studies have been conducted using only factor analytic methods (Abbady et al., 2021; Khosravani et al., 2021; Taylor et al., 2020b). Thus, the purpose of the current study was to address these two gaps in the literature by adapting the CSS (Taylor et al., 2020b) into the Turkish language (CSS-T) using confirmatory factor analysis (CFA) and Rasch analysis. Several other psychometric properties of the Turkish version of CSS were also examined. To this end, the present study examined the following research questions:

1. Does CSS-T yield sufficient validity evidence supported by Rasch model and CFA?
2. Do CSS-T items function differently for persons who have been tested positive for COVID-19?
3. Do CSS-T items function differently across gender groups?
4. Do CSS-T items function differently between students and non-students?

2. METHOD

2.1. Sample

A total of 546 adult volunteers participated in an online survey through a non-probability convenience sampling. The survey included demographic questions, questions about the participants' COVID-19 history, and CSS-T. After the Anadolu University ethics committee approved the project, the survey was disseminated using different platforms, such as cell phones, e-mails, and social networks (e.g., Twitter). The data were collected from participants located in different regions of Turkey in the same month, from March 16 to April 2, 2021. Based on the boxplot created to detect outliers, one participant was removed, and the analyses were conducted with 545 participants. Respondents were 18 to 64 years of age ($M = 30.1$ years, $SD = 9.8$). Most participants (69.9%) were women. About half of the participants were students (50.6 %).

Furthermore, 13.95% stated that they had previously tested positive for COVID-19. Twelve percent of the participants answered “no” to whether there is anyone other than themselves who has tested positive for COVID-19 in their family, relatives, friends, or family friends. Twenty-two percent stated that someone in their nuclear family had this disease. Sixty-six percent of

them stated that their close relatives, family friends, or friends had this disease. The descriptive information of the respondents is presented in [Table 1](#).

Table 1. *Demographics of the participants.*

Age (\bar{X})	Gender		Student Status		COVID-19 History		Infected People in Your Close Circle?	
	Female	Male	Student	Non-Student	Yes	No	Yes	No
30.1 (SD=9.8)	69.9%	30.1%	50.6%	49.4%	13.95%	86.05%	12%	88%

2.2. Instrument

2.2.1. Demographic Information

The demographic questionnaire contained items assessing the participants' age, gender, student status, and COVID-19 history. Items assessing COVID-19 history inquired about whether the participants themselves tested positive for COVID-19 disease. Subsequent questions asked whether someone in their nuclear family had COVID-19 and whether other relatives, family friends, or friends had this disease.

2.2.2. The COVID-19 Stress Scales (CSS)

The CSS is a self-report measure developed by Taylor et al. (2020b) to measure COVID-19 related stress symptoms and based on five factors associated with perceived threat and fear of disease. The first 24 items of the 36-item scale evaluate the extent to which individuals have experienced various kinds of worries over the last seven days on a 5-point scale ranging from “not at all” to “extremely.” Five of the remaining items assess the frequency with which they have experienced the situations described by the items in the last seven days, and seven items measure the frequency with which they have experienced the situations mentioned in the items in the last seven days. These 12 items were also rated on a 5-point scale ranging from “never” to “almost always.”

CSS theoretically consists of six domains, the dangerousness of COVID-19 (danger), fears about sources of COVID-19-related contamination (contamination), COVID-19-xenophobia (xenophobia), fears about the personal social and economic consequences of COVID-19 (socio-economic consequences), COVID-19-related checking (compulsive checking), and traumatic stress symptoms related to COVID-19 (traumatic stress), assessed with six items. On the other hand, a parallel analysis conducted through the same measurement tool revealed five factors in a Canadian sample revealed five factors (Taylor et al., 2020). Specifically, danger and contamination loaded on the same factor, while all other items worked theoretically as expected. The researchers preferred to preserve all 12 items underlying the two factors instead of reducing the number of items. The authors explained that keeping the number of items for each domain would allow subsequent studies to measure them separately (Taylor et al., 2020b). The researchers concluded that the theoretically expected 6-factor model should also be tested together with the 5-factor model in future studies.

In a follow-up study, the 5-factor structure was tested in the US sample (Taylor et al., 2020b). The fit statistics for the CFA were acceptable, RMSEA = .05, SRMR = .042, CFI = .93, and it was concluded that the model fit the data well. Internal consistency values, assessed by Cronbach's alpha, were above .80 for all subscales of the scale in both samples.

To determine the convergent validity, Taylor et al. (2020b) examined the correlations between the subscales of the CSS and pre-COVID health anxiety, obsessive-compulsive checking, and contamination symptoms. Moderate significant correlation coefficients supported the

convergent validity of the CSS. First, for discriminant validity, the correlations of five CSS subscales with the social desirability scale were all close to 0. On the other hand, correlations between most CSS subscales and current anxiety were significantly higher compared to correlations with current depression, supporting the discriminant validity of CSS.

To validate the Turkish version of the CSS, the existing translation that take place in the developers' web page (Psychology of Pandemics Network, n.d.) was examined first. However, since this translated version had many misconceptions, we decided to translate the CSS again with the permission of the scale developers. Two psychometricians and a psychological counselor who were the primary investigators of this study carried out the translation process. The researchers first translated the original version into Turkish independently, and subsequently, they came together and tried to reach an agreement on the different translated items. An English language expert was consulted for items that could not be agreed upon. This expert was given the original version of the articles, the researchers' translations, and the issues with which they disagreed. The translation was finalized according to this expert's opinions. A Turkish language expert also approved the resulting CSS-T.

2.3. Data Analysis

The analyses to determine the validity of the CSS-T in the Turkish sample were carried out in two parts. The CFA was conducted first to compare the fit indices for the theoretical 6-factor model of the scale and the 5-factor model obtained during the development phase. The goodness-of-fit indices were based on conventional guidelines introduced by Hu and Bentler (1998). We used Hu and Bentler's (1999) empirically derived cut-off values to interpret whether a given factor model fit the data well. Accordingly, $RMSEA < .06$, $SRMR < .08$ and $CFI \& TLI > .90$ were interpreted as good fit values. CFAs were performed with Mplus 7.0 (Muthen & Muthen, 2017). To establish the criterion-related validity, a one-dimensional Fear of COVID-19 Scale (FCV-19-S) developed by Ahorsu et al. (2020) and adapted into Turkish by Satici et al. (2020) was used. The correlations between the FCV-19-S and each subscale of CSS-T supported the CSS-T's criterion-related validity.

To determine internal consistency, Mc Donald's Omega was reported for each dimension along with Cronbach's Alpha. Stratified Alpha coefficient was also reported for the entire scale.

Another round of Rasch analyses was conducted to determine which items contribute to measurement of the scale dimensions identified through CFA. Analyses were conducted for each dimension separately. Construct unidimensionality was also checked for each dimension before proceeding to other Rasch-related analyses. A principal components analysis (PCA) of standardized residual correlations was conducted to determine if the extra variance was explained after the Rasch construct was extracted. The Rasch construct should account for at least 40% of the total variance, and the value of the first contrast in the "unexplained variance" (residual variance) should be less than or equal to 2.0 (Linacre, 2004).

We also evaluated the CSS-T using the following rating scale guidelines suggested by Linacre (1999, 2004).

- #1: At least ten frequencies should be observed for each category.
- #2: Observation distribution should be regular.
- #3: Average measures should advance monotonically with each category.
- #4: Outfit mean-squares should be less than 2.0.
- #5: Step calibrations should advance monotonically with each category.
- #6: Ratings should imply measures, and measures should imply ratings.
- #7: Step difficulties should advance by at least 1.4 logits and by less than 5.0 logits.

WINSTEPS Version 3.68.2 software (Linacre, 2009) was used to analyze Likert-type responses by calibrating a Rasch Rating Scale model (Andrich, 1978). The response distribution of the 36 items across six agreement options and their association with overall item variance (i.e., point-biserial correlation [PTMEA]) were analyzed. PTMEA correlations should all be positive and higher than .50. In addition to response category frequencies and distributions, the Rasch model was used to estimate item location parameters, step parameters, average measures, and fit statistics to evaluate the first five criteria of Linacre's guideline. Two item fit values (outfit and infit mean square statistics) were examined to determine which items should be flagged for revision due to misfit between items and the Rasch model (Bond & Fox, 2007; Sick, 2010). The acceptable range for infit (information-weighted fit) or outfit (outlier-sensitive fit) mean square statistics included values between 0.5 and 1.5 (Linacre, 2004; Sick, 2010). The values within this range indicate the fit between the item and model. Values below 0.5 may indicate less productive items, and items with values greater than 1.5 may indicate unproductive measurement construction. Person and item reliability indices along with separation indices were computed to determine the internal consistency of ratings. The "person reliability" obtained from WINSTEPS is equivalent to the traditional "test" reliability. Low values of the person and item reliability statistics may indicate a narrow range of person and item measures, respectively. Reliability values greater than .80 and separation indices greater than 2.0 are considered adequate (Crocker & Algina, 1986).

An item/person map (aka Wright map) was also created for each dimension to examine the item severity. This map shows the relationships between respondents' abilities (on the left side) and item difficulties (on the right side) on a linear scale in a unit logit to help us see whether the item difficulties were appropriate for the targeted respondents. Finally, differential item functioning (DIF) analyses were employed to examine the functioning of items across subgroups, including respondents' gender, student status, and COVID-19 history (having COVID-19 or not). When assessing DIF, two values (DIF contrast and p -value) were used to assess whether an item can be flagged as showing significant DIF. DIF contrast is calculated by taking the difference in item locations (item difficulty) between subgroups. Values greater than 0.5 logits may indicate a DIF situation (Linacre, 2006; Bond and Fox, 2015). The Rasch-Welch and the Mantel-Haenzel tests that are available in WINSTEPS software can be used to obtain a p -value for DIF analysis. Due to many items being compared, alpha (set at .05) was controlled when making comparisons using a Bonferroni correction. Therefore, p -values had to be less than .008 (i.e., .05 divided by six items) with a contrast greater than 0.5 logits to show evidence of DIF.

3. RESULT / FINDINGS

3.1. CFA and Criterion Related Validity Results

To evaluate the construct validity of CSS-T, the fit values of the theoretically predicted 6-factor model and the fit values of the 5-factor model obtained during the development process were compared. Since maximum likelihood (ML) requires multivariate normality assumption to be met, robust-maximum likelihood (MLR) was used as the estimator. The results obtained are shown in [Table 2](#).

Table 2. Fit indices obtained for five- and 6-factor model.

Model Solution	Modification Status	RMSEA	CFI	TLI	SRMR
5 Factors	Before Modification	.082 (.078 -.085)	.829	.816	.064
	After Modification	.063 (.060 .067)	.898	.889	.057
6 Factors	Before Modification	.072 (.069 .075)	.869	.857	.058
	After Modification	.056 (.053 .059)	.921	.913	.047

*Applied modifications: 3-4; 7-8; 22-23

As shown in [Table 2](#), the values obtained for the 6-factor model were slightly better, but the fit indices obtained for both structures were not acceptable. Therefore, three theoretically valid modifications based on correlating the error terms that significantly reduced the chi-square value in both models were conducted. These items, which are very similar in expression, are indicated at the bottom of the table. Item 3 and Item 4 assess worries about the healthcare system (Item 3: “I am worried that our healthcare system won’t be able to protect my loved ones;” Item 4: “I am worried that our healthcare system is unable to keep me safe from the virus.”). Item 7 and Item 8 assess worries about grocery stores (Item 7: “I am worried about grocery stores running out of food;” Item 8: “I am worried that grocery stores will close down.”). Finally, Item 22 and Item 23 assess monetary transactions (Item 22: “I am worried about taking change in cash transactions;” Item 23: “I am worried that I might catch the virus from handling money or using a debit machine.”). While only SRMR indicated that the 5-factor model had a good fit after modification, all fit indices obtained for the 6-factor model had good fit values when the same modifications were applied. These findings confirmed the theoretical 6-factor model of CSS-T in the Turkish sample. The results for the 6-factor model are given in [Appendices](#). Factor loadings are between .62 - .95 (see [Table 1A](#) in [Appendices](#)), and correlations between subscales are between .44 - .76 (see [Table 2A](#) in [Appendices](#)).

Cronbach’s Alpha and McDonald’s ω coefficients for each CSS-T subscale are shown in [Table 3](#). As shown, all values are greater than .80, indicating good-to-excellent internal consistency (Tavakol & Dennick, 2011). Stratified Alpha calculated for the whole scale is .97.

Table 3. Internal consistency measures of CSS-T.

Subscale	Cronbach’s Alpha	Mc Donald’s ω
COVID Danger (D)	.87	.85
COVID Socio-Economic Consequences (SEC)	.91	.91
COVID Xenophobia (X)	.93	.93
COVID Contamination (C)	.93	.93
COVID Traumatic Stress (TS)	.92	.92
COVID Compulsive Checking (CH)	.89	.89

To evaluate the criterion-related validity of CSS-T, the Turkish version of unidimensional FCV-19-S was administered to the participants. The correlation coefficients between this scale and the six subscale of CSS-T were calculated. Accordingly, five of the CSS-T subscales showed a moderate and statistically significant relationship with FCV-19-S and a relatively stronger relationship between Traumatic Stress. The obtained results support CSS-T's criterion-related validity.

3.2. Rasch Analysis

We also conducted Rasch analysis for each subscale's items to provide additional psychometric evidence for unidimensional structures. The Rasch analysis enables us to obtain different information that cannot be obtained with CFA. The Rasch analysis process included the evaluation of the rating scale functioning analysis, item fit, reliability, dimensionality, and differential item functioning analysis.

Table 4. Dimensionality results.

	D	SEC	X	C	TS	CH
Raw variance explained by measures	11.9 (66.4%)	7.5 (55.4%)	15.3 (71.9%)	13.3 (68.8%)	10.2 (63.1%)	7.2 (54.7%)
Raw variance explained by persons	7.3 (40.6%)	5.4 (40.0%)	10.6 (49.8%)	9.5 (49.5%)	5.5 (33.8%)	3.7 (28.0%)
Raw Variance explained by items	4.6 (25.8%)	2.1 (15.4%)	4.7 (22.1%)	3.7 (19.3%)	4.8 (29.3%)	3.5 (26.7%)
Raw unexplained variance (total)	6.0 (33.6%)	6.0 (44.6%)	6.0 (28.1%)	6.0 (31.2%)	6.0 (36.9%)	6.0 (45.3%)
Unexplained variance in 1st contrast	1.5 (8.5%)	2.0 (14.7%)	1.8 (8.4%)	2.5 (13.1%)	1.5 (9.0%)	2.2 (16.8%)
Unexplained variance in 2nd contrast	1.4 (7.7%)	1.2 (9.1%)	1.2 (5.7%)	1.2 (6.1%)	1.4 (8.5%)	1.8 (13.7%)

Note. D = Danger, SEC = Socio-Economic Consequences, X = Xenophobia, C = Contamination, TS = Traumatic Stress, CH = Compulsive Checking

3.2.1. Dimensionality analysis

For the present investigation, the dimensionality of six individual CSS-T subscales was assessed first by employing a PCA of standardized residual correlations. Individual PCAs were performed to determine whether another dimension is present in the residuals after estimating the primary measurement dimension. The amount of variance explained by each extracted principal component was computed based on separate PCAs, which is presented in Table 4 for each subscale. In all instances, the Rasch construct explained more than 50% of the variance. As shown in Table 4, the variance explained by the six subscales of CSS ranged from 54.7% (Compulsive Checking) to 71.9% (Xenophobia), just meeting the recommended level. The variance explained by the persons ranged from 28.0% to 49.8%, and the variance explained by the items ranged from 15.4% to 29.3%. In all instances, the eigenvalues of the first contrast were between 1.5 and 2.5. Only two subscales had values (2.2 and 2.5) greater than 2.0. These are above the cut-off value (i.e., 2.0) but not substantially higher than 2.0. The unexplained variances in the first extracted component were higher than the recommended lower bound of 5% for all subscales and less than 15%, except for Compulsive Checking. Dimensionality analyses showed that all six subscales had >50% of the variance explained by the Rasch dimensions, and that first contrasts of four subscales had eigenvalues less than 2. Therefore, it is concluded that all subscales could be considered as unidimensional.

3.2.2. Reliability analysis

Consistency and spread of persons or items on the measured variable were evaluated with reliability and separation indices. These measures were used to examine the degree to which measures are reproducible. Two different reliability and separation indices were estimated for each subscale, as presented in [Table 5](#).

Table 5. Reliability and separation estimates.

Scale	Real Reliability	Model Reliability	Real Separation	Model Separation
D				
Persons	.85	.87	2.37	2.54
Items	.99	.99	8.22	8.79
SEC				
Persons	.70	.71	1.51	1.58
Items	.96	.96	5.11	5.17
X				
Persons	.84	.86	2.32	2.51
Items	.99	.99	12.32	12.77
C				
Persons	.88	.90	2.72	2.95
Items	.99	.99	8.38	8.68
TS				
Persons	.82	.85	2.14	2.34
Items	.99	.99	11.20	11.49
CH				
Persons	.81	.83	2.05	2.22
Items	.98	.98	7.64	7.79

Real reliability refers to reliability at its worst. Model reliability refers to reliability at its best. True reliability falls somewhere in between. As shown in [Table 5](#), model reliability estimates ranged from .71 to .99, while real reliability estimates varied between .70 and .99. Item reliability estimates were found to be higher than person reliability estimates (see [Table 5](#)). Socio-Economic Consequences appeared to have the smallest person and item reliability estimates. The item reliability values in [Table 5](#) indicate high internal consistency, while person reliability values indicate moderate consistency, except for Socio-Economic Consequences. As shown in [Table 5](#), separation estimates for persons ranged from 1.51 (Socio-Economic Consequences) to 2.72 (Contamination). Item separation estimates varied between 5.11 (Socio-Economic Consequences) and 12.32 (Xenophobia). Item separation indices were higher than person separation indices. Except for Socio-Economic Consequences, person separation estimates indicated a reasonable spread and the scale’s ability to separate persons into different levels of ability. According to Bond and Fox (2015), an instrument with separation estimates greater than 1.0 can be considered minimally useful. All item and person separation measures exceeded cut-off in this study, indicating a sufficient spread of items across subscales.

3.2.3. Rating scale category effectiveness

Rasch-based estimates were computed to determine whether the rating scales are functioning properly according to Rasch model assumptions. [Table 6](#) shows the rating scale’s effectiveness, including the frequency and percentage values for each rating scale category and the scale’s inferential values such as infit and outfit, mean-square fit statistics, structure calibration, and category measure.

Table 6. Rating scale effectiveness.

Category	Count	%	Infit MNSQ	Outfit MNSQ	Structure Calibration	Category Measure
D						
1	452	16	1.38	1.40	NONE	-3.03
2	595	21	0.76	0.77	-1.75	-1.29
3	660	23	0.89	0.94	-0.56	0.00
4	617	22	0.84	0.87	0.57	1.30
5	532	19	1.06	1.05	1.73	3.02
SEC						
1	704	33	1.03	0.98	NONE	-3.44
2	842	39	0.73	0.82	-2.26	-1.31
3	379	18	0.95	1.04	-0.15	0.23
4	148	7	0.97	1.07	0.89	1.39
5	75	3	1.49	1.80	1.51	2.90
X						
1	936	32	1.01	1.01	NONE	-3.62
2	787	27	0.81	0.80	-2.39	-1.68
3	628	22	0.89	0.99	-0.82	0.04
4	341	12	0.97	1.00	0.94	1.68
5	206	7	1.54	1.56	2.28	3.54
C						
1	191	7	1.24	1.26	NONE	-4.27
2	502	18	0.96	0.94	-3.10	-2.04
3	740	26	0.97	1.02	-0.88	-0.02
4	934	33	0.84	0.86	0.84	2.04
5	501	17	1.08	1.05	3.14	4.31
TS						
1	732	24	1.14	1.14	NONE	-2.69
2	625	20	0.87	0.84	-1.31	-1.19
3	658	22	0.95	0.91	-0.65	-0.11
4	664	22	0.84	0.86	0.23	1.14
5	381	12	1.16	1.11	1.73	2.98
CH						
1	452	16	1.38	1.40	NONE	-3.03
2	595	21	0.76	0.77	-1.75	-1.29
3	660	23	0.89	0.94	-0.56	0.00
4	617	22	0.84	0.87	0.57	1.30
5	532	19	1.06	1.05	1.73	3.02

Counts and percentages were investigated to determine the extent to which survey respondents utilized the various rating scale categories. Infit and outfit mean square (MNSQ) fit statistics are used to determine if any rating scale category is “noisy” or produces calibrations that are not desirable for a productive measurement value. Structure calibration shows the transition between categories and how difficult it is to observe each category. As shown in Table 6, most of the criteria proposed by Linacre (2002) appeared to hold for the current scale: (a) each response category had a frequency count greater than 10, (b) average measures by each rating

scale category advanced from smallest to largest, (c) most response categories (except for Xenophobia and Contamination) had outlier- outfit MNSQ values less than 2, (d) step calibrations (distance between ratings) increased monotonically, and (e) advance in step difficulties between step calibrations were at least one logits (for a five-category rating scale) and less than five logits. Based on the collective evidence, we can conclude that the scale is functioning properly.

3.2.4. Item fit

Item quality of each subscale was investigated with several item parameter estimates as presented in Table 7, including item difficulty calibrations, standard errors, fit statistics (infit and outfit), and point–measure correlations. As can be seen in Table 7, item difficulty calibrations ranged from -0.82 to 0.78 logits for Danger, from -0.82 to 0.78 logits for Socio-Economic Consequences, from -1.30 to 1.25 logits for Xenophobia, from -1.15 to 0.81 logits for Contamination, from -0.71 to 1.04 logits for Traumatic Stress and from -0.68 to 0.58 logits for Compulsive Checking. These ranges indicated a good amount of spread in the item locations, which is desirable for Rasch measurement scales to cover the full theoretical range of the construct’s continuum. Standard errors ranged in size from 0.05 to 0.08 . Concerning the present data, the estimated infit and outfit MNSQ values were within the acceptable range, ranging from 0.52 to 1.55 (Table 7). As shown in Table 7, only two items measuring Danger and Xenophobia (Items 13 and 30) were identified as misfitting (infit MNSQ values >1.5 , Wright & Linacre, 1994). The remaining 34 items fit the criteria for noise-free calibrations. All the infit MNSQ values were within the suggested guidelines, illustrating an acceptable fit to the Rasch RSM. Infit is a weighted index while the outfit is unweighted. Thus, large outfit values are generally considered less problematic than large infit values (Bond & Fox, 2007). As shown in Table 7, point–measure correlations ranged from $.68$ to $.89$. All point-measure correlations were positive and above the suggested $.3$ cut-off for all 36 items, supporting item-level polarity and unidimensionality of each subscale. All scenarios demonstrated good properties based on the criteria proposed by Wright and Linacre (1994).

Table 7. *Item fit statistics.*

Item #	Logit (δ)	SE	Infit MNSQ	Outfit MNSQ	PTMEA
CH					
1	-0.07	0.06	1.06	1.08	.74
2	-0.68	0.05	1.09	1.10	.71
3	-0.26	0.05	0.95	0.92	.74
4	-0.01	0.05	0.80	0.79	.78
5	0.58	0.05	1.01	1.03	.76
6	0.45	0.05	1.09	1.12	.74
SEC					
7	0.74	0.08	1.01	1.05	.74
8	0.30	0.08	0.98	0.98	.76
9	-0.13	0.07	0.86	0.93	.82
10	-0.22	0.08	0.98	1.01	.82
11	-0.24	0.07	1.07	1.13	.78
12	-0.45	0.07	1.07	1.11	.82
D					
13	-0.37	0.06	1.38	1.51	.77
14	0.47	0.06	1.14	1.16	.81
15	-0.07	0.06	0.60	0.59	.88

Table 7. *Continues.*

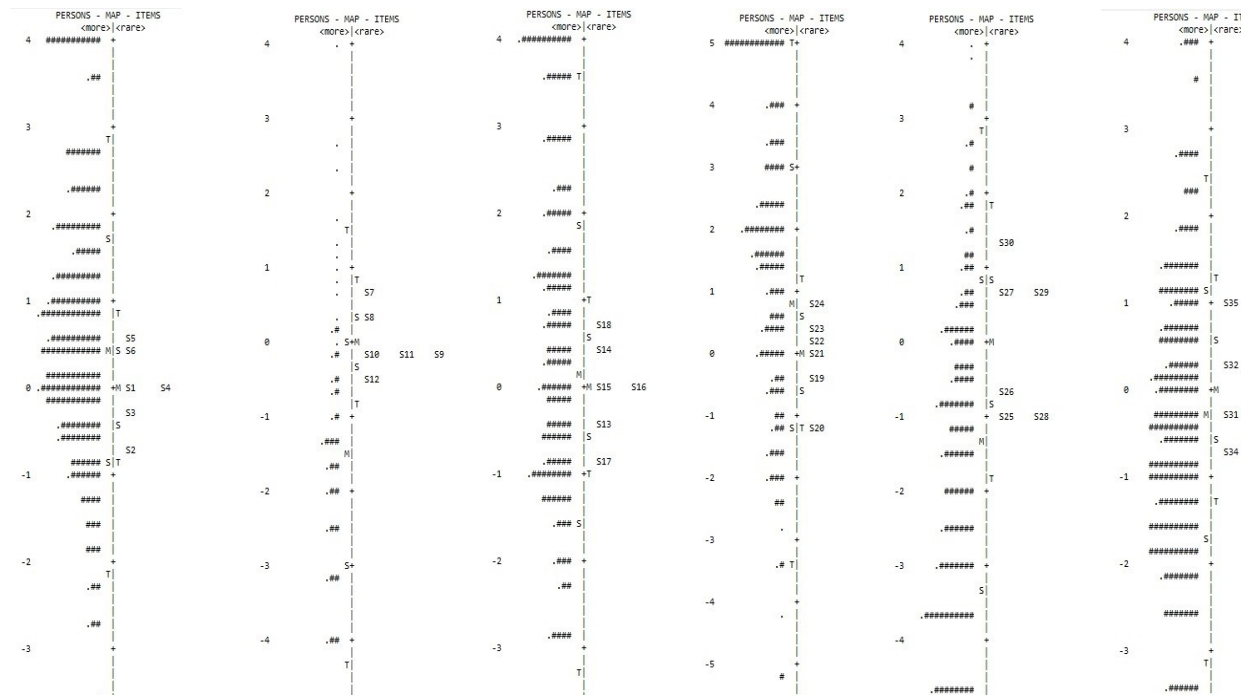
16	0.00	0.06	0.52	0.52	.89
17	-0.82	0.06	1.17	1.20	.80
18	0.78	0.06	1.17	1.11	.81
C					
19	-0.32	0.07	1.07	1.15	.83
20	-1.15	0.07	1.20	1.15	.81
21	0.03	0.07	1.14	1.16	.83
22	0.29	0.07	0.81	0.79	.87
23	0.34	0.07	0.77	0.76	.88
24	0.81	0.07	0.97	0.97	.87
X					
25	-1.30	0.07	0.97	0.96	.86
26	-0.60	0.07	0.97	0.96	.85
27	0.60	0.07	0.69	0.64	.83
28	-0.93	0.07	0.93	0.94	.86
29	0.70	0.07	1.12	1.45	.75
30	1.25	0.07	1.28	1.55	.68
TS					
31	-0.22	0.05	0.81	0.82	.81
32	0.28	0.05	1.03	1.03	.76
33	0.32	0.05	0.93	0.91	.79
34	-0.71	0.05	0.99	0.96	.78
35	1.04	0.06	1.17	1.05	.72
36	-0.71	0.05	1.11	1.11	.76

3.2.5. Results of person-item map

Figure 1 presents a person–item map in which information about the relation between item difficulty (easiest at the bottom to hardest at the top) and construct is shown on the right side. On the left side of the map are the person ability measures, showing the placement of respondents along the latent dimension. The map is centered at a score of 0 for the items, and because both sets of measures are on the same scale, meaningful comparisons can be made based on the map between items and persons. For person and item distributions, the mean (M) is provided in the center of the distribution with one (S) and two (T) standard deviations from the mean noted. The left side of the Wright map reports the distribution of measure scores for respondents, while the right side of the Wright map reports the calibrated scores for items of each subscale.

As shown in Figure 1, items are distributed mostly along the trait dimension, and the items appeared to be well-targeted to the sample. Most of the items align vertically across the logit scale. As shown in Table 7, Item 30 (item difficulty of 1.25) was the most difficult item for respondents to endorse, whereas Item 20 (item difficulty of -1.15) was the easiest to endorse (see Table 7). The mean of the person ability measures was close to the mean of the item difficulty measures for Subscales 1, 3, 4, and 6. As expected, the ability scores of each subscale were distributed normally, and the variance among the participants indicated a heterogeneous mix of responses.

Figure 1. Item/Person map.



3.2.6. Differential item functioning

The Rasch-Welch and the Mantel Haenzel tests, as well as the DIF contrast, were examined to detect DIF across subgroups by gender, having COVID-19 history and school enrollment status. Items that were problematic in terms of both p -values ($<.008$) and contrast (>0.5) were flagged, indicating bias between different subgroups. The p -value was nonsignificant, and DIF contrasts were less than 0.50 logits for all items across gender, suggesting the absence of DIF or invariance of the items across the subgroups. DIF results for gender showed that CSS-T items could be considered invariant based on the criteria used. We explored DIF for matched ability levels for the following variables: COVID-19 (having COVID-19 before vs. not having COVID-19 yet) and enrollment status (student vs. non-student). DIF was observed between students and non-students for Item 7 (“*I am worried about grocery stores running out of food.*”) and Item 13 (“*I am worried that foreigners are spreading the virus in my country.*”). For Item 7, the estimated item location parameter was higher for students compared to non-students, illustrating higher levels of stress ($p = .0034$, contrast = 0.50). For Item 13, the estimated item location parameter was lower for students compared to non-students, illustrating lower stress levels ($p < .0001$, contrast = -0.61). DIF was also observed between respondents who tested positive for COVID-19 and not tested positive for COVID-19 before for Item 7. For Item 7, the estimated item location parameter was higher for COVID-19 patients compared to non-patients, illustrating higher levels of stress ($p = .0150$, contrast = 0.62).

Overall, the results of Rasch analyses provided evidence to support the substantive aspect of validity and the appropriateness of the CSS-T as a measure of COVID-19 stress level in a Turkish sample.

4. DISCUSSION and CONCLUSION

In this study, the English version of the CSS developed using American and Canadian samples was adapted into Turkish culture. In addition to CFA-related analyses, Rasch analysis was performed to provide additional evidence for the validity of the Turkish form. Within the scope of the study, translation and back-translation were performed, and subsequently, the CSS-T was

administered online to an adult Turkish sample. Its validity and reliability were examined with several methods.

The internal consistency coefficients for each subscale and the overall scale were relatively high. While the reliability coefficients of the original scale varied between .85 and .95 (Taylor et al., 2020b), the CSS-T in this study varied between .85 and .93.

To reveal the construct validity of CSS-T, the results of CFA and criterion-related validity findings were reported. In addition, Rasch analysis was performed to determine the contribution of the items to relevant dimensions as well as the entire scale.

As Taylor et al. (2020b) suggested testing both the proposed 6-factor model and the 5-factor model that emerged in their study, this study compared the 5-factor structure of the original CSS found for the Canadian and American samples with the expected 6-factor model. The CFAs revealed acceptable item fit statistics for the 6-factor model but not for the 5-factor model.

In the original form of CSS, COVID Danger and COVID Contamination subscales formed a single factor (Taylor et al., 2020b). This 5-factor model was confirmed in Persian (Khosravani et al., 2021) and Arabic (Abbady et al., 2021). However, in this study, the theoretical 6-factor model showed a better fit. While the items related to COVID Danger measure the person's concerns about getting infected directly, COVID Contamination, on the other hand, measures the sensitivity to factors that may cause transmission of the disease. Therefore, it was expected that these items would load on two different scales. It is noteworthy that the predicted 6-factor model was confirmed in the Turkish sample but not in the US and Canadian samples. It appears that the items of these two factors are more comprehensible in the Turkish Language.

The Turkish version of FCV-19-S (Satici et al., 2020) was used for criterion-related validity. The results indicated a moderate correlation between five of the CSS-T subscales and FCV-19S, but a relatively higher correlation was observed with Traumatic Stress. A meta-analysis showed a strong relationship between fear of COVID-19 and traumatic stress (Şimşir et al., 2021). Based on this finding, it can be concluded that the Traumatic Stress subscale and FCV-19S measure constructs are more similar, unlike the other subscales. Overall, the study proved that CSS-T is a valid measurement tool that can be used to assess COVID-19 anxiety more comprehensively compared to FCV-19S.

Rasch analysis was performed separately for each dimension to further support the validity of the CSS-T. First, the unidimensionality of each dimension of CSS-T was tested using PCA results. In the next step, reliability analysis was performed, and it was observed that all item and person separation measures exceeded cut-off values, indicating a sufficient spread of items across subscales. Item fit analysis was performed by evaluating item difficulty calibrations, standard errors, fit statistics (infit and outfit), and point-measure correlations. These values indicated a good amount of spread in the item locations, which is necessary for Rasch measurement scales to cover the full theoretical range of the construct's continuum. While the infit values of all items indicated a good fit, the outfit values were high only for Item 13 ("I am worried that foreigners are spreading the virus in my country.") and Item 30 ("I had bad dreams about the virus.") (1.51 and 1.55, respectively). However, these values are close to the 1.50 cut-off value suggested by Wright and Linacre (1994). Moreover, since infit is a weighted index while the outfit is unweighted, large outfit values are generally considered less problematic than large infit values (Bond & Fox, 2007).

Point-measure correlation values supporting were all above the cut-off point, supporting item-level polarity and unidimensionality of each subscale. When the person-item map was examined, the person's ability scores for each subscale appeared to be distributed normally, and the between-person variance indicated a heterogeneous mix of responses. Finally, Rasch-Welch and the Mantel Haenzel tests were applied to determine whether any items showed DIF according to gender, student status, and COVID-19 history. The results indicated that no item

showed DIF by gender, but DIF was observed for COVID-19 history Item 7. This Item's location parameter was higher for persons who had COVID-19 before compared to non-patients, illustrating higher stress levels. The fear of catching the virus and the fear of losing one's life due to COVID-19 brings are exacerbated by being separated from loved ones due to quarantine, withdrawal from social life, and exclusion. These situations can have traumatic effects on individuals. A previous study revealed that individuals who experience quarantine experience show more symptoms of anxiety and depression compared to those who were not quarantined (Wang et al., 2021). Therefore, the possibility of running out of food in the markets may affect individuals who tested positive for COVID-19 more compared to those who did not. DIF was also observed between students and non-students for Item 7 ("I am worried about grocery stores running out of food.") and Item 13 ("I am worried that foreigners are spreading the virus in my country."). Accordingly, for Item 7, the item location parameter indicated a higher stress level for students compared to non-students. One of the most important concerns related to COVID-19 is the possibility of running out of basic needs, such as food and cleaning materials in the markets, and not being able to meet these needs. This anxiety can also lead to panic buying (Mertens et al., 2020). Individuals can easily communicate disaster scenarios to others, especially through social media, which can increase panic. University students participating in the study were among the youngest participants. Therefore, they are likely to use social media more actively compared to their older counterparts (see Turkish Statistical Institute, 2020).

Finally, for Item 13, the item location parameter had higher stress levels for non-students than for students. This item measures the level of hostility towards foreigners. This finding can also be explained by the difference in attitudes and behaviors between generations, as in item 7. As individuals age, they may not tolerate change and thus develop more negative attitudes towards individuals who differ from them, such as immigrants and foreigners. Indeed, a previous study reveals that today's youth may be more tolerant of strangers compared to middle age and older age groups (Janmaat & Keating, 2019). Therefore, the finding that university students, who are the youngest participants, had lower stress levels compared to non-students in this study would be expected.

4.1. Strengths, Limitations, and Future Directions

There are many strengths and some limitations of this study. An obvious strength is that the study sample comprised individuals from a wide age range living in different regions of Turkey. Another strength of the study is that compared to the original CSS (Taylor et al., 2020b) and other adaptation studies (Abbadly et al., 2021; Khosravani et al., 2021), more sophisticated analyses were utilized in this study to support the validity of CSS-T.

This study validated a 6-factor model of the CSS-T. This finding shows that CSS-T is more compatible with the originally proposed structure rather than the 5-factor one that was supported in different cultures. Despite this strength, there is a need to compare the differences between cultures with further analysis, such as cross cultural measurement invariance and DIF.

As a result, a reliable and valid measurement tool was obtained in this study to measure adults' anxiety about COVID-19 across different factors. Another important strength of the CSS-T is that it can measure the long-term effects of COVID-19 that researchers and mental health practitioners can use to determine the long-term effects of COVID-19 on individuals. Taylor et al. (2020b) stated that the original CSS could be easily adapted to future pandemic situations. In this respect, the Turkish version of the scale could be used in future pandemics.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE

belongs to the authors. **Ethics Committee and Protocol No:** Anadolu University Ethics Committee, 50194.

Authorship Contribution Statement

Murat Dogan Sahin: Investigation, Scale Translation, Data Collection, Resources, Methodology, Software, Formal Analysis and Writing-original draft. **Sedat Sen:** Investigation, Scale Translation, Data Collection, Resources, Methodology, Software, Formal Analysis and Writing-original draft. **Deniz Guler:** Investigation, Literature Review, Data Collection, Scale Translation and Writing-original draft.

Orcid

Murat Dogan Sahin  <https://orcid.org/0000-0002-2174-8443>

Sedat Sen  <https://orcid.org/0000-0001-6962-4960>

Deniz Guler  <https://orcid.org/0000-0001-6006-5795>

REFERENCES

- Abbady, A.S., El-Gilany, A. H., El-Dabee, F.A., Elsadek, A.M., ElWasify, M., & Elwasify, M. (2021). Psychometric characteristics of the of COVID Stress Scales-Arabic version (CSS-Arabic) in Egyptian and Saudi university students. *Middle East Current Psychiatry*, 28(1), 1-9. <https://doi.org/10.1186/s43045-021-00095-8>
- Ahorsu, D.K., Lin, C.Y., Imani, V., Saffari, M., Griffiths, M.D., & Pakpour, A.H. (2020). The Fear of COVID-19 Scale: Development and Initial Validation. *International Journal of Mental Health and Addiction*, 20(1), 1537–1545. <https://doi.org/10.1007/s11469-020-00270-8>
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573. <https://doi.org/10.1007/BF02293814>
- Asmundson, G.J.G., Paluszek, M.M., Landry, C.A., Rachor, G.S., McKay, D., & Taylor, S. (2020). Do pre-existing anxiety-related and mood disorders differentially impact COVID-19 stress responses and coping? *Journal of Anxiety Disorders*, 74(1), 1-6. <https://doi.org/10.1016/j.janxdis.2020.102271>
- Baenas, I., Caravaca-Sanz, E., Granero, R., Sánchez, I., Riesco, N., Testa, G., Vintro-Alcaraz, C., Treasure, J., Jiménez-Murcia, S., & Fernández-Aranda, F. (2020). COVID-19 and eating disorders during confinement: Analysis of factors associated with resilience and aggravation of symptoms. *European Eating Disorders Review*, 28(6), 855–863. <https://doi.org/10.1002/erv.2771>
- Bernardo, A.B.I., & Mendoza, N.B. (2020). Measuring hope during the COVID-19 outbreak in the Philippines: development and validation of the state locus-of-Hope scale short form in Filipino. *Curr Psychol.*, 40(11), 5698–5707. <https://doi.org/10.1007/s12144-020-00887-x>
- Biçer, İ., Çakmak, C., & Demir, H. (2020). Coronavirus Anxiety Scale Short Form: Turkish Validity and Reliability Study. *Anadolu Kliniği Tıp Bilimleri Dergisi*, 25(1), 216–225. <https://doi.org/10.21673/anadoluklin.731092>
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.
- Boyras, G., & Legros, D.N. (2020). Coronavirus Disease (COVID-19) and Traumatic Stress: Probable Risk Factors and Correlates of Posttraumatic Stress Disorder. *Journal of Loss and Trauma*, 25(6–7), 503–522. <https://doi.org/10.1080/15325024.2020.1763556>
- Bueno-Notivol, J., Gracia-García, P., Olaya, B., Lasheras, I., López-Antón, R., & Santabárbara, J. (2021). Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies. *International Journal of Clinical and Health Psychology*, 21(1), 100196. <https://doi.org/10.1016/j.ijchp.2020.07.007>

- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.
- Evren, C., Evren, B., Dalbudak, E., Topcu, M., & Kutlu, N. (2020). Measuring anxiety related to COVID-19: A Turkish validation study of the Coronavirus Anxiety Scale. *Death Studies, 46*(5), 1052-1058. <https://doi.org/10.1080/07481187.2020.1774969>
- Gallagher, M.W., Zvolensky, M.J., Long, L.J., Rogers, A.H., & Garey, L. (2020). The impact of COVID-19 experiences and associated stress on anxiety, depression, and functional Impairment in American Adults. *Cognitive Therapy and Research, 44*(6), 1043–1051. <https://doi.org/10.1007/s10608-020-10143-y>
- Gavin, B., Lyne, J., & McNicholas, F. (2020). Mental health and the COVID-19 pandemic. *Irish Journal of Psychological Medicine, 37*(3), 156-158. <https://doi.org/10.1017/ipm.2020.72>
- Hu, L., & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453. <https://doi.org/10.1037/1082-989X.3.4.424>
- Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Janmaat, J.G., & Keating, A. (2019). Are today's youth more tolerant? Trends in tolerance among young people in Britain. *Ethnicities, 19*(1), 44-65. <https://doi.org/10.1177/1468796817723682>
- Khosravani, V., Asmundson, G.J.G., Taylor, S., Sharifi Bastan, F., & Samimi Ardestani, S. M. (2021). The Persian COVID stress scales (Persian-CSS) and COVID-19-related stress reactions in patients with obsessive-compulsive and anxiety disorders. *Journal of Obsessive-Compulsive and Related Disorders, 28*, 100615. <https://doi.org/10.1016/j.jocrd.2020.100615>
- Lee, S.A. (2020). Coronavirus Anxiety Scale: A brief mental health screener for COVID-19 related anxiety. *Death Studies, 44*(7), 393-401. <https://doi.org/10.1080/07481187.2020.1748481>
- Lee, S.A., Jobe, M.C., Mathis, A.A., & Gibbons, J.A. (2020). Incremental validity of coronaphobia: Coronavirus anxiety explains depression, generalized anxiety, and death anxiety. *Journal of Anxiety Disorders, 74*(1), 6-9. <https://doi.org/10.1016/j.janxdis.2020.102268>
- Linacre, J.M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.
- Linacre, J.M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*(1), 95-110.
- Linacre, J.M. (2009). *Winsteps* (Version 3.68.2) [Computer software].
- Liu, S., Lithopoulos, A., Zhang, C. Q., Garcia-Barrera, M.A., & Rhodes, R.E. (2021). Personality and perceived stress during COVID-19 pandemic: Testing the mediating role of perceived threat and efficacy. *Personality and Individual Differences, 168*. <https://doi.org/10.1016/j.paid.2020.110351>
- Mazza, C., Ricci, E., Biondi, S., Colasanti, M., Ferracuti, S., Napoli, C., & Roma, P. (2020). A nationwide survey of psychological distress among Italian people during the COVID-19 pandemic: Immediate psychological responses and associated factors. *International Journal of Environmental Research and Public Health, 17*(1), 1–14.
- Mertens, G., Gerritsen, L., Duijndam, S., Salemink, E., & Engelhard, I.M. (2020). Fear of the coronavirus (COVID-19): Predictors in an online study conducted in March 2020. *Journal of Anxiety Disorders, 74*(1), 102258. <https://doi.org/10.1016/j.janxdis.2020.102258>
- Pakpour, A.H., Griffiths, M.D., & Lin, C.Y. (2020). Assessing psychological response to the COVID-19: The fear of COVID-19 scale and the COVID Stress Scales. *International*

- Journal of Mental Health and Addiction*, 19(1), 2407-2410. <https://doi.org/10.1007/s11469-020-00334-9>
- Porcelli, P. (2020). Perspective article Fear, anxiety and health-related consequences after the COVID-19 epidemic Piero Porcelli. *Clinical Neuropsychiatry*, 17(2), 103-111. <https://doi.org/10.36131/>
- Psychology of Pandemics Network (2020). URL: <https://coronaphobia.org/>
- Qiu, J., Shen, B., Zhao, M., Wang, Z., Xie, B., & Xu, Y. (2020). A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: Implications and policy recommendations. *General Psychiatry*, 33(2), 1-4. <https://doi.org/10.1136/gpsych-2020-100213>
- Satici, B., Gocet-Tekin, E., Deniz, M.E., & Satici, S.A. (2020). Adaptation of the fear of COVID-19 scale: Its association with psychological distress and life satisfaction in Turkey. *International Journal of Mental Health and Addiction*, 19(1), 1980–1988 <https://doi.org/10.1007/s11469-020-00294-0>
- Seçer, İ., & Ulaş, S. (2020). An investigation of the effect of COVID-19 on OCD in youth in the context of emotional reactivity, experiential avoidance, depression and anxiety. *International Journal of Mental Health and Addiction*, 19(1), 2306-2319. <https://doi.org/10.1007/s11469-020-00322-z>
- Sick, J. (2010). Assumptions and requirements of Rasch measurement. *SHIKEN: JALT Testing & Evaluation SIG Newsletter*, 14(2), 23–29.
- Şimşir, Z., Koç, H., Seki, T., & Griffiths, M.D. (2021). The relationship between fear of COVID-19 and mental health problems: A meta-analysis. *Death Studies*, 46(3), 515-523. <https://doi.org/10.1080/07481187.2021.1889097>
- Taylor, S., Landry, C.A., Paluszek, M.M., Fergus, T.A., McKay, D., & Asmundson, G.J.G. (2020a). COVID stress syndrome: Concept, structure, and correlates. *Depression and Anxiety*, 37(8), 706–714. <https://doi.org/10.1002/da.23071>
- Taylor, S., Landry, C.A., Paluszek, M.M., Fergus, T.A., McKay, D., & Asmundson, G.J.G. (2020b). Development and initial validation of the COVID Stress Scales. *Journal of Anxiety Disorders*, 72, 102232. <https://doi.org/10.1016/j.janxdis.2020.102232>
- Turkish Statistical Institute. (2020). Hanehalkı Bilişim Teknolojileri (BT) Kullanım Araştırması, 2020 [Household Information Technologies (IT) Usage Survey, 2020]. Retrieved from [https://data.tuik.gov.tr/Bulten/Index?p=Hanehalki-Bilisim-Teknolojileri-\(BT\)-Kullanim-Arastirmasi-2020-33679](https://data.tuik.gov.tr/Bulten/Index?p=Hanehalki-Bilisim-Teknolojileri-(BT)-Kullanim-Arastirmasi-2020-33679)
- Valiente, C., Contreras, A., Peinado, V., Trucharte, A., Martínez, A.P., & Vázquez, C. (2021). Psychological Adjustment in Spain during the COVID-19 Pandemic: Positive and negative mental health outcomes in the general population. *Spanish Journal of Psychology*, 24(1), 1–13. <https://doi.org/10.1017/SJP.2021.7>
- Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., McIntyre, R.S., Choo, F.N., Tran, B., Ho, R., Sharma, V.K., & Ho, C. (2020). A longitudinal study on the mental health of general population during the COVID-19 epidemic in China. *Brain, Behavior, and Immunity*, 87, 40–48. <https://doi.org/10.1016/j.bbi.2020.04.028>
- Wang, Y., Shi, L., Que, J., Lu, Q., Liu, L., Lu, Z., Xu, Y., Liu, J., Sun, Y., Meng, S., Yuan, K., Ran, M., Lu, L., Bao, Y., & Shi, J. (2021). The impact of quarantine on mental health status among general population in China during the COVID-19 pandemic. *Molecular Psychiatry*, 26(1), 4813-4822. <https://doi.org/10.1038/s41380-021-01019-y>
- WHO. (2020). *Mental Health and Psychosocial Considerations During COVID-19 Outbreak*. World Health Organization, January, 1-6. Retrieved from: <https://www.who.int/docs/default-source/coronaviruse/mental-health-considerations.pdf>
- Wright, B.D., & Linacre, M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370-371.

APPENDIX

Table 1A. STDYX Estimate of the items according to CFA results.

Subscale	Item #	STDYX Estimate	SE
D	1	.772	.030
	2	.731	.034
	3	.614	.035
	4	.677	.032
	5	.719	.036
	6	.689	.038
SEC	7	.704	.040
	8	.726	.041
	9	.831	.028
	10	.839	.023
	11	.778	.032
	12	.816	.023
X	13	.733	.027
	14	.788	.021
	15	.921	.012
	16	.945	.008
	17	.786	.021
	18	.769	.023
C	19	.836	.021
	20	.820	.019
	21	.859	.017
	22	.811	.024
	23	.816	.024
	24	.810	.024
TS	25	.828	.021
	26	.840	.019
	27	.862	.015
	28	.844	.017
	29	.750	.027
	30	.686	.032
CH	31	.823	.020
	32	.744	.024
	33	.769	.024
	34	.758	.024
	35	.693	.028
	36	.715	.026

All *p* values < .001

Table 2A. Correlations among the CSS-T dimensions.

Subscale	D	SEC	X	C	TS
D	-				
SEC	.54				
X	.51	.43			
C	.76	.43	.64		
TS	.66	.47	.47	.60	
CH	.53	.32	.37	.54	.64

All p values < .001

Exploring how the use of a simulation technique can affect EFL students' willingness to communicate

Houman Bijani^{1,*}, Masoumeh Abbasi¹

¹Islamic Azad University, Zanjan Branch, Zanjan, Iran

ARTICLE HISTORY

Received: Aug. 27, 2021

Revised: May 26, 2022

Accepted: June 15, 2022

Keywords:

EFL learners,
Elementary learners,
Simulation,
Speaking proficiency,
Willingness to
communicate.

Abstract: This study is intended to explore an applicable and effective model of simulated situation for English as a Foreign Language (EFL) learners and also investigate the effects of the simulated environment on Willingness to Communicate (WTC) of the learners. To carry out this study, 300 elementary level EFL learners were chosen. A Key English Test (KET) was administered to ensure homogeneity on the learners. They were divided into two groups of experimental and control. A WTC questionnaire developed by Macintyre, Baker, Clement, and Conrod (2001) was used, after validation through Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA) and Modeling, as an instrument to obtain primary data. The results of Mann-Whitney U test revealed that simulated environment had positive effects on the participants' WTC. The findings of this study suggest that understanding how a simulated environment affects EFL learners' success in speaking proficiency can help institutes to provide such environments for EFL learners and instructors. This method can be presented at different levels of English proficiency. The focus of this study was mainly on speaking skill; therefore, similar studies can be conducted regarding other language skills, e.g., writing, listening and reading.

1. INTRODUCTION

Nowadays, it is undeniable that learning a foreign language has become a significant part of people's lives. In this increasingly globalized world, learning a foreign language can help people progress in their career, become aware of other cultures and help them to increase understanding and knowledge of their own language. The global significance of English education has affected the society of Iran. Therefore, Iranians try to learn English and improve their English proficiency to get a good job, achieve better employment prospect, enhance social status, immigrate to other countries, etc. In Iran English is particularly a means of having access to new information and technology, though there is emphasis on reading comprehension. (Kiany et al., 2011). In the past, the emphasis of teaching English was on teaching grammar rules and vocabularies. But nowadays the emphasis is on teaching oral aspects of the target language. According to Dörnyei (2005), the goal of teaching is "the learners' communicative competence in the target language" (p.207). Speaking is an important skill in learning foreign language, since EFL learners should use that skill to convey messages and express ideas.

*CONTACT: Houman Bijani ✉ houman.bijani@gmail.com 📍 Islamic Azad University, Zanjan Branch, Zanjan, Iran

According to Lazarton (2001), when a learner is able to communicate orally, it means that s/he knows the given language because speaking is the primary tool for communication. Aleksandrak (2011) states that the main problem in EFL learning process is the insufficient speaking varieties and lack of appropriate chances of speaking in the classrooms compared to opportunities in the real-life situations. Teaching English in learning contexts is so helpful for learners. Most Iranian EFL learners are taught grammar rules, vocabularies and pragmatic features without immersion in contexts. According to Witmer and Singer (1998) immersion is the psychological response to the technology. Based on several recent studies, immersion is like a product of technology in which the user is provided by the production of multimodal sensory “input” (Bystrom et al., 1999; Draper et al., 1998; Slater & Wilbur, 1997).

One way to create such a situation for learners in EFL classrooms is through simulation which simulate EFL classes to real world life. Since classroom is a small symbol of the real world, simulation will help EFL learners to have many opportunities to engage in communication as if they were in a real situation.

Few studies have suggested an appropriate and cheap simulated design. Wang, Petrina and Feng (2015) have suggested 3D virtual environment in their study for immersion in a real life. However, providing computers for learners in the classroom is a costly way for most institutes. This computer-assisted environment cannot be a real and tangible environment for learners and they do not have real interactions with other peers. To fill these gaps, one purpose of this study is to suggest an applicable and effective model of simulated situation for EFL learners. In this model teachers can change the environment of the class according to the context of the lesson and the learner would use their background language knowledge in that simulated area.

This study investigates the provision of a simulated environment for learners to make them feel that they are in the target country and encourage them to express themselves to other peers and make communication with them. The important point is that the learners do not need to go to the target country physically to be in such situation. The simulation may provide simulated situations or scenes for EFL learners.

This study aimed at describing the process of learning a foreign language in a simulated environment in the classroom among 14-17-year-old female children. This study attempted to investigate the effects of simulation of real-life situations on EFL learners considering different influential factors, such as autonomy of learners, willingness to communicate and their speaking proficiency.

1.1. Review of Literature

1.1.1. Willingness to communicate

One of the factors which has recently been presented in Second Language Acquisition (SLA) studies is willingness to communicate (WTC). MacIntyre, Baker, Clement and Donovan (2002) characterize WTC as “...the inclination toward or absent from communicating, given the choice” (p.538).

WTC is one of the emotional variables which is expected to impact success in language learning. According to Richmond and Roach (1992), in case a speaker has high WTC, s/he is more likely to be successful in learning a second language. That demonstrates the noteworthy part of WTC in learning foreign languages. Hashimoto (2002) examined the impacts of WTC and motivation on second language in a Japanese setting. The results appeared that, in the event that the learners’ competence information expanded and his/her anxiety decreased, WTC and using the second language expanded within the classroom. Instructors can increment WTC of learners by creating less threatening environment within the classroom and propelled learners to extend their perceived competence.

As the emphasis in L2 teaching and learning has been moving to communication, both as an essential process and as an objective of learning a L2, a way to account for individual contrasts in L2 communication is required. Zarei et al., (2019) illustrated the plausibility by combining insights from two disciplines, L2 acquisition and communication.

In Japan, as the Ministry of Education, Culture, Sports, Science, and Technology's rules for foreign language (generally English) educating inside the school instruction curriculum (Monbusho, 1989; 1999a; 1999b) have put expanding emphasis on communication, a more noteworthy portion of reading material and classroom exercises has centered on face-to-face interaction in theoretical intercultural contact circumstances. There is expansion to inspiration and states of mind toward the individuals with whom students will communicate. In this respect, WTC, psychology of communication, and intercultural stances have to be inspected as factors that influence communication results. Amirian, Karamifar and Youhanaee (2020) stated that second language learners' WTC can be expanded by giving opportunities to form an environment for learners that they would feel comfortable to communicate with each other since the learners with high WTC use second language in authentic communications. They expressed that WTC in language settings exists as personal physiological variables and situational factors. Assuming that numerous variables impact a person's readiness to communicate, such as fear of talking, need of self-esteem and the issue of introversion and extroversion (McCroskey, 1992), the significance of assessing the degree of the impact of WTC in success in SLA becomes clear.

Yashima (2002) demonstrates a direct connection between students' WTC and their attitude toward worldwide community within the EFL (English as a Foreign Language) context. In the ESL (English as a Second Language) context, Clement et al., (2003) appear an indirect connection through linguistic self-confidence between WTC and attitude toward the other language group.

MacIntyre et al., (1998) characterized WTC as "the likelihood of engaging in communication when free to select to do so" (p.546). However, Sheybani (2019) did not treat WTC in L2 as an identity characteristic but as a situational variable that has both temporal and enduring impacts. In addition, they theorized that WTC impact not only talking mode but also listening, writing and reading modes.

Wen and Clement (2003) examined Chinese social context impacts on learners' WTC. Further analysis shows that features including interpersonal relations such as other-directed self and submissive way of learning are the components which shape the Chinese students' learning behaviors within the course. In another study, Mallahi and Hosseini (2020) explored Iranian EFL learners' unwillingness to communicate. The results appeared that unwillingness to communicate is related to language anxiety, language proficiency and access to English.

1.1.2. WTC in a L2

One of the factors which has recently been presented in SLA investigate is WTC. Communication apprehension in a L1 and its negative impact on communication have been a matter of insightful attention by communication analysts (Daly & McCroskey, 1984; McCroskey, 1977).

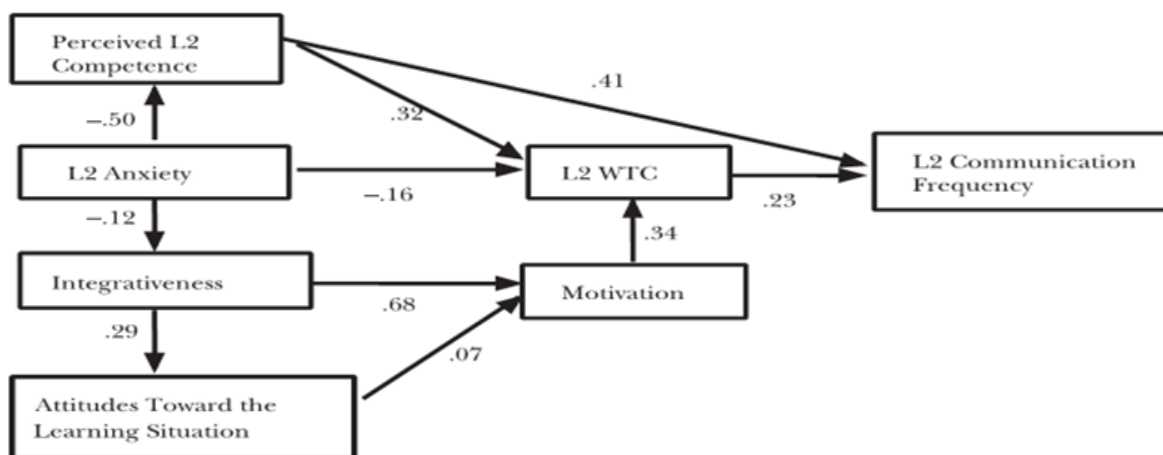
McCroskey and associates (e.g., McCroskey, 1992; McCroskey & Richmond, 1990) proposed the construct, WTC, that captures the major suggestion that communication apprehension, introversion, hesitance, and modesty have for communicative behavior. MacIntyre (1994) created a path model that hypothesizes that WTC is based on a combination of more prominent seen communicative competence and a lower level of communication anxiety (Figure 1).

Figure 1. Portion of MacIntyre’s (1994) WTC model.



Studies conducted in different Canadian contexts combined the WTC show with Gardner’s socio-educational model to look at the relations among factors underlying WTC in a L2. In these studies, WTC was a predictor of frequency of communication in a L2, though motivation was a predictor of WTC, frequency of communication in a L2, or both (MacIntyre & Charos, 1996; MacIntyre & Clément, 1996; see Figure 2 as an example). MacIntyre did not regard WTC in a L2 as a simple manifestation of WTC in a L1; a much greater range of communicative competence is obvious in a L2 than in a L1. In addition, according to MacIntyre et al. (1998) “L2 use carries a number of intergroup issues, with social and political implications, that are usually irrelevant to L1 use” (p. 546).

Figure 2. MacIntyre and Charos’ (1996) model of L2 WTC applied to monolingual university students.



MacIntyre et al. (1998) conceptualized WTC in a L2 in a hypothetical model. In this model, learner identity, inter-group climate, intergroup attitudes, inter-group motivation, L2 self-confidence, and communicative competence, among other components, are interrelated in affecting WTC in L2.

1.1.3. Simulation

Simulations as a language learning approach/tactic have been interpreted in different resources and/or by diverse authors in different ways. The terms utilized within the role playing/simulation literature are regularly utilized interchangeably as well, such as: “simulation”, “game”, “simulation game”, “role-playing game”, “role-play simulation”.

In spite of the fact that the word “simulation” and its definition in a dictionary may suggest that in simulations the participants simulate (act, play, and pretend), the simulations in language educating and learning displayed are not the same as a role-play or game. They are based on Jones’s (1986) definition of a simulation “as reality of function in a simulated and structured environment” (p. 173). In other words, rather than a role to play, students have a real-life task to achieve.

1.1.3.1. Difference Between Simulations and Role Plays. As simulations are most frequently mixed up with Role plays, the main differences between these two language learning activities are shown in the table below (Table 1).

Table 1. *Difference between simulations and role plays.*

Simulations	Role Plays
The (simulated) environment is provided, using text, audio or video input.	Participants have to create (imagine) aspects of the environment.
Key facts are provided for the background (sex, age, job etc.)	Participants invent key facts or have to act descriptions according to a specific script or no script provided “You are angry because your friend broke your watch.”
Participants take on a role (accept duties/responsibilities and perform task according to their own personalities)	Participants play/act out a pre-defined role. (Pretend to be someone else according to the provided role-card)
Imagination may be involved. Invention is not allowed.	Participants are encouraged to invent/create whatever is necessary to play the role.
Real communication in a controlled realistic situation.	Dialogues in a fixed context of speech in an improvisational and imaginary one.

In a Role Play, one student might be told that she is a supermarket checkout assistant whilst another is a customer. Students might also be given fairly tight guidelines outlining the nature of their exchange or the language points they are expected to cover. Role Play involves participants to ‘act’ in a given role which is clearly defined on a role-card. It is very much akin to acting in a play. Simulations, however, allow students to express themselves to their peers in a group setting (3 or 4 students in a group) where they retain their own personalities and are not required to pretend to be someone else. Or, as Wang (2005) says:

“... simulations, where simple or complex, do not specify the role a person has to play. On the contrary, a task is given which requires participants to resolve a problem of some kind using their own life experience and character. Simulation mimics real life situation as closely as possible. For example, if you have a group of doctors learning English as a second language and they need to practice in a “real life” context, you would set up a simulated situation in a hospital or health center in which doctors have to meet ‘patients’ and diagnose their problem, and give treatment or prescriptions. The ‘patients’ may be given (or create themselves) their symptoms, and the doctors have to find out the cause of the illness (using their own experience) by interacting with the patients. The problem is resolved when the doctor diagnoses the problem and prescribes therapy (p. 20).”

1.1.3.2. Simulation in Teaching English. These days, English learning is more coordinated at the communications function. Learning English is aiming that students are able to utilize English to communicate not only learn the science of language itself. This is in line with the communicative approach which emphasizes that learning a language is learning to communicate (Richards, 1985) however, in practice, learning models are still not able to supply many opportunities for learners to utilize language that has been examined. This is proved from the number of students who are still afraid to talk English even in spite of the fact that they have sufficient lexicon. This is not a portion of the methods and strategies utilized by instructors. Numerous instructors are split between teaching materials and the ultimate objective of learning, for example, teachers might give their own lexicon without entering the lexicon of a context for communication.

Another problem that is found in learning process is that the students are in classroom environments that provide few opportunities to engage in communication in realistic situations, whereas practice plays an important role in improving communication skills, but learners have

a lack of opportunities to do so language teaching is ideally suited to language practice. Language teaching can be an interesting process when teachers make the effort to explore a variety of approaches. However, unfortunately, only a few teachers can do it. It can be caused by the lack of experience and knowledge about the varieties of teaching methods and techniques. There are many techniques can be applied in teaching English for elementary school students, one of them is simulation. Simulation is a language learning model which allows students to express themselves to their peers in a group setting, groups comprising usually three or four. Some benefits of simulation allow students to experiment with new vocabulary and structures and gives students the chance to carry out a task or solve a problem together Simulation technique follows from the interactional view. This view sees language as a vehicle for the realization of interpersonal relations and for the performance of social transactions between individuals. Language teaching content, according to this view, may be specified and organized by “patterns of exchange and interaction or may be left unspecified, to be shaped by the inclinations of learners as interactors.” (Richards & Rodgers, 2001, p.17)

Simulation clearly promotes effective interpersonal relations and social transactions among participants. "In order for a simulation to occur the participants must accept the duties and responsibilities of their roles and functions and do the best they can in the situation in which they find themselves" (Jones, 1982, p.113).

The problem with English classes in Iran is that there are not many opportunities for learners to utilize what they have already learned, practically in the class. In some cases, the learners have the knowledge of English grammar or vocabulary but due to the lack of opportunities they cannot use them to speak fluently and accurately. One way to solve this problem is that learners should experience living in an English-spoken country to learn English effectively since they would immerse in that situation. According to Kemp (2003), people who have mastered a foreign language believed that the most beneficial way of learning a foreign language occurred while the person is immersed in the target-language spoken environment. However, it is quite difficult for EFL learners to travel or immigrate to target countries in order to learn English and enhance speaking proficiency and it is almost impossible for most English language institutes in Iran to send EFL learners to target language countries due to financial challenges. Therefore, creating a simulated environment of the target country situations in the classrooms can be an alternative approach instead of costly and time-consuming way of sending learners to target countries. In simulated environment, EFL learners would feel that they are in a real and tangible situation, so they will inevitably communicate in English with each other and even create conversations according to the environment in which they are.

Although many role-playing researchers support the effects of digital games on EFL learners' performances (Liang, 2012; Peterson, 2011; Thorne et al., 2009), few studies focused on the virtual contexts for elementary students using language effectively in real situations. Using digital games would not be possible to use in every class and there should be special facilities to be used by students.

Few studies have suggested an appropriate and cheap simulated design. Wang, Petrina and Feng (2017) have suggested 3D virtual environment in their study for immersion in a real life. However, providing computers for learners in the classroom is a costly way for most institutes. This computer-assisted environment cannot be a real and tangible environment for learners and they do not have real interactions with other peers.

Although many researchers (Lan, 2017; Li & Topolewski, 2002; Wang et al., 2015) have focused on the benefits of using simulation in EFL classrooms by using real objects or visual games by providing these objects and facilities through changing the environment for different contexts, it is costly and difficult for teachers. For instance, Lyu (2006) believed that using simulation for basic level classes by creating simple simulations with less complicated

processes is a good idea. He suggested that teachers can provide real objects for learners to simulate the environment. For example, learners can use some maps in order to learn the directions. Using map can be a useful technique for simulation, however it cannot simulate the whole context for them to immerse in it.

Many studies investigated the effects of the environment or simulated situation on learning a foreign language (e.g., Lan, 2017; Wang, Petrina & Feng, 2017); however, a few studies have considered the effects of simulated environment on WTC of the EFL learners.

Response to the aforementioned gaps will definitely have important implications for second language learning research in general, and EFL teachers in particular. Thus, the following research question guided the present study:

RQ: What effects does language-learning simulation have on willingness to communicate of Iranian elementary EFL learners?

2. METHOD

2.1. Design

This research is a survey study to find the effect of a simulated environment on WTC of Iranian elementary EFL learners. To this end, quantitative data was gathered based on related questionnaires. The study design was quasi-experimental. Regarding the grouping procedures, the participants were assigned to two groups of experimental groups using the treatment (simulated environment) and control group (traditional method). The experimental group included 150 participants and the control group had 150 participants, too. There were pre-tests and post-tests for both groups. Elementary EFL learners were selected as the population of this study. A simple random sampling technique was used in this study. The dependent variables included WTC of the learners. The independent variable was simulated environment. There were some control variables like age, social class, background of language knowledge and bilingualism.

2.2. Participants

To carry out the present study, two elementary classes were selected from a private language institute in Zanjan, Iran. Their level was elementary i.e. A2 on the Common European Framework of Reference (CEFR). They were specified to this level based on institute's criteria and the placement test. In order to homogenize the participants, they all participated in a sample of Cambridge KET (2007). The ones who took the test and their scores fell between one standard deviation above or below the mean were chosen to participate in the study.

The participants were 300 elementary EFL learners that were divided into two equal groups, i.e., experimental (n=150) and control (n=150). Each group was assigned to five classes each containing 30 students; in other words, there were altogether 10 classes for both the experimental and control group participants. All the students were selected from various state high schools in Zanjan, Iran; however, as indicated before, they were all homogenized through Cambridge KET (2007) Placement Test. All the students were non-native speakers of English who had learned English as a foreign language in a non-English learning context. They were bilingual speakers of Persian and Turkish. The learners in the experimental group were exposed to the treatment (simulated environment), whereas the learners in the control group were involved in traditional method. They were all female and the age range of the learners was 14-17. All the learners studied Pearson's Top Notch fundamentals (Saslow & Ascher, 2005) during the term. The population of this study was monolingual and bilingual (Persian & Turkish). Their first language was Persian. The social class of participants was middle-class.

2.3. Instruments

The course book which was used in this study was Pearson's Top Notch. This series are co-authored by Saslow and Ascher and came out in 2005. The whole series are divided into three classes of volumes according to the proficiency of the learners (two Fundamentals, six Top Notches and four Summits) and include different units and subsections like warm-ups, starting conversation (dialogue), grammar spot, structural drills and lexical exercises, reading and checkpoints. The participants studied Top Notch fundamentals during the term. This book was used as the main course book in the institute.

To show different pictures or video clips on the walls, two video projectors were used. Some professional 3D computer graphics program such as 3D Max, Revit and Maya were used for making 3D images and animations related to the contents. A detailed explanation of the application of the treatment is given in the procedure section.

The instruments used in this study included: A) Key English Test (KET), and B) WTC questionnaire, which will be discussed in more details below:

2.3.1. Key English test

The first instrument which was used in this study was the Cambridge KET also known as 'Key'. Cambridge tests include all four skills – listening, speaking, reading and writing. They are arranged around four necessary qualities: “validity, reliability, impact and practicality” (university of Cambridge ESOL Examinations, 2008, p.2). KET is an English language test which is in sync with A2 level on the Common European Framework of Reference (CEFR) which is used to demonstrate the communication ability in simple situations. The only purpose of running KET test was to ascertain the homogeneity of the students before applying the treatment in order to assign the participants to experimental and control groups. Besides, since the focus of this was on oral communication (to be published in forthcoming papers), only the speaking section of KET was administered to the participants.

2.3.2. WTC questionnaire

To measure the students' WTC, a modified version of the Likert-type WTC questionnaire was used. This questionnaire includes 27 items and was developed by Macintyre, Baker, Clement, and Conrod (2001) to measure the learners' WTC both inside and outside the classroom. The questionnaire which is used in this study includes 23 items which range from 1 to 4 (1= almost never willing, 2= sometimes willing, 3 = willing half of the time, 4 = almost always willing). The reason for the elimination of four items in the questionnaire and reducing item numbers from 27 to 23 was due the fact that those four items contained western cultural issues and concepts which did not comply with the students' background who were raised in the context of Iran. Thus, in order not to cause confusion for the participants of this study, those items were crossed out to ensure the suitability and understanding of the questionnaire in the context of administration. The learners were asked to demonstrate that how much willing they were to communicate both during the class time and outside of the classroom. The questionnaire was translated to Persian, their first language, since their proficiency level was elementary and understanding the questions might be difficult to them. Having translated the questionnaire into Persian language, it was pilot-run and the reliability of the questionnaire was measured through Cronbach Alpha $r = 0.696$ (see Table 3 for a detailed analysis). According to Cohen's Table of Effect Size, the reliability measure was found much larger than typical. Therefore, it can be concluded that the questionnaire possessed an acceptable internal reliability. The validity of the questionnaire was confirmed through CFA, EFA and modeling. For measuring EFA and CFA, the researcher had to obtain the scree plot figure of the questionnaire in order to determine how many factors could be identified as the significant factors under which the items of the

questionnaire could be loaded. This will pave the way to specify the number of influential significant factors of the WTC questionnaire.

2.4. Data Collection Procedures

Data were collected by two teachers in order to determine the impact of simulation of WTC of the participants. In the beginning, to homogenize the participants, KET was applied. They were homogenized according to institute criteria. However, to make sure they are at the same level, they were tested by KET once more. Ultimately, 300 participants were classified as experimental and control groups (see [Table 2](#) for data analysis).

Before running the treatment, Macintyre et al.'s (2001) WTC questionnaire was given to the participants of both groups and they were asked to fill them in after 20 sessions of treatment, the WTC questionnaire was given again to both groups to complete. The outcomes of the questionnaire were compared to those of the same questionnaire distributed to participants before the treatment.

In the next phase, the treatment began. The two classes were taught by the same instructor. Both groups were taught Top-Notch fundamentals. The number of sessions for each group was 20 sessions. Treatment was conducted in June 2019 and continued for two months. It is noteworthy to mention that the instructor used the same course book, materials and activities for both classes, but in the experimental group, the dialogues and related contents, tasks and conversation-related activities were exposed in the simulated designed environment. During the treatment, appropriate images and animations that were related to their course book's contents were displayed through video projectors every session. The video projection was designed in a form of a four-sided OHP (Overhead Projector) that was used to project the environmental screens to which the content of the course was related on the wall. For example, if the content of the course was related to the hospital setting, a four-sided projection representing the hospital environment would be displayed on the walls of the class making the students feel like they are in the real context. This would simulate the real situation using virtual reality technique to better help the students visualize themselves in the real context to better comprehend the related materials.

Learners were taught the input and became aware of the content of the lesson, then they saw the related images or videos on the screens around them. It was like that they were in the real environment. Gradually they started to speak and make communication with each other by using what they have learned.

After 20 sessions of treatment, WTC questionnaire was given to both groups again to fill it in. The results of the questionnaire were compared to the results of the same questionnaire, which was distributed to participants before the treatment. The learner autonomy questionnaire was given to the learners after the treatment.

2.5. Data Analysis

Several analyses were done by SPSS software, such as Mann-Whitney U test, exploratory factor analysis (EFA) and Pearson Correlation. Confirmatory factor analysis (CFA) was done by Amos (version 22.0).

To examine whether the difference found between two groups regarding WTC was statistically meaningful or not, a Mann-Whitney U test was employed as the data were non-parametric and the data distribution was not normal. In the first phase of validating process, EFA was conducted to construct the validity of WTC questionnaire and also to identify factor structures. The correlation between the extracted factors was determined by Pearson correlation. In the next phase, CFA was run to neutralize the loading effects of the items on the factors and to keep the significant loadings of the significant items on the factors.

3. FINDINGS

As one of the most important assumptions for running parametric statistical analyses, is the assumption of the normal distributions of the scores, the researcher ran Kolmogorov-Smirnov and Shapiro-Wilk tests on the obtained scores to ensure the normality of the distributions based on the participants' performance on the speaking section of KET test. Table 2 presents the results.

Table 2. Tests of normality for participants' fluency and accuracy across control and experimental groups obtained from KET.

		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Fluency	Control	.119	15	.200*	.967	15	.804
	Experimental	.114	15	.200*	.937	15	.346
Accuracy	Control	.207	15	.084	.908	15	.126
	Experimental	.218	15	.053	.893	15	.074

As is illustrated in Table 2, the significant for each set of scores is higher than 0.05; therefore, all sets of scores have normal distributions which indicated that the participants were homogeneous prior to running the treatment. Based on this homogeneity, the participants were assigned randomly to experimental and control groups afterwards.

Cronbach alpha reliability coefficient, Kaiser-Meyer-Olkin (KMO) test, Bartlett's test, Varimax Rotation and Maximum likelihood method were used to determine the reliability and validity of this questionnaire. To examine the reliability of the WTC questionnaire, the Cronbach alpha was used.

Table 3. Reliability statistics of WTC questionnaire.

Cronbach's Alpha	N of Items
.696	23

According to Table 3, the Cronbach's Alpha reliability for the WTC questionnaire is 0.696. According to Cohen's Table of Effect Size, the reliability measure was found much larger than typical. Therefore, it can be concluded that the questionnaire possessed an acceptable internal reliability.

Having made sure about the reliability of the instruments in use, we can deal with the analysis of the research question.

RQ: What effects does language-learning simulation have on willingness to communicate of Iranian elementary EFL learners?

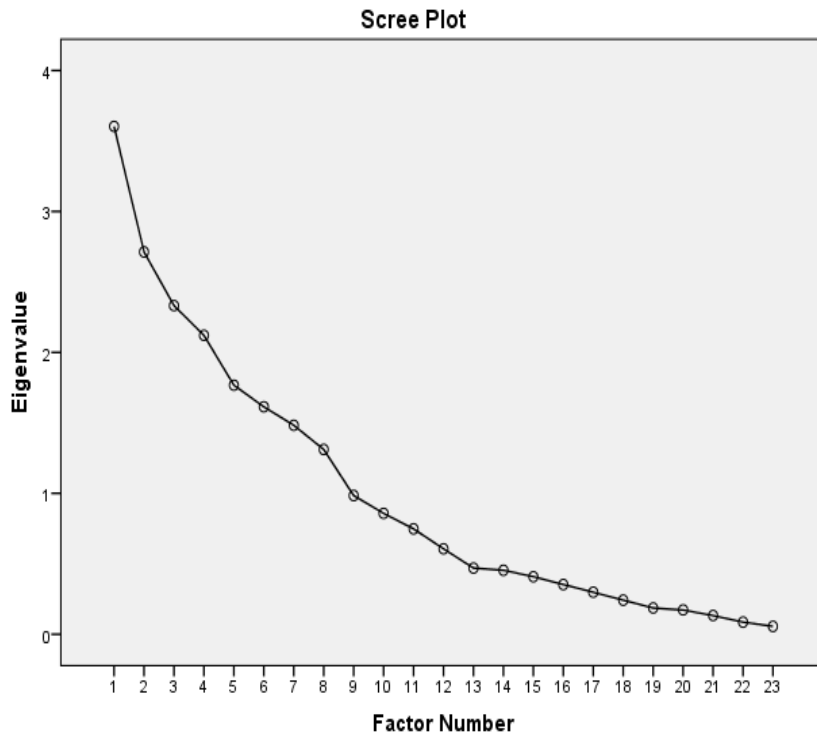
Kaiser Meyer Olkin (KMO) and Bartlett Test (see Table 4) were carried out to check the appropriateness of the data for factor analysis. The significant valid for the validity must be lower than 0.05 degree of probability and in the current questionnaire, the validity measure was found to be 0.000. It shows that the validity could be measured.

Table 4. The results of the KMO and Bartlett's test for WTC questionnaire.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.397
	Approx. Chi-Square	619.740
Bartlett's Test of Sphericity	df	253
	Sig.	.000

Exploratory factor analysis (EFA) was conducted to extract the new factor structure and to examine the construct validity. Since confirmatory factor analysis (CFA) was carried out after EFA, maximum likelihood method was conducted. A scree plot was used to confirm that current scale includes eight factors. The scree plot graph in Figure 3 indicates that there are eight components at the elbow. It means that the questionnaire items were loaded in eight significant factors.

Figure 3. Scree plot of WTC questionnaire.



Pearson's correlation coefficient is used to measure of the strength of the association between the eight factors which were extracted.

Table 5. The results of the KMO and Bartlett's test for WTC questionnaire.

Factor	1	2	3	4	5	6	7	8
1	1.00	-.003	.068	.014	.054	.108	.219	-.022
2	-.003	1.00	.120	-.083	-.061	-.033	.025	.051
3	.068	.120	1.00	-.079	-.017	.069	.164	-.082
4	.014	-.083	-.079	1.00	.049	.072	.002	.134
5	.054	-.061	-.017	.049	1.00	-.057	.145	.007
6	.108	-.033	.069	.072	-.057	1.00	.088	-.081
7	.219	.025	.164	.002	.145	.088	1.00	-.104
8	-.022	.051	-.082	.134	.007	-.081	-.104	1.00

Extraction method: Maximum likelihood, rotation method: Oblimin with Kaiser Normalization.

Table 5 presents that all the correlation between factors is below 0.300. It means that there is a weak correlation between factors. The correlations between factors was weak and rotated solution was obtained through Varimax rotation in order to clarify the factor structure. According to the results of EFA, confirmatory factor analysis (CFA) was conducted to determine whether the current model is confirmed or not. The model obtained from the analysis can be seen in Figure 4.

Figure 4. Confirmatory factor analysis model of WTC questionnaire.

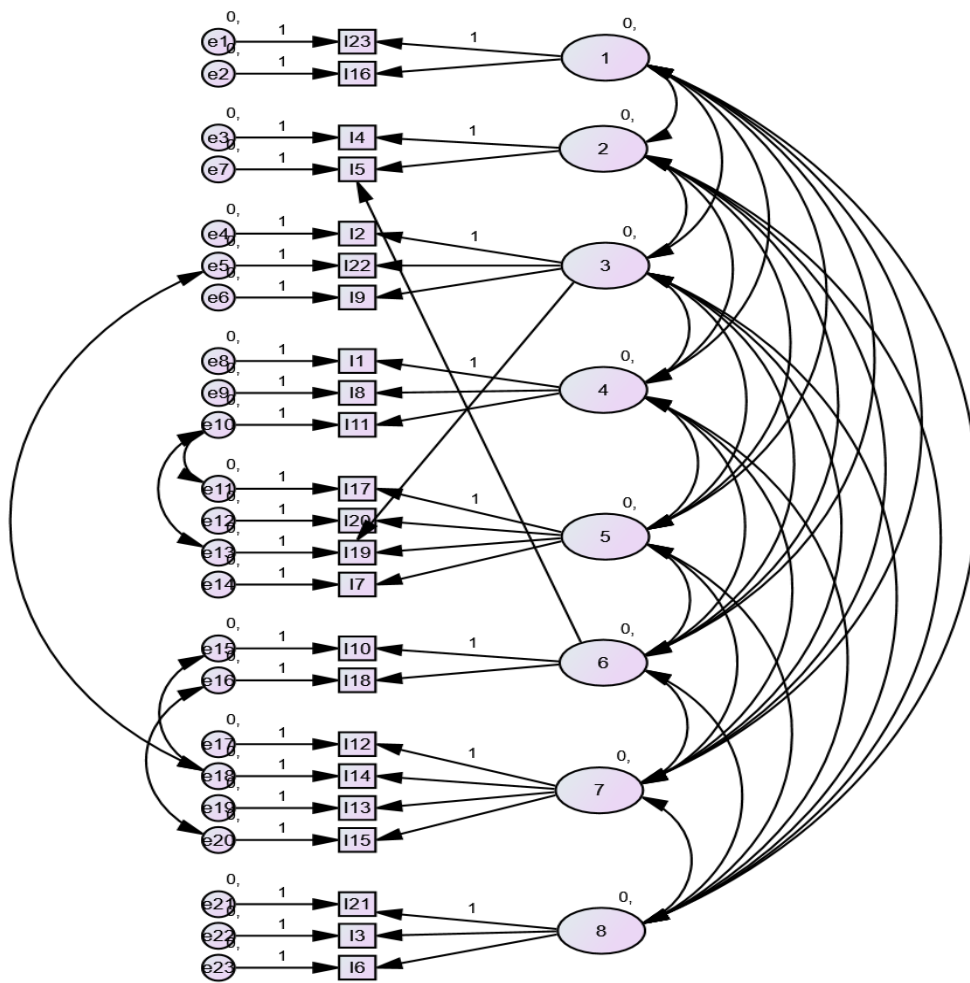


Figure 4 shows the items have loadings higher than 0.3. it means all the items of questionnaire were significant. As illustrated, items (16) and (23) loaded onto Factor 1. Items (4) and (5) loaded onto Factor 2. Items (19), (2), (22) and (9) loaded onto Factor 3. Items (1), (8) and (11) loaded onto Factor 4. Items (17), (20), (19) and (7) loaded onto Factor 5. Items (5), (10) and (18) loaded onto Factor 6. Items (12), (14), (13) and (15) loaded onto Factor 7. Items (21), (3) and (6) loaded onto Factor 8.

Table 6. The results of chi-square analysis of goodness of fit index.

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	104	370.605	195	.000	1.901
Saturated model	299	.000	0		
Independence model	46	724.052	253	.000	2.862

Table 6 shows the chi-square fit statistics. As it can be seen, the chi-square statistics of $\chi^2=370.605$, $df=195$, $p=0.000$ and relative chi-square (CMIN/df) = 1.901 which is smaller than 5.0 indicating an acceptable fit. P-value is 0.000 which is less than 0.5. It means that it is significant. Significant value shows that this model is different than the default one.

A good model fit has some criteria for goodness-of-fit indices. The TLI, CFI and NFI should be 1. However, according to Ho (2006), a cut-off value close to 0.90 is commonly used for

these incremental fit indices. For current model, CFI is 0.827, TLI is 0.816 and NFI is 0.812. The results indicate an acceptable fit. All the results were summarized in [Table 7](#).

Table 7. Baseline comparisons for goodness of fit index.

Model	NFI Delta1	RFI rho1	IFI Delta2	TLI rho2	CFI
Default model	.812	.336	.768	.816	.827
Saturated model	1.00		1.00		1.00
Independence model	.000	.000	.000	.000	.000

It can be said that one of the most commonly used goodness of fit indexes in CFA is RMSEA. The RMSEA value of this model is 0.074 which is below 0.07 and it indicates an acceptable model (see [Table 8](#)).

Table 8. Goodness of fit values obtained from CFA.

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.074	.104	.143	.000
Independence model	.178	.163	.193	.000

If the Standardized RMR value is lower than 0.08, it can be considered as an acceptable value that indicates the model-data fit (Byrne, 2001). In this current model standardized RMR is 0.078 which is an acceptable value for a good model. According to these results, it can be confirmed that model-data fit is acceptable.

Table 9. Descriptive statistics of the results of WTC questionnaire.

	Group	N	Mean Rank	Sum of Ranks
WTC_Pre	Control	150	18.27	2740.00
	Experimental	150	12.73	1910.00
	Total	300	31.00	
WTC_Post	Control	150	11.17	1670.50
	Experimental	150	19.83	2970.50
	Total	300	31.00	

[Table 9](#) represents the mean rank of the control and experimental groups in pre-test and post-test, for their WTC. To see whether there is significant difference between control and experimental groups in pre-test and post-test, regarding their WTC, Mann-Whitney U test was run. [Table 10](#) displays the Mann-Whitney U test examining the difference between control and experimental groups in pre-test and post-test.

As it can be seen in [Table 10](#), the significant (0.085) of control and experimental groups in pre-test is higher than 0.05. It is concluded that there is no significant difference between the groups in pre-test. In other words, they were homogenous regarding WTC before the study. Based on the results ([Table 10](#)), the significant (0.007) in post-test is less than 0.05 meaning that there is a significant difference between experimental and control groups regarding WTC. It can be concluded that the treatment had significant effects on the level of WTC of the participants.

Table 10. *The Mann-Whitney U test exploring the effect of simulation on WTC.*

	WTC_Pre	WTC_Post
Mann-Whitney U	71.000	47.500
Wilcoxon W	191.000	167.500
Z	-1.724	-2.707
Asymp. Sig. (2-tailed)	.085	.007
Exact Sig. [2*(1-tailed Sig.)]	.089	.006

Table 9 presents mean ranks of the groups in pre-test and post-test, regarding their level of WTC. The mean rank of post-test shows that the mean of control group is 11.17 and the mean of experimental is 19.83. The difference is 8.66. The results indicated that the students in the experimental group has a significant higher mean rank compared to the students in the control group. In other words, after the treatment the learners in experimental group became more willing to communicate in English learners in control group. It means that the treatment positively influenced learners' WTC.

4. DISCUSSION and CONCLUSION

The results of the current study illustrate that simulated environment affected learners' WTC. To evaluate the WTC measure of learners, a WTC questioner was used. This questionnaire was developed by Macintyre, et al. (2001). The findings of pre-test indicated that the learners WTC level was almost equal. However, the post-test results illustrated that the difference between both groups is meaningful. In other words, learners become more willing to communicate in simulation environment classes. Simulation has significant effect on WTC of the learners. In the simulated environments, EFL learners get more willing to communicate with others. Simulation provides more chances for them to speak in target language. Through simulation, learners are encouraged to participate and be involved in class interactions and it is an opportunity to practice a full range of communication skills (Jones, 1982; Jones, 1983).

This finding is in line with a paper by Lyu (2006) which was entitled as "Simulation and Second/ Foreign Language learning: Improving communication skills through simulations" has witnessed the practical effects of language simulations in improving communication skills. In this research, many useful suggestions were shown on how simulations can be used in EFL class of basic level, intermediate level and advanced level, which was also implicitly indicated as the significance of the current study. Moreover, the results of this study showed that through learning in the light of the simulation technique, learners perceived their classroom activities joyfully. It inherently tends to promote real-life and authentic communication (Crookall et al., 1987; Crookall & Oxford, 1990; Nemitcheva, 1995), "enhances instrumental motivation by making the coursework more engaging" (Jones, 1986, p. 10), lowers affective barriers to acquisition by reducing the fear of making mistakes (Nemitcheva, 1995), and presents real time scenarios and instantaneous feedback (Jones, 1986).

The outcomes of this study also match up with the findings of researchers (Shankar et al., 2012), which reveal that this technique can expose students to different situations they are likely to face in their future career. From this approach, the students have a chance to explore different situations of real life that enable them to speak confidently and fluently in their second language.

In fact, this study states that learners were highly willing to communicate in simulated environment inside the classroom. They feel that they are in real-life situations, so they are eager to make communications. The reason is that Iranian EFL learners don't always have the chance to talk to some native speakers or travel to target-language countries. They can

communicate in English only in English classes. Therefore, it can be said that Iranian EFL learners get more willing to communicate in situations in which they experience communicating in their daily life. Creating such familiar environments for them inside the classrooms affect their willingness to express their ideas, thoughts and feeling in target language.

In simulated environment the learners become more comfortable and safer to speak in a foreign language. In fact, they psychologically make connections with the environment around them, which make it simple and accessible for them to make communication in target language. In formal and strict environment of some English classes, they feel unsecure and uncomfortable to speak in foreign language, however, by presenting such situations to them, they gradually feel they are in target language-speaking countries. Thus, they communicate with others and don't be afraid of speaking in a foreign language.

The main purpose of using simulation for EFL learners is to provide an environment where learners have ample opportunity for creating communication. Moreover, it provides a simulated of real-life situations in which learners experience real communications of real world. Since simulations focus on communication rather than language itself, they are real communicative activities. It can be concluded that simulation has significant effect on WTC of the learners. In the simulated environments, EFL learners get more willing to communicate with others. Simulation provides more chances for them to speak in target language. Through simulation, learners are encouraged to participate and be involved in class interactions.

The findings of the present study have important implications for second language learning research in general, and EFL teachers in particular. Considering the positive outcomes of this study, it is recommended that there is no denying the fact that this modern teaching methodology is quite instrumental and helpful in teaching English language as second or foreign language. The findings of this study suggest that understanding how a simulated environment affects EFL learners' success in speaking and communicating in foreign language can help institutes to provide such environments for EFL learners and instructors. The results of this study may be useful to the organizations of language testing and assessment. The organizations and institutes which administer testing exams for EFL learners can use the results of this study to use in speaking tests.

As pointed out earlier, simulation environment class is completely modern design. EFL learners, who are sick of traditional methods or dull atmosphere in classes, get motivated to learn English through this novel treatment. It promotes learners' engagement and enjoyment in learning. Results of this study, in accordance with previous studies, illustrate that using simulation help EFL learners shape their perceptions and conceptions toward learning English. Furthermore, the students experience being in simulated environment of the target country situations and they are provided many opportunities to engage in different communications. Some EFL learners always complain about the lack of opportunities or situations whereby they could use input that they have already learned. These classes are the best choice for these students.

This study has focused on only female students. The same study can be performed for male and female students. Moreover, the elementary participants were chosen to conduct this study. This method can be presented in different levels of English proficiency. The focus of this study was mainly on speaking skill. The same study can be conducted regarding other language skills (writing, listening and reading). This research was confined to the adjustment during the institute's semester. Research may be extended to a longer period to see the effects of simulated environment on the students in variety of contexts. Research may be conducted on the native students' WTC to see how the simulated environment could affect shy students. This study can be performed at high schools in order to evaluate the effects of simulated environment on their

English-speaking proficiency. Furthermore, this study could be done in a facilitated institute, in a room with blank walls and four video projectors. If future research compensates these items, they will access to more valid data. In this study one variable was considered, i.e., WTC; however, the effect of simulated environment could be investigated on different variables like non-verbal behaviors of learner, which demands specialized psychological investigations.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** University of Technology and Applied Sciences-Ibra, 22S41OIC

Authorship Contribution Statement

Houman Bijani analyzed the data and wrote the discussion and the conclusion of the study. **Masoumeh Abbasi** wrote the literature review and collected the data. Both authors finally reviewed and edited the final manuscript.

Orcid

Houman Bijani  <https://orcid.org/0000-0002-4305-7977>

Masoumeh Abbasi  <https://orcid.org/0000-0002-5310-0553>

REFERENCES

- Aleksandrak, M. (2011). Problems and challenges in teaching and learning speaking at advanced learning. *Glottodidactica*, 37(1), 37–48. <https://doi.org/10.14746/gl.2011.37.3>
- Amirian, Z., Karamifar, Z., & Youhanaee, M. (2020). Structural equation modeling of EFL learners' willingness to communicate and their cognitive and personality traits. *Applied Research on English Language*, 9(1), 103-136. <https://doi.org/10.22108/are.2019.116248.1451>
- Byrne, B.M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Lawrence Erlbaum Associates.
- Bystrom, K.E., Barfield, W., & Hendrix, C. (1999). A conceptual model of sense of presence in virtual environments. *Teleoperators and Virtual Environments*, 5(1), 109–121. <https://doi.org/10.1162/105474699566107>
- Clement, R., Baker, S.C., & MacIntyre, P.D. (2003). Willingness to communicate in a second language: The effect of context, norms, and vitality. *Journal of Language and Social Psychology*, 22(2), 190-209. <https://doi.org/10.1177/0261927X03022002003>
- Crookall, D., Greenblatt, C.S., Coote, A., Klabbers, J., & Watson, D. (Eds.) (1987). *Simulation-gaming in the late 1980s* (pp. 57-63). Pergamon.
- Crookall, D., & Oxford, R.L. (1990). *The island game*. In D. Crookall & R.L. Oxford (Eds.), *Simulation, gaming, and language learning* (pp. 251-259). Newbury House.
- Crookall, D., Coote, A., Dumas, D., & Le Gat, A. (1987). *The ISAGA GAME: Inquisitive speaking and gameful acquaintance: A mix of tongues and communicating across cultures*. In D. Crookall, C.S. Greenblatt, A. Coote, J. Klabbers & D. Watson (Eds.), *Simulation-gaming in the late 1980s* (pp. 57-63). Pergamon.
- Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Lawrence Erlbaum.
- Draper, J.V., Kaber, D. B., & Usher, J.M. (1998). Telepresence. *Human Factors*, 40(3), 354–375. <https://doi.org/10.1518/001872098779591386>
- Hashimoto, Y. (2002). Motivation and willingness to communicate as predictors of reported L2 use: The Japanese ESL context. *Second Language Studies*, 20(2), 29-70.
- Ho, R. (2006). *Handbook of univariate and multivariate data analysis and interpretation with SPSS*. Chapman & Hall/CRC.

- Jones, G. (1986). Computer simulations in language teaching-The kingdom experiment. *System*, 14(2), 171-178. [https://doi.org/10.1016/0346-251X\(86\)90007-2](https://doi.org/10.1016/0346-251X(86)90007-2)
- Jones, K. (1982). *Simulations in language teaching*. Cambridge University Press.
- Jones, L. (1983). *Eight simulations*. Cambridge University Press.
- Kemp, K.S. (2003). *Simulation and communicative language teaching in the Spanish II classroom*. Unpublished Master's Thesis. University of Toledo.
- Kiany, Gh.R., Mahdavy, B., & Ghafar Samar, K. (2013). Motivational changes of learners in a traditional context of English education: A case study of high school students in Iran. *International Journal of Research Studies in Language Learning*, 2(1), 3-16. <https://doi.org/10.5861/ijrsl.2012.92>
- Lazarton, A. (2001). *Teaching oral skills*. In M. Celce-Murcia (Ed.). *Teaching English as a second or foreign language* (pp. 103–115). Heinle & Heinle.
- Li, R.C., & Topolewski, D. (2002). ZIP & TERRY: A new attempt at designing language learning simulation. *Simulation & Gaming: An Interdisciplinary Journal*, 33(2), 181–186. <https://doi.org/10.1177/1046878102332006>
- Liang, M.Y. (2012). Foreign lucidity in online role-playing games. *Computer-Assisted Language Learning*, 25(5), 455–473. <https://doi.org/10.1080/09588221.2011.619988>
- Liu, M. & Jackson, J. (2008). An exploration of Chinese EFL learners' unwillingness to communicate and foreign language anxiety. *Modern Language Journal*, 92(1), 71–86. <https://doi.org/10.1111/j.1540-4781.2008.00687.x>
- Lyu, Y. (2006). *Simulations and Second / Foreign Language Learning: Improving communication skills through simulations*. (Electronic Thesis or Dissertation). Retrieved from <https://etd.ohiolink.edu/>
- MacIntyre, P.D. (1994). Variables underlying willingness to communicate: A causal analysis. *Communication Research Reports*, 11(2), 135-142. <https://doi.org/10.1080/08824099409359951>
- MacIntyre, P.D., & Charos, C. (1996). Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology*, 15(1), 3-26. <https://doi.org/10.1177/0261927X960151001>
- MacIntyre, P.D., & Clément, R. (1996). *A model of willingness to communicate in a second language: The concept, its antecedents, and implications*. Paper presented at the 11th World Congress of Applied Linguistics, Jyväskylä, Finland.
- MacIntyre, P.D., Baker, S.C., Clément, R., & Conrod, S. (2001). Willingness to communicate, social support, and language-learning orientations of immersion students. *Studies on Second Language Acquisition*, 23(3), 369-388. <https://doi.org/10.1017/S0272263101003035>
- MacIntyre, P.D., Dörnyei, Z., Clément, R., & Noels, K.A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, 82(4), 545-562. <https://doi.org/10.1111/j.1540-4781.1998.tb05543.x>
- Mallahi, O., & Hosseini, S. (2020). Predictors of performance of Iranian English language learners on speaking skill: A study of socially oriented personal attributes. *Iranian Journal of Learning and Memory*, 3(9), 75-86. <https://doi.org/10.22034/iepa.2020.237449.1181>
- McCroskey, J.C. (1992). Reliability and validity of the willingness to communicate scale. *Communication Quarterly*, 40(1), 16-25. <https://doi.org/10.1080/01463379209369817>
- McCroskey, J.C. (1997). *An introduction to rhetorical communication*, (7th ed.). Needham Heights, Allyn and Bacon.
- McCroskey, J.C., & Richmond, V.P. (1987). *Willingness to communicate and interpersonal communication*. In J.C. McCroskey & J.A. Daly (Eds.), *Personality and interpersonal communication* (pp. 129-159). Sage.

- Nemitcheva, N. (1995). *The Psychologist and games in the intensive foreign language game-based course*. In D. Crookall & K. Arai (Eds.), *Simulation and gaming across disciplines and Cultures* (pp. 70-74). Sage Publications.
- Peterson, M. (2011). Towards a research agenda for the use of three-dimensional virtual worlds in language learning. *CALICO Journal*, 29(1), 67-80. <https://doi.org/10.11139/cj.29.1.67-80>
- Richards, J.C. (1985). *The context of language learning*. Cambridge University Press.
- Richards, J.C., & Rodgers, T.S. (2001). *Approaches and methods in language teaching* (2nd ed.). Cambridge University Press.
- Richmond, V.P., & Roach, D.K. (1992). Willingness to communicate and employee success in U.S. organizations. *Journal of Applied Communication Research*, 20(1), 95-112. <https://doi.org/10.1080/00909889209365321>
- Saslow, J., & Ascher, A. (2006). *Top-Notch* (2nd ed.). Pearson Education, Inc.
- Shankar, P.R., Piryani, R.M., Singh, K.K., & Karki, B.M. (2012). Student feedback about the use of role plays in Sparshanam: A medical humanities module. *F1000Research*, 1(6), 1-11. <https://doi.org/10.12688/f1000research.1-65.v1>
- Sheybani, M. (2019). The relationship between EFL learners' willingness to communicate (WTC) and their teacher immediacy attributes: A structural equation modelling. *Cogent Psychology*, 6(1), 1-15. <https://doi.org/10.1080/23311908.2019.1607051>
- Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments: Speculations on the role of presence in virtual environments. *Teleoperators and Virtual Environments*, 6(6), 603-616. <https://doi.org/10.1162/pres.1997.6.6.603>
- Thorne, S.I., Black, R.W., & Sykes, J.M. (2009). Second language use, socialization, and learning in internet interest communities and online gaming. *The Modern Language Journal*, 93(1), 802-821. <https://doi.org/10.1111/j.1540-4781.2009.00974.x>
- Wang, H. (2005). *Comprehensible input in the real world and its implications for ESL teaching* [Unpublished Masters Thesis]. University of Toledo.
- Wang, Y.F., Petrina, S., & Feng, F. (2017). VILLAGE-Virtual immersive language learning and gaming environment: Immersion and presence. *British Journal of Educational Technology* 48(2), 431-450. <https://doi.org/10.1111/bjet.12388>
- Wen, W.P., & Clement, R. (2003). A Chinese conceptualisation of willingness to communicate in ESL. *Language, Culture and Curriculum*, 16(1), 18-38. <https://doi.org/10.1080/07908310308666654>
- Witmer, B.G., & Singer, M.J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7(3), 225-240. <https://doi.org/10.1162/105474698565686>
- Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *The Modern Language Journal*, 86(1), 54-66. <https://doi.org/10.1111/1540-4781.00136>
- Zarei, N., Saeidi, M., & Ahangari, S. (2019). Exploring EFL teachers' socio-affective and pedagogic strategies and students' willingness to communicate with a focus on Iranian culture. *Education Research International*, 9(1), 1-11. American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). <https://doi.org/10.1037/0000165-000>

Differential item functioning across gender with MIMIC modeling: PISA 2018 financial literacy items

Fatima Munevver Saaatcioglu ^{1,*}

¹Ankara Yildirim Beyazit University, Rectorate, Ankara, Türkiye

ARTICLE HISTORY

Received: Feb. 20, 2022

Revised: June 27, 2022

Accepted: July 04, 2022

Keywords:

Differential item functioning,
Latent class analysis,
Measurement invariance,
Mixture modeling,
PISA 2018.

Abstract: The aim of this study is to investigate the presence of DIF over the gender variable with the latent class modeling approach. The data were collected from 953 students who participated in the PISA 2018 8th-grade financial literacy assessment in the USA. Latent Class Analysis (LCA) approach was used to identify the latent classes, and the data fit the three-class model better in line with fit indices. In order to obtain more information about the characteristics of the emerging classes, uniform and non-uniform DIF sources were identified by using the Multiple Indicator Multiple Causes (MIMIC) model. The findings are very important in terms of contributing to the interpretation of latent classes. According to the results, the gender variable was a source of DIF for latent classes. It is important to include direct effects by gathering unbiased estimates for the measurement and structural parameters. Disregarding these effects can lead to incorrect identification of implicit classes. A sample application of MIMIC model was performed in a latent class framework with a stepwise approach in this study.

1. INTRODUCTION

One of the basic aims of measurement studies is to develop and construct valid items measuring latent variable. In many studies, Differential Item Functioning (DIF) can be a threat to the validity of a test or a scale. DIF concept is defined that the situation in which “different groups of test takers with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item” in AERA & NCME (2014). Definition of two types of DIF called uniform and non-uniform DIF was emphasized in the literature (Ackerman, 1992; Mellenbergh, 1982; Millsap & Everson, 1993; Swaminathan & Rogers, 1990). Uniform DIF occurs while students in one group consistently have a better chance of giving a correct answer than those with the similar ability level in another group. If the relationship is not consistent, in this situation non-uniform DIF occurs (Swaminathan & Rogers, 1990).

The concept of DIF is directly related to the concept of fairness and bias. Fairness means that for different groups of students, inferences made according to test scores are valid (ETS, 2019). Thus, fairness in the test is related to bias. If a fair test is applied, students with the similar level of competence have the similar probability of answering an item correctly. Therefore, items

*CONTACT: Fatima Munevver Saaatcioglu ✉ fmsaatcioglu@ybu.edu.tr 📍 Ankara Yildirim Beyazit University, Rectorate, Ankara, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

having DIF cause bias, which is a problem in psychological and educational testing. Bias is concerned to construct-irrelevant factors such as education, gender, culture, age, although students have the same trait or ability (Lee & Zhang, 2017; Messick, 1989). The construct is accepted as the test of interest and can explain the variance of student's performance in a test. According to Messick (1989), construct-irrelevant variance refers to variables unrelated to the construct, and it can occur when the test scores are affected by factors that are irrelevant to the construct. Test preparation, test development and administration, scoring, students' background knowledge, personality, answering strategies, and cognitive ability can be construct-irrelevant, and efforts are needed to minimize such effects (Gallagher et al., 2002; Haladyna & Downing, 2004). In addition, The Standards state that any bias causing students' scores in systematically high or low is construct-irrelevant variance (AERA & NCME, 2014).

Studies to examine item and test bias are based on two fundamental perspectives in measurement theory. First, from the Classic Test Theory (CTT) perspective, the Multi-group Confirmatory Factor Analysis (MG-CFA) method is tested for the relationship between observed variable and the latent trait for measurement invariance across groups. The second one is evaluated according to whether the ability levels of individuals in separate groups are equivalent on substance behavior with the Item Response Theory (IRT) and DIF approach (Embretson & Reise, 2000). In IRT framework for detecting DIF; differences in the probability to reply the item correctly for two groups are taken into account. Therefore, IRT methods focus on comparing item parameters of the groups i.e item characteristic curves (Thissen et al., 1993). In DIF studies that were conducted according to manifest grouping approach, assume that the groups come from a homogenous subgroups, and this homogeneity means that items do not have DIF within the subgroups (De Ayala, et al., 2002). Latent classes can occur whether all students do not have homogeneous response patterns (De Ayala et al., 2002; Samuelsen, 2008). On the other hand, it is debated that DIF results obtained from groups may be biased (Rupp & Zumbo, 2006). Hence, it is proposed to use mixture models that reject the homogeneity of the data for DIF in latent classes. Mixture models consider a mixture of latent classes to compose the sample (Mislevy & Verhelst, 1990; Rost, 1990; De Ayala & Santiago, 2017). According to this mixture modeling approach, invariance assumption is no longer essential, and thus, item parameters are estimated for each latent class (Cho, 2007; Cohen & Bolt, 2005; De Mars & Lau, 2011; Oliveri, et al., 2013; Park et al., 2016; Rupp & Zumbo). Thanks to these studies, , DIF studies should be examined between latent classes. The MIMIC modeling is used by researchers within mixture modeling approach to explore the latent classes (Nylund-Gibson & Choi, 2018). In these studies, the researchers examined the effect of covariates on latent class variable. With this perspective, the direct effects can be examined from covariates to items determining possible sources of DIF, which is called MIMIC modeling (Masyn, 2017). Moreover, it can be examined if the identified latent classes are invariant when the students in a class have the similar responses (Kankaraš et al., 2011).

The MIMIC model can be defined as a form of Structural Equation Modeling (SEM). The model combines covariates into a CFA model. MIMIC model includes a measurement model enabling to detect the link within latent variable and items, and also a structural model bringing out the direct effect of a covariate. There are studies stating that MIMIC models are more useful compared to other techniques such as multigroup CFA in examining DIF (Vandenberg & Lance, 2000; Millsap, 2011). MIMIC modeling contributes to external validity by examining the relationship between covariate and latent structure, and to internal validity by estimating IRT parameters (Tsaousis et al., 2020).

MIMIC modeling allows us to see the effect of covariates. In addition, estimates can be obtained from all other grouping variables (covariates) in the model (Asparouhov & Muthén, 2015; Cheng et al., 2016). These variables can be observed or latent, and they can also be categorical

or continuous (Glockner-Rist & Hoijtink, 2003). These flexibilities support the MIMIC model for DIF studies.

Next, in international large scale assessments like Program for International Student Assessment (PISA; OECD, 2019a) and Trends in International Mathematics and Science Study (TIMSS; IEA, 2017a) DIF analysis requires having the scores that are fairly comparable across countries. In international large scale assessments, IRT models are used to estimate item parameters. However, invariance assumption of the IRT models cannot be met in a heterogenous population which contains latent classes. Hence, the aim of this study is to investigate the presence of DIF over the gender variable with a stepwise procedure conducted with a MIMIC modeling framework that has been developed by Masyn, (2017). The MIMIC approach is a method to test measurement invariance, and since its introduction (Masyn, 2017), a study conducted with real data by Tsaousis et al., (2020) but there is no study with international large scale assessment data in which this method was used. Consequently, in this study, following the stepwise procedure outlined by Masyn (2017), to explore sources of DIF over gender using large scale assessment data (i.e., PISA 2018 financial literacy test). The results of this study are expected to have vital implications for measurement research by examining DIF between latent classes.

2. METHOD

2.1. Data

PISA is an international survey assessing competency of 15-year-old students in the basic domains of reading, mathematics and science literacy. PISA was first administered in 2000, and it cycles every three years. In each cycle, one of the basic domains is specified as the major domain, which is administered to all participants. The other domains are considered minor domains which are not administered to all participants. In addition, financial literacy was added in the PISA 2012 assessment, and has been provided as an international choice in the two PISA assessments (2015 and 2018). Financial literacy categories are money and transactions, planning and managing finance, risk and rewards, and an understanding of the financial landscape. These categories are measured by several open-ended and multiple-choice items (OECD, 2019a). In this study, 16 multiple choice items were used to detect uniform and non-uniform DIF items in Booklet 6. This booklet was used in the analysis because the number of items in the 6th booklet is more than the others in the booklets.

2.2. Sample

In PISA 2018, a total of 20 countries participated in financial literacy testing, including 13 OECD countries and seven partner countries. Since the purpose of this research is to show an application of the MIMIC model, the sample of this study includes 953 students from the USA who replied booklet 6. This country was chosen because the aim of this study is to show an application of the DIF study with MIMIC modeling and the sample size of the USA data is large. For the USA sample, 479 (50.3%) were females and 474 (49.7%) were males.

2.3. Data Analysis

The stepwise procedure has been developed by Masyn (2017), and in this study, the original source are used, and models are shown in Figure 1 with diagrams for each step. First LCA was carried out to determine the number of latent classes. A procedure based on comparing the fit of models that have different numbers of latent classes and using model fit information criteria is applied in LCA. In simulation studies, it has been found that BIC outperforms in determining the number of latent classes (Nylund et al, 2007). In addition, sample-adjusted BIC is among the recommended indexes for consistent AIC (CAIC; Bozdogan, 1987) model fit. Next, VLMR and BLRT tests results are interpreted. LCA is performed with the Mplus (Muthén and Muthén,

1998-2021) software program using Robust Maximum-Likelihood (MLR) and expectation maximization (EM) algorithm as an estimation method.

This first step, i.e. Step 0 includes the process of deciding on the number of classes by finding the model that best fits the data with an exploratory approach. Considering the model fit indices, the number of latent classes are identified taking the covariate as an auxiliary variable so that it does not have any effect on the determining latent classes (Masyn, 2013; Nylund-Gibson & Masyn, 2016). Only class indicators are included in the model as observed variables.

In Step 1, two models are compared. The first model, called A1.0, contains only the regression of the covariate over the latent class variable, which evaluates the model-fit of a non-DIF model. This model is compared to a model (non-uniform DIF) that items and latent variable are regressed to the covariate (A1.1) model in which the effects of the covariate on the items are released to differ between classes. The models compared with likelihood ratio test should supply proof on behalf of the A1.1 model as compared to A1.0 model, in the presence of DIF. If A1.0 model is the chosen model, there is no significant proof of DIF owing to the covariate. However, the choice of the A1.1 model requires further examination on the location of the invariance due to the covariate.

In Step 2, the purpose is to evaluate the presence of non-uniform DIF running models to detect the effects of the covariate on items. The models involve a no-DIF model (A2.0.1) that the latent class is regressing on the covariate and DIF model that an item regressing on the covariate A2.1.1 model from the first item to the last item. In model comparison, the likelihood ratio difference tests were utilized. Proof on behalf of the later model indicated the existence of non-uniform DIF.

In Step 3, the purpose is to select most parsimonious non-uniform DIF model (A3.0). This model helps to estimate a latent class model including non-uniform paths in which statistically significant. This model (A3.0) is first compared no-DIF model (A1.0) with the prospect that A3.0 would be excellent to model A1.0. The next comparison was between A3.0 and A1.1 (the all DIF model) with the expectancy that A3.0 would be no worse than A1.1.

In Step 4, we test hypothesis that non-uniform DIF effects items do not indicate uniform effects. Nonsignificant differences between models A3.0 and A4.1–A4.5, show proof of non-uniform DIF effects. Analyses were conducted with Mplus software (Muthén and Muthén, 1998-2019). The syntax codes for analyses can be found in the [Appendix](#). In [Figure 1](#), the model diagrams as stepwise procedure is given.

2.4. Effect Size

Several studies investigated the effect size metrics for DIF (Raju, 1990; Penfield & Lam, 2000; Zwick, 2012) and among them, the most considerable are the ETS criteria, transforming the difference logit parameter onto the delta metric system (Dorans & Holland, 1993). The Educational Testing Service (ETS) defined a three-level category sizes of DIF that are negligible, medium and large. For the negligible DIF level, the size of DIF should be 0.43 and below; for medium DIF, the size of the DIF should be ≥ 0.44 and for the large DIF, the size of the DIF should be ≥ 0.64 on logit scale (Lin & Lin, 2014).

Figure 1. Stepwise procedure for DIF detection using mixture modeling.

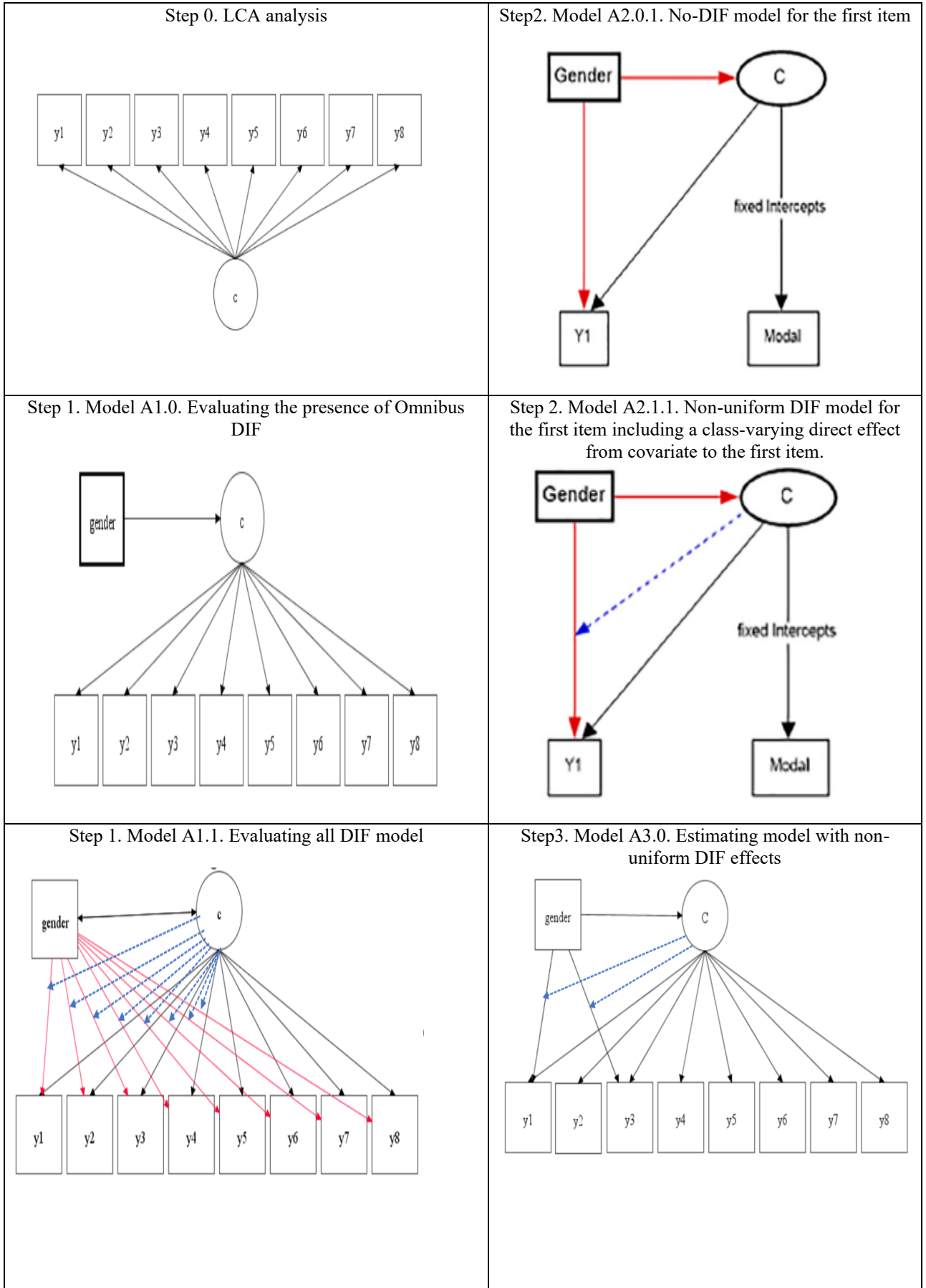
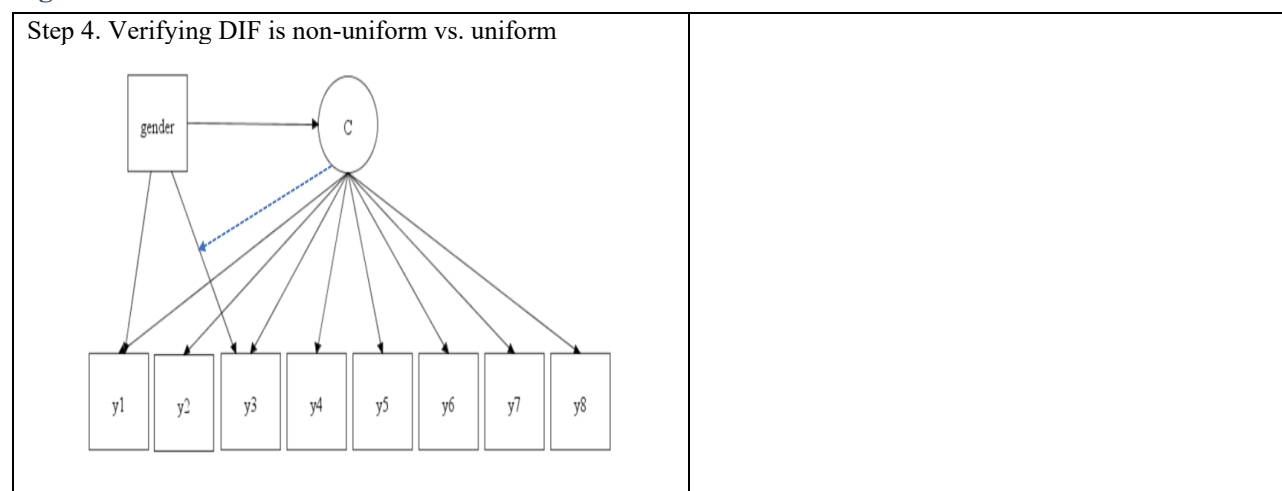


Figure 1. *Continues*

3. RESULTS

3.1. Step 0

In this step, 1, 2, 3 and 4-class models were tested, respectively, and the model fit indices were presented in Table 1.

Table 1. *Fit indices of models tested for data from the USA.*

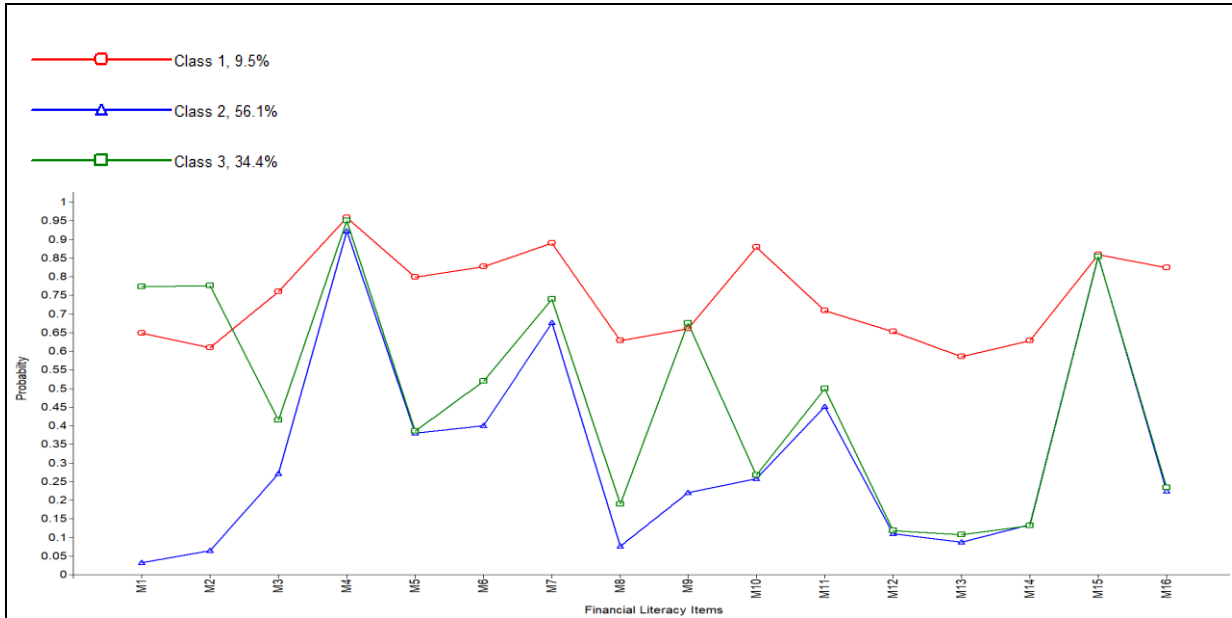
Fit indices	1-class	2-class	3-class	4-class
npar	16	33	50	67
LL	-8292.439	-7961.477	-7727.144	-7814.792
AIC	16616.878	15988.954	15588.287	15729.585
BIC	16694.598	16149.251	15913.741	15972.460
SSA-BIC	16643.783	16044.445	15700.952	15813.663
LR Chi-Square Test	2825.081	2463.731	2440.565	2275.961
LR Chi-Square <i>p</i> value	1.0000	1.0000	1.0000	1.0000
VLMR Test	-	656.295	293.369	175.297
VLMR <i>p</i> value	-	0.0000	0.0000	0.0649
BLRT Test	-	656.295	293.369	175.297
BLRT <i>p</i> value	-	0.0000	0.0000	0.0000

npar, number of free parameters; LL, log likelihood; AIC, Akaike's information criterion; BIC, Bayesian information criterion; SA-BIC, Sample-Size Adjusted BIC; VLMR, Vuong-lo-mendell-rubin test; BLRT, Bootstrapped likelihood ratio test; $p < 0.05$.

When the fit indexes were examined in the Table 1, the LR Chi-Square test, has an insignificant p value for data from the USA, showing that the model data fit was achieved. When the AIC, BIC and SSA-BIC values of the relative fit indices were examined, the three-class model had the lowest values among other models. The results of VLMR and BLRT showed a statistically significant difference between the 2-class and 3-class models. When the 3-class model was compared with the 4-class model, it was observed that the p value was not significant ($p > 0.05$). This finding means that to add one more class to the 3-class model does not improve the model-data fit. As a result, it was seen that a three-class model was fit to the data. The model classified 9.5% students into Class 1 which had high probability of ability, 56.1 % subjects into Class 2 which had moderate probability of ability and 34.4% in Class 3 with low probability of ability.

The value of classification accuracy was 0.78. It can be stated that the three-class model is useful in assigning students to the correct classes as the entropy value was obtained to be greater than .70 (Nagin, 2005). The graph of the results obtained for these three classes were presented in Figure 2.

Figure 2. Latent class profile plots for PISA 2018 financial literacy subtest.



After determining the best-fitting model, it was examined for DIF effects on the financial literacy items, MIMIC modeling results were given step by step.

3.2. Step 1

In this step, Null model (A1.0) assuming no DIF and an alternate model (A1.1) assuming DIF for all items were compared. The results from the likelihood ratio test statistics (LRTS) indicated that gender was a source of DIF rejecting the null model A1.0 (LRTS = 130.40, $df = 48$, $p = 0.0001$). Therefore, the analyses in Step 2 were performed to explore the item-level effects of the DIF source based on the result from the omnibus DIF finding.

3.3. Step 2

In this step, null model (A2.0.1- no DIF model) and an alternate model (A2.1.1 non-uniform DIF model) for a specific item were compared. The results obtained from LRTS were presented in Table 2. According to results for five items (9, 12, 13, 15, 16), the no DIF model was rejected on behalf of the alternate model. These results mean that the non-uniform DIF items can be differentiated over gender.

3.4. Step 3

In this step, A3.0 model constructed from items displaying non-uniform DIF from step 2. A3.0 model was compared to model A1.0 (no-DIF), and the latter showed fit (LRTS = 56.634, $df = 15$, $p = 0.0001$). When A3.0 model was compared to A1.1 model (all DIF), significant differences were found (LRTS = 73.766, $df = 33$, $p = 0.0001$). In addition, the BIC values for A3.0 model, it was 16054.657, and for A1.1 model, it was 16207.187. Thus, the model A3.0 that has lower BIC value was the preferred model compared to A1.1 model. The results were presented in Step 3 part of Table 2. Finally, results from this step recommended that A3.0 model was the last latent class MIMIC model.

Table 2. Model comparisons for DIF by stepwise procedure.

Steps	Model	Model description	LogL	npar	Model Comparison	LRTS	df	p-value
1	A1.0	MIMIC: No DIF	-7815.632	55	A1.0 vs. A1.1	130.40	48	0.0001
	A1.1	MIMIC: All DIF	-7750.432	103				
2	A2.0.1	Item1: No DIF	-1128.703	7	A2.0.1 vs. A2.1.1	3.408	3	Ns
	A2.1.1	Item1: NON-U DIF	-1126.99	10				
	A2.0.2	Item2: No DIF	-1202.695	7	A2.0.2 vs. A2.1.2	6.872	3	Ns
	A2.1.2	Item2: NON-U DIF	-1199.259	10				
	A3.0.3	Item3: No DIF	-1434.869	7	A2.0.3 vs. A2.1.3	2.94	3	Ns
	A3.1.3	Item3: NON-U DIF	-1434.722	10				
	A4.0.4	Item4: No DIF	-1081.621	7	A2.0.4 vs. A2.1.4	1.910	3	Ns
	A4.1.4	Item4: NON-U DIF	-1080.666	10				
	A5.0.5	Item5: No DIF	-1459.193	7	A2.0.5 vs. A2.1.5	1.644	3	Ns
	A5.1.1	Item5: NON-U DIF	-1458.371	10				
	A6.0.6	Item6: No DIF	-1476.005	7	A2.0.6 vs. A2.1.6	0.848	3	Ns
	A6.1.6	Item6: NON-U DIF	-1475.581	10				
	A7.0.7	Item7: No DIF	-1405.517	7	A2.0.7 vs. A2.1.7	1.216	3	Ns
	A7.1.7	Item7: NON-U DIF	-1404.909	10				
	A8.0.8	Item8: No DIF	-1205.034	7	A2.0.8 vs. A2.1.8	0.892	3	Ns
	A8.1.8	Item8: NON-U DIF	-1204.588	10				
A9.0.9	Item9: No DIF	-1382.867	7	A2.0.9 vs. A2.1.9	10.658	3	0.01	
A9.1.9	Item9: NON-U DIF	-1377.538	10					
A10.0.10	Item10: No DIF	-1379.947	7	A2.0.10 vs. A2.1.10	3.460	3	Ns	
A10.1.10	Item10: NON-U DIF	-1378.217	10					
A11.0.11	Item11: No DIF	-1486.230	7	A2.0.11 vs. A2.1.11	5.94	3	Ns	
A11.1.11	Item11: NON-U DIF	1483.998	10					
A12.0.12	Item12: No DIF	-1202.668	7	A2.0.12 vs. A2.1.12	10.792	3	0.01	
A12.1.12	Item12: NON-U DIF	-1197.272	10					
A13.0.13	Item13: No DIF	-1176.637	7	A2.0.13 vs. A2.1.13	10.180	3	0.01	
A13.1.13	Item13: NON-U DIF	-1171.547	10					
A14.0.14	Item14: No DIF	-1242.057	7	A2.0.14 vs. A2.1.14	1.142	3	Ns	
A14.1.14	Item14: NON-U DIF	-1241.486	10					
A15.0.15	Item15: No DIF	-1238.851	7	A2.0.15 vs. A2.1.15	9.382	3	0.02	
A15.1.15	Item15: NON-U DIF	-1234.160	10					
A16.0.16	Item16: No DIF	-1343.979	7	A2.0.16 vs. A2.1.16	8.698	3	0.03	
A16.1.16	Item16: NON-U DIF	-1339.630	10					
3	A3.0	MIMIC with all NON-U DIF items	-7787.315	70	A1.0 vs. A3.0	56.634	15	0.0001
					A3.0 vs. A1.1	73.766	33	0.0001
4	A4.1	Item9 (U- DIF) all other (NON-U DIF)	-7791.601	67	A4.1 vs. A3.0	8.572	3	0.035
	A4.2	Item12 (U-DIF) all other (NON-U DIF)	-7796.839	67	A4.2 vs. A3.0	19.048	3	0.0001
	A4.3	Item13 (U- DIF) all other (NON-U DIF)	-7795.662	67	A4.3 vs. A3.0	16.694	3	0.0001
	A4.4	Item15 (U- DIF) all other (NON-U DIF)	-7792.602	67	A4.4 vs. A3.0	10.574	3	0.014
	A4.5	Item16 (U -DIF) all other (NON-U DIF)	-7792.077	67	A4.5 vs. A3.0	9.524	3	0.023

LL: log likelihood; *df*.: degrees of freedom; LRTS: likelihood ratio test statistic, ;npar: number of free parameters, UN-DIF: uniform DIF, NON-U DIF: Non-uniform DIF, Ns: not significant; $p < 0.05$.

3.5. Step 4

In this step, MIMIC models (A4.1-A4.5) including the items which displayed non-uniform DIF. In these models, all other direct effects were allowed to free all across classes but the direct effect to each item was constrained to be invariant. Hence, each model (A4.1-A4.5) was compared with the non-uniform DIF model (A3.0 model). According to the results, it was found

that models were statistically worse than A3.0 model, and DIF effects were non-uniform DIF (items 9, 12, 13, 15, 16).

Table 3. Statistics for non-uniform DIF items over gender for PISA 2018 financial literacy subtest.

Item no	Latent Class 1			Latent Class 2			Latent Class 3		
	Estimates	95% CIs (UL/LL)	Effect size	Estimates	95% CIs (UL/LL)	Effect size	Estimates	95% CIs (UL/LL)	Effect size
9	1.315	0.927/14.965	Large	-0.298	0.451/1.222	Negligible	-0.433	0.371/1.135	Medium
12	1.657	0.397/69.242	Large	1.080	1.428/6.076	Large	0.179	0.541/2.648	Negligible
13	0.978	0.496/14.269	Large	1.326	1.367/10.377	Large	0.206	0.523/2.884	Negligible
15	-1.216	0.011/7.669	Large	-0.598	0.321/0.943	Medium	-0.565	0.283/1.141	Medium
16	0.618	0.231/14.910	Medium	0.268	0.826/2.070	Negligible	0.748	1.156/3.860	Large

UL: upper level; LL: low level.

Table 3 presents estimates in logits, 95% CIs and effect size values for each classes. According to ETS criteria, the size of DIF effects were interpreted (Lin & Lin, 2014). For Class 1 (high performing) item 9, 12, 13 and 15 exhibited large level DIF, and item 16 showed medium level DIF over gender. Moreover, males scored higher than females (positive values mean that males have higher values) on all items except item 15. For Class 2 (average performing), the DIF effect was negligible for item 9 and 16; item 12 and 13 showed large level DIF, and item 16 showed medium level DIF over gender. Also males scored higher than females on item 12 and 13. For Class 3 (low performing), the DIF effect was negligible for item 12 and 13, and it was medium for item 9 and 15 with males scoring higher than females; and it was large for item 16 with females scoring higher than males.

4. DISCUSSION and CONCLUSION

The aim of this study was to investigate the presence of DIF over the gender variable with a MIMIC modeling including a stepwise procedure (Masyn, 2017). In the first step, a LCA was conducted to detect group of heterogeneity. According to the indices, data fit the three-class model better. The model classified 9.5% of the students into Class 1 (high performing), 56.1 % of the students into Class 2 (average performing) and 34.4% in Class 3 (low performing).

In addition to the above classification of the students into the three classes, this analysis could provide further information about the specific items that performed across the different classes. For example, item 4 was an easy item and had a high probability of ability for each class. A similar pattern was observed with item 7 and 15. Also item 5 and 10 seem to be difficult items that discriminate Class 1 (high performing) from the Class 2 (average performing) and Class 3 (low performing) but not differentiate Class 2 and Class 3 (average and low performing). It can be seen that the majority of the items differentiate students across classes.

Then, it was investigated if there had been direct effects from the latent class to items. Thus, DIF test was conducted by comparing no DIF model with all-DIF model considering no DIF model was statistically worse than all DIF model. So it can be stated that gender is a source of DIF. This is an important result showing that gender should be added in the regression model. Studies reveal that ignoring the effects of covariates may lead to misspecifications for the latent classes (Asparouhov & Muthén, 2014; Clark & Muthén, 2009; Masyn, 2017).

Next, uniform and non-uniform DIF effects were investigated for financial literacy items. According to the results, five items displayed non-uniform DIF with significant p values. Next, the effect size of non-uniform DIF items was investigated over gender. For Class 1, item 9, 12, 13 and 15 exhibited large level DIF effect, and item 16 exhibited medium level DIF effect, when males scored higher. For Class 2, the DIF effect was negligible for item 9 and 16; item 12 and 13 exhibited large level DIF effect, and item 16 exhibited medium level DIF effect over gender. Furthermore, males also scored higher than females on item 12 and 13. For Class 3, the DIF effect was negligible for item 12 and 13, medium for item 9 and 15 with males who had higher scores; large for item 16 with females who had higher scores.

This study showed that what may be the cause for DIF in a latent class framework. Ignoring DIF effects in LCA could lead to the misinterpretation of the analysis and getting biased estimates in identifying classes and estimating relationship between latent class variable and covariate. Previous studies have shown that ignoring these effects can lead to biased estimated parameters for both measurement and structural model of the latent class analysis, and in this situation latent classes cannot be used for class comparisons (Clark & Muthén, 2009; Masyn, 2017; Nylund-Gibson & Choi, 2018). The results showed that MIMIC modeling was an essential procedure to find items displaying DIF effects between females and males. Thus, the nature of latent classes may be investigated by considering in each latent class membership. In addition in LCA, direct effects examinations must be a standard procedure to investigate direct effects of covariates on latent class indicators.

This study also revealed that response probabilities across latent classes were not the same for all latent class indicators. In this context, students within a class could have different response probabilities depending on a specific characteristic (in terms of gender). Hence, it can be pointed out that assuming that all latent class indicators have the same expected responses across classes and across different levels of a demographic variable can lead to the misinterpretation of latent classes. Thus, identifying latent classes by inspecting the manifest characteristics in each latent class membership is so important to have the right information about classes.

Throughout this article, the analyses were conducted on logit scale in Mplus, and the effect sizes were interpreted according to logit scale. The MIMIC model can be conducted for logistic or normal-ogive link functions. Thus, analysis can be run on probit link.

MIMIC modeling contributes to external validity by examining the relationship between covariate and latent structure, and to internal validity by estimating the parameters. Contributing to validity studies, other demographic variables can be included in the analysis. Next, various distal outcomes could be used to detect latent classes displaying statistically significant differences.

Continuous or categorical variables and the mixture of both can be used in MIMIC model approach. In this study, dichotomous variables were used. Future studies can be conducted with continuous variables, and the models can be compared with information criteria like SRMR, TLI, CFI etc. model fit statistics (Kang & Cohen, 2007).

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

F. Munevver Saaticioglu  <https://orcid.org/0000-0003-4797-207X>

REFERENCES

- Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67-91. <https://doi.org/10.1111/j.1745-3984.1992.tb00368.x>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370. <https://doi.org/10.1007/BF02294361>
- Cheng, Y., Shao, C., & Lathrop, Q.N. (2016). The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement*, 76(1), 43-63. <https://doi.org/10.1177/0013164415576187>
- Cho, S.J. (2007). *A multilevel mixture IRT model for DIF analysis* [Unpublished doctoral dissertation]. University of Georgia: Athens.
- Choi, Y., Alexeev, N., & Cohen, A.S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing*, 15(3), 239-253. <https://doi.org/10.1080/15305058.2015.1007241>
- Clark, S.L., & Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. Available online at: <http://www.statmodel.com/download/relatinglca.pdf>
- Cohen, A.S., & Bolt, D.M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42(2), 133–148. <https://doi.org/10.1111/j.1745-3984.2005.00007>
- De Ayala, R.J., Kim, S.H., Stapleton, L.M., & Dayton, C.M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, 2(3-4), 243-276. <https://doi.org/10.1080/15305058.2002.9669495>
- De Ayala, R.J., & Santiago, S.Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, 60(1), 25-40. <https://doi.org/10.1016/j.jsp.2016.01.002>
- Educational Testing Service. (2019). Standards for Quality and Fairness. Retrieved from <https://www.ets.org/s/about/pdf/standards.pdf>
- Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Erlbaum.
- Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel–Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, 29, 278–295. <https://doi.org/10.1177/0146621605275728>
- Finch, W.H., & French, B.F. (2012). Parameter estimation with mixture item response theory models: A Monte Carlo comparison of maximum likelihood and Bayesian methods. *Journal of Modern Applied Statistical Methods*, 11(1), 14.
- Gallagher, A., Bennett, R.E., Cahalan, C., & Rock, D.A. (2002). Validity and fairness in technology-based assessment: detecting construct-irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment*, 8(1), 27-41. https://doi.org/10.1207/S15326977EA0801_02
- Glockner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, 10(4), 544-565. https://doi.org/10.1207/S15328007SEM1004_4
- Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Assessment Issues and Practice*, 23 (1), 17-27. <https://doi.org/10.1111/j.1745-3992.2004.tb00149.x>

- IEA. (2017a). *TIMSS 2015 user guide for the international database*. Chestnut, MA: Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement (IEA).
- Kang, T., & Cohen, A.S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*(1), 331-358. <https://doi.org/10.1177%2F0146621606292213>
- Kankaraš, M., Moors, G., & Vermunt, J.K. (2011). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, J. Billiet, & B. Mueleman (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 359–384). Routledge.
- Lee, Y., & Zhang, J. (2017). Effects of differential item functioning on examinees' test performance and reliability of test. *International Journal of Testing, 17*(1), 23–54. <https://doi.org/10.1080/15305058.2016.1224888>
- Lin, P.-Y., and Lin, Y.-C. (2014). Examining student factors in sources of setting accommodation DIF. *Educational and Psychological Measurement 74*(1), 759–794. <https://doi.org/10.1177%2F0013164413514053>
- Masyn, K. (2013). "Latent class analysis and finite mixture modeling," in *The Oxford handbook of quantitative methods in psychology*, Vol. 2, ed. T. D. Little (Oxford University Press), 551–611.
- Masyn, K. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2), 180-197. <https://doi.org/10.1080/10705511.2016.1254049>
- Messick, S. (1989). "Validity," in *Educational Measurement*. Editor R. L. Linn 3rd ed. (NewYork: American Councilon Education and Macmillan), 13–103.
- Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. Taylor & Francis.
- Mislevy, R.J., & Verhelst, N.D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(1), 195-215.
- Nylund, K.L., Asparouhov, T., & Muthén, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal, 14*(4), 535-569. <https://doi.org/10.1080/10705510701575396>
- Nagin, D. (2005). *Group-based modeling of development*. Harvard University Press.
- Nylund-Gibson, K., & Masyn, K.E. (2016). Covariates and mixture modeling: results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling, 23*(1), 782-797. <https://doi.org/10.1080/10705511.2016.1221313>
- OECD. (2019a). *PISA 2018 assessment and analytical framework*. PISA, OECD Publishing.
- OECD. (2019b). *Technical report of the Survey of Adult Skills (PIAAC) (3rd Edition)*. OECD Publishing.
- Oliveri, M.E., & von Davier, M. (2017). Analyzing the invariance of item parameters used to estimate trends in international large-scale assessments. In H. Jiao & R. W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 121–146). Information Age Publishing, Inc.
- Oliveri, M., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing, 13*(3), 272–293. <https://doi.org/10.1080/15305058.2012.738266>
- Oliveri, M.E., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education, 24*(4), 349-366. <https://doi.org/10.1080/08957347.2011.607063>

- Penfield, R.D., & Lam, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, 19(1), 5–15. <https://doi.org/10.1111/j.1745-3992.2000.tb00033.x>
- Raju, N. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14(1), 197–207. <https://doi.org/10.1177/014662169001400208>
- Rost, J. (1990). Rasch Models in Latent Classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271–282.
- Samuelsen, K.M. (2008). Examining differential item functioning from a latent mixture perspective. In Hancock, G.R., & Samuelsen, K.M. (Eds.) *Advances in latent variable mixture models*, Information Age.
- Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370.
- Tsaousis, I., Sideridis, G.D., & AlGhamdi, H.M. (2020). Measurement invariance and differential item functioning across gender within a latent class analysis framework: evidence from a high-stakes test for university admission in Saudi Arabia. *Frontiers in Psychology*, 11, 1-13. <https://doi.org/10.3389/fpsyg.2020.00622>
- Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-69. <https://doi.org/10.1177/109442810031002>
- Yun, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* [Unpublished doctoral dissertation]. University of California, Los Angeles.
- Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, 2012(1), i-30. <http://dx.doi.org/10.1002/j.2333-8504.2012.tb02290.x>

APPENDIX
A.0

TITLE: Stepwise MIMIC Model DIF Detection

DATA: file is b16-gender.dat;

VARIABLE:

NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

MISSING ARE ALL (99);

Auxiliary = gender;

USEVARIABLES ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15
m16;

CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

CLASSES = c(3);

ANALYSIS:

type=mixture;

MODEL:

%OVERALL%

[c#1*-1.291];

[c#2* 0.489];

c on gender;

%C#1%

[m1\$1* -0.612];

[m2\$1* -0.444];

[m3\$1* -1.149];

[m4\$1* -3.141];

[m5\$1* -1.383];

[m6\$1* -1.561];

[m7\$1* -2.091];

[m8\$1* -0.524];

[m9\$1* -0.668];

[m10\$1* -1.982];

[m11\$1* -0.891];

[m12\$1* -0.631];

[m13\$1* -0.345];

[m14\$1* -0.524];

[m15\$1* -1.815];

[m16\$1* -1.545];

%C#2%

[m1\$1* 3.386];

[m2\$1* 2.674];

[m3\$1* 0.990];

[m4\$1* -2.473];

[m5\$1* 0.490];

[m6\$1* 0.404];

[m7\$1* -0.740];

[m8\$1* 2.497];

[m9\$1* 1.261];

[m10\$1* 1.058];

```
[ m11$1* 0.194 ];
[ m12$1* 2.084 ];
[ m13$1* 2.346 ];
[ m14$1* 1.864 ];
[ m15$1* -1.777 ];
[ m16$1* 1.240 ];
```

%C#3%

```
[ m1$1* -1.230 ];
[ m2$1* -1.240 ];
[ m3$1* 0.344 ];
[ m4$1* -2.971 ];
[ m5$1* 0.469 ];
[ m6$1* -0.079 ];
[ m7$1* -1.043 ];
[ m8$1* 1.456 ];
[ m9$1* -0.730 ];
[ m10$1* 1.008 ];
[ m11$1* 0.004 ];
[ m12$1* 2.004 ];
[ m13$1* 2.112 ];
[ m14$1* 1.885 ];
[ m15$1* -1.763 ];
[ m16$1* 1.186 ];
```

OUTPUT:

TECH1 TECH8;

PLOT: type=plot3;

series = m1 (1) m2 (2) m3 (3) m4 (4) m5 (5) m6 (6) m7 (7) m8 (8)

m9 (9) m10 (10) m11 (11) m12 (12) m13 (13) m14 (14) m15 (15) m16 (16);

! how the variables are presented in the X axis

! (*) separate them by a space

SAVEDATA:

file = data_savedata.txt;

save = cprob;

missflag = 9999;

format = free;

A1.0

TITLE: Stepwise MIMIC Model DIF Detection

DATA: file is b16-gender.dat;

VARIABLE:

NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

MISSING ARE ALL (99);

USEVARIABLES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14
m15 m16;

CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

CLASSES = c(3);

ANALYSIS:

type=mixture;

MODEL:

%OVERALL%

[c#1*-1.291];

[c#2* 0.489];

c on gender;

%C#1%

[m1\$1* -0.612];

[m2\$1* -0.444];

[m3\$1* -1.149];

[m4\$1* -3.141];

[m5\$1* -1.383];

[m6\$1* -1.561];

[m7\$1* -2.091];

[m8\$1* -0.524];

[m9\$1* -0.668];

[m10\$1* -1.982];

[m11\$1* -0.891];

[m12\$1* -0.631];

[m13\$1* -0.345];

[m14\$1* -0.524];

[m15\$1* -1.815];

[m16\$1* -1.545];

%C#2%

[m1\$1* 3.386];

[m2\$1* 2.674];

[m3\$1* 0.990];

[m4\$1* -2.473];

[m5\$1* 0.490];

[m6\$1* 0.404];

[m7\$1* -0.740];

[m8\$1* 2.497];

[m9\$1* 1.261];

[m10\$1* 1.058];

[m11\$1* 0.194];

[m12\$1* 2.084];

[m13\$1* 2.346];

[m14\$1* 1.864];

[m15\$1* -1.777];

[m16\$1* 1.240];

%C#3%

[m1\$1* -1.230];

[m2\$1* -1.240];

[m3\$1* 0.344];

[m4\$1* -2.971];

[m5\$1* 0.469];

[m6\$1* -0.079];

```
[ m7$1* -1.043];  
[ m8$1* 1.456];  
[ m9$1* -0.730];  
[ m10$1* 1.008];  
[ m11$1* 0.004];  
[ m12$1* 2.004];  
[ m13$1* 2.112];  
[ m14$1* 1.885];  
[ m15$1* -1.763];  
[ m16$1* 1.186];
```

A1.1

TITLE: Stepwise MIMIC Model DIF Detection

DATA: file is b16-gender.dat;

VARIABLE:

NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

MISSING ARE ALL (99);

USEVARIABLES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14
m15 m16;

CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

CLASSES = c(3);

ANALYSIS:

type=mixture;

MODEL:

%OVERALL%

[c#1*-1.291];

[c#2* 0.489];

c on gender;

m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 on gender;

%C#1%

[m1\$1* -0.612];

[m2\$1* -0.444];

[m3\$1* -1.149];

[m4\$1* -3.141];

[m5\$1* -1.383];

[m6\$1* -1.561];

[m7\$1* -2.091];

[m8\$1* -0.524];

[m9\$1* -0.668];

[m10\$1* -1.982];

[m11\$1* -0.891];

[m12\$1* -0.631];

[m13\$1* -0.345];

[m14\$1* -0.524];

[m15\$1* -1.815];

[m16\$1* -1.545];

m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 on gender;

%C#2%

[m1\$1* 3.386];
 [m2\$1* 2.674];
 [m3\$1* 0.990];
 [m4\$1* -2.473];
 [m5\$1* 0.490];
 [m6\$1* 0.404];
 [m7\$1* -0.740];
 [m8\$1* 2.497];
 [m9\$1* 1.261];
 [m10\$1* 1.058];
 [m11\$1* 0.194];
 [m12\$1* 2.084];
 [m13\$1* 2.346];
 [m14\$1* 1.864];
 [m15\$1* -1.777];
 [m16\$1* 1.240];

m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 on gender;

%C#3%

[m1\$1* -1.230];
 [m2\$1* -1.240];
 [m3\$1* 0.344];
 [m4\$1* -2.971];
 [m5\$1* 0.469];
 [m6\$1* -0.079];
 [m7\$1* -1.043];
 [m8\$1* 1.456];
 [m9\$1* -0.730];
 [m10\$1* 1.008];
 [m11\$1* 0.004];
 [m12\$1* 2.004];
 [m13\$1* 2.112];
 [m14\$1* 1.885];
 [m15\$1* -1.763];
 [m16\$1* 1.186];

m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16 on gender;

A2.0.1

TITLE: Stepwise MIMIC Model DIF Detection

DATA: file is data_savedata.txt;

VARIABLE:

NAMES ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16
 gender cprob1 cprob2 cprob3 cmod ;

MISSING ARE ALL (9999);

USEVARIABLES ARE m1 cmod gender;

CATEGORICAL ARE m1;

NOMINAL are cmod;

CLASSES = c(3);

ANALYSIS:

```
type=mixture;
STARTS=0;
processors = 7;
MODEL:
%OVERALL%
[ c#1*-1.291 ];
[ c#2* 0.489 ];
c on gender;

%C#1%
[cmo#1@2.610 cmo#2@-4.036];
%C#2%
[cmo#1@-3.434 cmo#2@1.739];
%C#3%
[cmo#1@-1.477 cmo#2@-2.631];
```

A2.1.1

TITLE: Stepwise MIMIC Model DIF Detection

DATA: file is data_savedata.txt;

VARIABLE:

NAMES ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16

gender cprob1 cprob2 cprob3 cmod ;

MISSING ARE ALL (9999);

USEVARIABLES ARE m3 cmod gender;

CATEGORICAL ARE m3;

NOMINAL are cmod;

CLASSES = c(3);

ANALYSIS:

```
type=mixture;
```

```
STARTS=0;
```

```
processors = 7;
```

MODEL:

```
%OVERALL%
```

```
[ c#1*-1.291 ];
```

```
[ c#2* 0.489 ];
```

```
c on gender;
```

```
m1 on gender;
```

```
%C#1%
```

```
[cmo#1@2.610 cmo#2@-4.036];
```

```
m1 on gender;
```

```
%C#2%
```

```
[cmo#1@-3.434 cmo#2@1.739];
```

```
m1 on gender;
```

```
%C#3%
```

[cmod#1@-1.477 cmod#2@-2.631];
m1 on gender;

OUTPUT: CINTERVAL;

A3.0.

TITLE: Stepwise MIMIC Model DIF Detection

DATA: file is b16-gender.dat;

VARIABLE:

NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

MISSING ARE ALL (99);

USEVARIABLES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14
m15 m16;

CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

CLASSES = c(3);

ANALYSIS:

type=mixture;

MODEL:

%OVERALL%

[c#1*-1.291];

[c#2* 0.489];

c on gender;

m9 m12 m13 m15 m16 on gender;

%C#1%

[m1\$1* -0.612];

[m2\$1* -0.444];

[m3\$1* -1.149];

[m4\$1* -3.141];

[m5\$1* -1.383];

[m6\$1* -1.561];

[m7\$1* -2.091];

[m8\$1* -0.524];

[m9\$1* -0.668];

[m10\$1* -1.982];

[m11\$1* -0.891];

[m12\$1* -0.631];

[m13\$1* -0.345];

[m14\$1* -0.524];

[m15\$1* -1.815];

[m16\$1* -1.545];

m9 m12 m13 m15 m16 on gender;

%C#2%

[m1\$1* 3.386];

[m2\$1* 2.674];

[m3\$1* 0.990];

[m4\$1* -2.473];

```
[ m5$1* 0.490 ];  
[ m6$1* 0.404 ];  
[ m7$1* -0.740 ];  
[ m8$1* 2.497 ];  
[ m9$1* 1.261];  
[ m10$1* 1.058];  
[ m11$1* 0.194 ];  
[ m12$1* 2.084 ];  
[ m13$1* 2.346 ];  
[ m14$1* 1.864 ];  
[ m15$1* -1.777 ];  
[ m16$1* 1.240 ];  
m9 m12 m13 m15 m16 on gender;
```

```
%C#3%  
[ m1$1* -1.230 ];  
[ m2$1* -1.240 ];  
[ m3$1* 0.344 ];  
[ m4$1* -2.971 ];  
[ m5$1* 0.469 ];  
[ m6$1* -0.079 ];  
[ m7$1* -1.043 ];  
[ m8$1* 1.456 ];  
[ m9$1* -0.730 ];  
[ m10$1* 1.008 ];  
[ m11$1* 0.004 ];  
[ m12$1* 2.004 ];  
[ m13$1* 2.112 ];  
[ m14$1* 1.885 ];  
[ m15$1* -1.763 ];  
[ m16$1* 1.186 ];  
m9 m12 m13 m15 m16 on gender;
```

A.4.1.

TITLE: Stepwise MIMIC Model DIF Detection

DATA: file is b16-gender.dat;

VARIABLE:

NAMES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

MISSING ARE ALL (99);

USEVARIABLES ARE gender m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14
m15 m16;

CATEGORICAL ARE m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12 m13 m14 m15 m16;

CLASSES = c(3);

ANALYSIS:

type=mixture;

MODEL:

```
%OVERALL%
```

```
[ c#1*-1.291 ];
```

```
[ c#2* 0.489 ];
```

c on gender;
m12 m13 m15 m16 on gender;

%C#1%

[m1\$1* -0.612];
[m2\$1* -0.444];
[m3\$1* -1.149];
[m4\$1* -3.141];
[m5\$1* -1.383];
[m6\$1* -1.561];
[m7\$1* -2.091];
[m8\$1* -0.524];
[m9\$1* -0.668];
[m10\$1* -1.982];
[m11\$1* -0.891];
[m12\$1* -0.631];
[m13\$1* -0.345];
[m14\$1* -0.524];
[m15\$1* -1.815];
[m16\$1* -1.545];

m12 m13 m15 m16 on gender;

%C#2%

[m1\$1* 3.386];
[m2\$1* 2.674];
[m3\$1* 0.990];
[m4\$1* -2.473];
[m5\$1* 0.490];
[m6\$1* 0.404];
[m7\$1* -0.740];
[m8\$1* 2.497];
[m9\$1* 1.261];
[m10\$1* 1.058];
[m11\$1* 0.194];
[m12\$1* 2.084];
[m13\$1* 2.346];
[m14\$1* 1.864];
[m15\$1* -1.777];
[m16\$1* 1.240];

m12 m13 m15 m16 on gender;

%C#3%

[m1\$1* -1.230];
[m2\$1* -1.240];
[m3\$1* 0.344];
[m4\$1* -2.971];
[m5\$1* 0.469];
[m6\$1* -0.079];
[m7\$1* -1.043];

[m8\$1* 1.456];
[m9\$1* -0.730];
[m10\$1* 1.008];
[m11\$1* 0.004];
[m12\$1* 2.004];
[m13\$1* 2.112];
[m14\$1* 1.885];
[m15\$1* -1.763];
[m16\$1* 1.186];
m12 m13 m15 m16 on gender;

The study of the effect of item parameter drift on ability estimation obtained from adaptive testing under different conditions

Merve Sahin Kursad^{1,*}, Omay Cokluk Bokeoglu², Rahime Nukhet Cikrikci²

¹National Defense University, Department of Measurement and Evaluation, Ankara, Türkiye

²Ankara University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

³Istanbul Aydın University, Faculty of Science and Literature, İstanbul, Türkiye

ARTICLE HISTORY

Received: Feb. 09, 2022

Revised: July 25, 2022

Accepted: Aug. 08, 2022

Keywords:

Item parameter drift,
Computer adaptive test,
Measurement precision,
Test information function.

Abstract: Item parameter drift (IPD) is the systematic differentiation of parameter values of items over time due to various reasons. If it occurs in computer adaptive tests (CAT), it causes errors in the estimation of item and ability parameters. Identification of the underlying conditions of this situation in CAT is important for estimating item and ability parameters with minimum error. This study examines the measurement precision of IPD and its impacts on the test information function (TIF) in CAT administrations. This simulation study compares sample size (1000, 5000), IPD size (0.00 logit, 0.50 logit, 0.75 logit, 1.00 logit), percentage of items containing IPD (0%, 5%, 10%, 20%), three time points and item bank size (200, 500, 1000) conditions. To examine the impacts of the conditions on ability estimations; measurement precision, and TIF values were calculated, and factorial analysis of variance (ANOVA) for independent samples was carried out to examine whether there were any differences between estimations in terms of these factors. The study found that an increase in the number of measurements using item bank with IPD items results in a decrease in measurement precision and the amount of information the test provides. Factorial ANOVA for independent samples revealed that measurements precision and TIF differences are mostly statistically significant. Although all IPD conditions negatively affect measurement precision and TIF, it has been shown that sample size and item bank size generally do not have an increasing or decreasing effect on these factors.

1. INTRODUCTION

Computer adaptive tests (CAT) produce more reliable results in ability estimations of individuals compared to paper-and-pencil tests and have many advantages. CAT administrations based on Item Response Theory (IRT) place each individual's ability on the same scale with item difficulty values by employing a variety of computer algorithms and measuring the probability that 50% of individuals will provide a correct response to the relevant item (Lord, 1980; Reckase, 2011). This way, tests can be conducted that are more efficient than paper-and-pencil tests in terms of cost and time, but are just as valid and reliable as paper-and-pencil tests by providing individuals with suitable items in line with their ability levels (Çikrikçi-Demirtaşlı, 1999; Kaptan, 1993; Wainer, 1993; Weiss & Kingsbury, 1984).

*CONTACT: Merve ŞAHİN KÜRŞAD ✉ sahinmerv@gmail.com 📍 Devlet Mahallesi, Kara Harp Okulu Caddesi, National Defense University, Department of Measurement and Evaluation, Ankara, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

Creating a large item bank consisting of high-quality items in CAT administrations is the primary step and an important factor for obtaining valid and reliable results. During the administration of tests, it is important for these items to be of high quality and to maintain this characteristic in successive administrations to obtain accurate results (Bock et al, 1988). Maintaining the item bank's continuity is important for test reliability and observing the changes in item parameters (Risk, 2015). The long-term use of items in the item bank may negatively affect the quality of items, because the repeated use of certain items in administration results in individuals becoming familiar to with these items. Even if the reliability of the item bank is ensured, the frequent encounter of individuals with the same items becomes a factor that compromises reliability and causes item parameters to change or deviate from their original values over time. This change is called item parameter drift (IPD) (Bock et al., 1988). First introduced to the literature in the 1980s, IPD is defined as the differentiation of item parameters over time in successive administrations of tests (Hatfield & Nhoyvanisvong, 2005; McCoy, 2009). This differentiation may occur in one or more parameters of an item (Goldstein, 1983).

Item parameter drift may even occur in situations where the security of the item bank is ensured, and high-quality items are prepared. There are several reasons for the occurrence of IPD in items. Some of these reasons may be listed as: historical and cultural changes, incorrect item calibration, miscalculation of item location on the scale, changes in knowledge, skills, and educational activities, overuse of items, changes in policy or curricula, cheating or security (Li, 2008; Stahl & Muckle, 2007). IPD arising from these reasons may increase or decrease item difficulty or simultaneously increase and decrease item difficulty or other item parameters. Certain negative results thus may arise. The most significant of these negative results is the violation of the invariance assumption, one of the basic assumptions of IRT. If the invariance assumption is ensured, the differences between scores accurately reflect ability differences between individuals or individuals' development over time. However, the occurrence of IPD leads to errors in measurement results, and the test may measure something outside of the construct it intends to measure. Validity also decreases when variables that are irrelevant to the measured construct get mixed into measurement results (McCoy, 2009). This leads to certain problems in the administrations of tests that require the invariance property in item parameters, including test equating, test developing/parallel test developing and CAT (Li, 2008). For instance, in the event of IPD in pre-test items of CAT administrations, errors may occur in item calibration (Meng et al., 2010).

When the scores of two individuals are close to each other, or an individual's score is close to the cut-off score, IPD may lead to incorrect pass-fail decisions and deviations in ability estimations (Rupp & Zumbo, 2006). Apart from that, IPD can also occur when there is not enough time to answer the questions in a test. Not providing sufficient time results in individuals being unable to reach certain items at the end of the test and these items appear more difficult than they actually are. If this problem persists in successive tests administrations, errors pile up and the measurement using previous test items is negatively affected. This leads to the measurement scale to drift (Wise & Kingsbury, 2000).

Another impact of IPD can be observed in CAT administrations. Similar to paper-and-pencil tests, both item and ability parameters are negatively affected in CAT administrations. In terms of item parameters, using previous item parameter estimations to scale new test items leads to errors in item parameter calibration. This results in the deviation of item parameters. The deviation of items from their original parameter values results in the incorrect calibration of pre-test items, leading to errors in individuals' ability estimations (Deng & Melican, 2010). As CAT administrations become more frequently used, the occurrence of IPD in these administrations negatively affects the accuracy of ability estimations and the validity of inferences from test scores.

IPD is a condition that affects the accuracy of individuals' ability estimations and pass/fail decisions. Examining the effect of IPD on measurement precision and TIF is crucial for the safe combination of exams as a whole and the validity of inferences to be made from test scores. Although the use of CAT is widespread, the presence of IPD in these applications negatively affects the accuracy of ability estimations and the validity of inferences made from test scores. Therefore, an examination of the impact of IPD on ability estimations for CAT administrations is significant for the validity of inferences from test scores. Additionally, items containing IPD may have different effects in groups participating in different test administrations. This is a significant issue for CAT administrations since it violates the invariance assumption, one of the basic assumptions of IRT (Babcock & Albano, 2012). The presence of IPD in test applications, where large item banks are used, and especially important decisions are made about the test takers, causes variables unrelated to the structure to interfere with the measurement results, thus reducing the validity. This issue negatively affects measurement precision of scores and validity when interpreting scores in particular (Risk, 2015). For this reason, carrying out IPD studies of item banks in CAT administrations serves to counter this issue, posing threats to construct validity (Wainer et al., 2010). The results of this study are also important to see how the direction, amount and size of the deviations in the item difficulty parameter affect measurement precision and TIF for future CAT applications. In this direction, it is expected that the research findings will provide psychometric information about the organization of the CAT, the sustainability, and updating of the item bank to the institutions and organizations serving in the field of measurement and evaluation.

The overuse of items in successive CAT administrations is a significant cause of IPD occurrence (Bock et al., 1988). For this reason, the item bank should regularly be inspected and updated. IPD studies should therefore be conducted for CAT administrations. However, few studies in the literature examine the impacts of IPD on estimation of ability and item parameters in CAT administrations (Aksu Dünya, 2017; Deng & Melican, 2010; Guo & Wang, 2003; Han & Guo, 2011; Risk, 2015). When some of these studies were closely examined, Guo and Wang (2003) examined the effect of scale drift on the CAT application. The study was conducted with real and simulative data, and the bias in ability estimations and the change in test scores were calculated. Bias, test characteristic curves, and item characteristic curves were compared. As a result of the research, it was stated that a low amount of bias was observed, and this was not important in practical terms. In addition, it was determined that scale drift affects test scores, but this change between two time points is very low. Deng and Melican (2010) studied IPD at multiple time points in CAT applications. The adaptive ACCUPLACER® test was evaluated as part of the scope of the study. Four time points were analyzed using a 3-parameter logistic model (3PLM), and the IPD at parameters a, b, and c was examined. In the evaluation, the item and test characteristic curves were compared. As a result of study, very few items were found to have IPD, but none of the items showed IPD due to its frequent occurrence.

Han and Guo (2011) studied IPD in the context of CAT, resulting from practice and curriculum change. In the study, the effect of IPD on item calibration and ability estimation was examined, using both real and simulative data. Items were calibrated according to 3PLM. According to the results of the study, it was determined that the effect of IPD on item calibration and ability estimations was high, but this effect was not statistically significant. A similar result was obtained by Risk (2015) who examined the effect of IPD on ability estimations in CAT application under various simulative conditions. The Rasch model was used in the study, and the effect of IPD on measurement precision and test effectiveness was examined. When the findings obtained from all conditions are evaluated in general, it is concluded that there are negligible differences between the baseline data set and the conditions that create IPD. However, the most important finding that emerged as a result of the study was; that IPD size has a greater effect on measurement precision than the number of items showing IPD.

Aksu Dünya (2017) investigated the effect of IPD on ability estimations and classification accuracy in the CAT under the condition that IPD affects subgroups with Rasch dichotomous model. According to the study's findings, classification accuracy was significantly affected when a certain group of individuals were exposed to items with IPD. At the same time, average ability estimates were less affected by IPD. In summary, these studies generally focus on the impacts of IPD on item and ability parameters. While some studies find that IPD has a significant effect on CAT-obtained ability estimates (Abad et al., 2010; Hagge et al., 2011; Risk, 2015), others argue that its effect on CAT-obtained ability estimates is small and insignificant (Aksu Dünya, 2017; Deng & Melican, 2010; Guo & Wang, 2003; Han & Guo, 2011; Jiang et al., 2009; McCoy, 2009). These studies mostly examine the impacts of IPD for two time points. However, to be able to observe the impacts of IPD, measurements should be taken for more than two time points. Because it is stated in the literature that if there is an IPD, its effect can be observed clearly after two time points, the IPD's effect can be observed after two time points. Therefore, more than two time points are needed (Babcock & Albano, 2012; Chan et al, 1999; Deng & Melican, 2010; Kim & Cohen, 1992). During the literature review, we could not find any study examining the impact of this issue on ability estimations and test information function while accounting for sample size, item bank size, and various conditions of IPD. Therefore, the impacts of IPD on factors as mentioned above in CAT administrations are not fully known. The aim of this study is to investigate the impact of IPD on measurement precision and TIF in CAT administrations. To this end, answers to the following research questions are sought:

1. When the sample size is 1000, IPD size is 0.00, 0.50, 0.75, 1.00 logit, percentage of items containing IPD is 0%, 5%, 10%, 20%, item bank size is 200, 500, 1000, and measurements are taken for three time points, how do the values of measurement precision and TIF vary in CAT administrations?
2. When the sample size is 5000, IPD size is 0.00, 0.50, 0.75, 1.00 logit, percentage of items containing IPD is 0%, 5%, 10%, 20%, item bank size is 200, 500, 1000, and measurements are taken for three time points, how do the values of measurement precision and TIF vary in CAT administrations?

2. METHOD

2.1. Research Model

This is a simulation-based study that utilized simulated data. Simulation studies are frequently favored in real-world situations involving relatively complex processes, implementation issues, or when real data suited to the type of problem are unavailable. Simulation studies consist of data generating and analysis processes appropriate to situations encountered in real life (Burton et al., 2006; Ranganathan & Foster, 2003). Simulated data are frequently preferred, given the fact that most CAT administrations have implementation problems and require a large sample size and a large item bank (Barrada et al., 2010; Kalender, 2011; McDonald, 2002; Patton et al., 2013; Scullard, 2007; Wang et al., 2012). In this study, because small, medium, and especially large item pools and small and especially large sample sizes are used and drawing on IRT, examines certain IPD situations under controlled conditions in CAT administrations, it is a simulative research.

2.2. Data Generation and Analysis

This study used the R programming language and carried out analyses by generating data using the R Studio 3.3.2 CRAN package (Nydick, 2015). The characteristics of CAT administrations and large-scale assessments were considered when generating data. IPD size and the percentage of items containing IPD were considered when creating conditions for IPD. Also, data was created for taking measurements at three time points. The initial data set that does not contain

IPD was used as the baseline data set during data generation, and data sets containing IPD were compared to this baseline data set. Table 1 displays the controlled and manipulated conditions used in data generation.

Table 1. *Controlled and manipulated conditions in simulated data generation.*

Controlled Conditions	Manipulated Conditions
1. Distribution of ability parameters	1. Sample size (1000, 5000)
2. IRT model and distribution of item parameters	2. IPD size (0.00, 0.50, 0.75, 1.00)
3. Direction and type of IPD	3. Percentage of items containing IPD (0%, 5%, 10%, 20%)
4. CAT Conditions	4. Three time points
• Method of ability estimation	5. Item bank size (200, 500, 1000)
• Starting Rule	
• Method of item selection	
• Termination Rule	

2.3. Controlled Conditions in Simulated Data Generation

Since the CAT administration in this study used the Bayesian Expected A Posteriori (EAP) estimate for ability estimation, the distribution of ability parameters was generated with normal distribution with a mean of zero and standard deviation of one. Rasch was chosen as the IRT model because it is favored in large-scale assessments such as Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA) (Schulz & Frallion, 2009) and IPD studies for large-scale assessments (Babcock & Albano, 2012; Bergstrom et al., 2001; Hagge et al., 2011; Jones & Smith, 2006; Kingsbury & Wise, 2011; McCoy, 2009; Meyers et al., 2009; Witt et al., 2003). Taking into account the characteristics of CAT administrations and studies in the relevant literature (Filho et al., 2014; Svetina et al., 2013), the distribution of item difficulty parameters was generated with normal distribution with a mean of zero and a standard deviation of one.

The study examines the impact of the item difficulty parameter drift towards the easier condition on ability estimations. There are several reasons for examining this condition. These reasons may be listed as: this situation being encountered more frequently (Babcock & Albano, 2012; Hagge et al., 2011; Risk, 2015; Stahl & Muckle, 2007) and situations with drift in item difficulty parameter being more significant than other parameters (Bock et al., 1988; Donoghue & Isham, 1998; Song & Arce-Ferrer, 2009). Another reason is that although frequent exposure to items or factors such as cheating are observed more frequently, these situations negatively affect ability estimations by causing deviation towards the easier (Risk, 2015; Wells et al., 2012).

After IPD conditions were prepared, conditions for CAT administration were formed. Expected A Posteriori (EAP) was used as the ability estimation method. The ability estimation methods frequently used in CAT applications are the Maximum Likelihood Estimation (MLE), EAP and Maximum A Posteriori (MAP) methods. In most of the studies, the EAP ability estimation method yielded better results than the other two methods (Eroğlu, 2013; Kezer, 2013; Keller, 2000; Kingsbury & Zara, 2009; Wang et al., 2012), with a lower standard error (Wang, 1997) and lower bias value than the MLE method (Eroğlu, 2013). The MLE method was not specified as effective because it estimates ability with more items than EAP and MAP methods (Kezer, 2013). For these reasons, the EAP method was used as an ability estimation method in the CAT application.

Prior θ distributions according to scores individuals acquired in pre-tests were used as starting rule. When the Bayesian approach is used as an ability estimation method, the initial θ level is

estimated from the pre-test before estimating the individuals' real abilities. Thus, the first item to be applied will be the item that gives the most information at the initial θ level (Eroğlu, 2013; Kezer, 2013; Segall, 2004). Accordingly, in this study, as the ability estimation method, one of the Bayes methods, EAP, was used, and the prior θ distributions were used as the starting rule according to the scores of the individuals from the pre-test. The Kullback–Leibler divergence was used for the item selection method. Basic item selection methods used in CAT applications; Maximum Fisher Information, Kullbak-Leibler Information, Interval Information Criterion, Likelihood Weighted Information Criterion, a-stratification, Gradual Maximum Information Ratio, Optimal -b Value (Sulak, 2013). In studies comparing the performance of these methods, -a stratification and Kullbak-Leibler item selection methods have better performances in ability estimations than other methods (Barrada et al., 2010; Chang & Ying; 1999; Chen et al., 2000; Deng et al., 2010; Eggen, 1999; Linda 1996; Sulak, 2013; Veldkamp & van der Linden, 2006; Yao, 2013). However, since the analyzes were made based on the Rasch model within the scope of this study, the -a stratification method is not suitable because the discrimination values of all items are constant. For this reason, the Kullbak-Leibler item selection method was preferred.

Lastly, the minimum number of items rule, one of the variable-length termination rules, and standard error were used as the termination rule. For the minimum number of items rule, the termination rule was set as minimum 10 items and standard error at less than 0.40. Higher error and bias values are obtained when the minimum number of items applied is less than 10 (Babcock & Weiss 2012; Erolu, 2013), and the normal distribution is compromised when the minimum number of items applied is low (Blais & Raiche, 2002). Therefore, in this study, a minimum of 10 items was preferred for the minimum number of items rule. In the standard error termination rule between the [-3.00; +3.00] ability interval, a standard error equal to or less than 0.40 is suitable for measurement precision (Babcock & Weiss, 2012; Blaise & Raiche, 2002). Therefore, these termination rules were preferred.

2.4. Manipulated Conditions in Simulated Data Generation

While a sample size of 1000-2000 is required to make accurate estimations of item parameters based on IRT (Rudner & Guo, 2011; Stahl & Muckle, 2007), lower standard error values are obtained when the sample size is 5000 (Şahin, 2012). The sample size of 1000 was thus treated as the small sample size and 5000 as the large sample size. One of the most important factors affecting the estimation of ability is the size of the IPD (Risk, 2015). IPD size of 0.50 logit or more significantly affects parameter estimations (Donoghue & Isham, 1998; Han & Wells, 2007; Wollack et al., 2005). Therefore 0.00, 0.50, 0.75, and 1.00 logit were generated as IPD magnitude to examine the impact of IPD magnitude.

As one of the factors negatively affecting ability estimations, the IPD percentage (Hagge et al., 2011; Huang & Shyu, 2003; Wells et al., 2002) was found to range between 5 and 20–25% in the relevant literature (Hagge et al., 2011; Stahl et al., 2002; Song & ArceFerrer, 2009; Wells et al., 2002). This study examines IPD-containing items with 0%, 5%, 10%, and 20%. To fully reveal the impact of IPD, more than two time points or measurements are needed (Babcock & Albano, 2012; Chan et al., 1999; Deng & Melican, 2010; Kim & Cohen, 1992). For this reason, this study uses parameter estimations at three time points. In line with some studies in the literature regarding item bank size in the CAT application (Han & Guo, 2011; Risk, 2015; Veldkamp & Linden, 2006; Wise & Kingsbury, 2000), this study set item bank sizes of 200, 500 and 1000 for small, medium and large item banks respectively.

Given the controlled and manipulated conditions, simulated data were generated for 288 situations, calculated as 2 (sample sizes) \times 3 (item bank sizes) \times 4 (IPD sizes) \times 4 (IPD percentages) \times 3 (time points). For every situation, a total of 100 replications were carried out and 28,800 analyses were performed. In simulation studies, replication numbers must be kept higher to see the effect of the variables on the situations to be observed more clearly (Köse &

Başaran, 2021). As Evans (2010) quoted, to eliminate bias caused by sample size, at least 25 replications were recommended (Harwell, 1996). Consequently, 100 replications were favored. To examine the effect of condition on estimations of ability, values for measurement precision (bias and root-mean-square error -RMSE-) and TIF were calculated. The calculation formulas are displayed in Table 2 below.

Table 2. Assessment criteria for item parameter drift.

Criteria		Description	Formula
Measurement Precision	Bias	Systematic deviation of real ability from estimated ability.	$\frac{\sum_{j=1}^n (\hat{\theta}_i - \theta_i)}{n}$
	RMSE	Root mean square error	$\sqrt{\frac{\sum_{j=1}^n (\hat{\theta}_i - \theta_i)^2}{n}}$
Test Information Function	TIF	The test information function is equal to the total information function of items individuals obtain from the relevant test. This value is calculated using standard error values.	$S_{em}(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$

θ_i : Ability of individuals, $\hat{\theta}_i$: Estimated ability of individuals, n : Total number of individuals, $I(\theta)$: Item information function. Also, measurement precision and TIF are correlated with each other by SEM with $RMSE^2 = BIAS^2 + S_{em}^2$ formula

After calculating values for measurement precision and TIF for 100 replications using the formula in Table 2, a three-factor analysis of variance (ANOVA) was performed for independent samples to examine whether the obtained values displayed statistically significant differences. In the analysis, the independent variables consisted of IPD size (0.00 logit, 0.50 logit, 0.75 logit, 1.00 logit), IPD percentage (0%, 5%, 10%, 20%), and measurements using item banks with IPD (3 measurements), while the dependent variables consisted of bias, RMSE, and TIF values. Along with ANOVA, the Eta squared (η^2) effect size was also reported. When interpreting the effect size, .01 was taken as small, .09 as a medium, and .25 as large effect sizes (Cohen, 1988). When calculating the impact of IPD for every condition, the initial data set that did not contain IPD was taken as the baseline data set. After forming IPD conditions using this data set, data sets containing IPD and the baseline data set were compared, and the results were interpreted.

3. FINDINGS

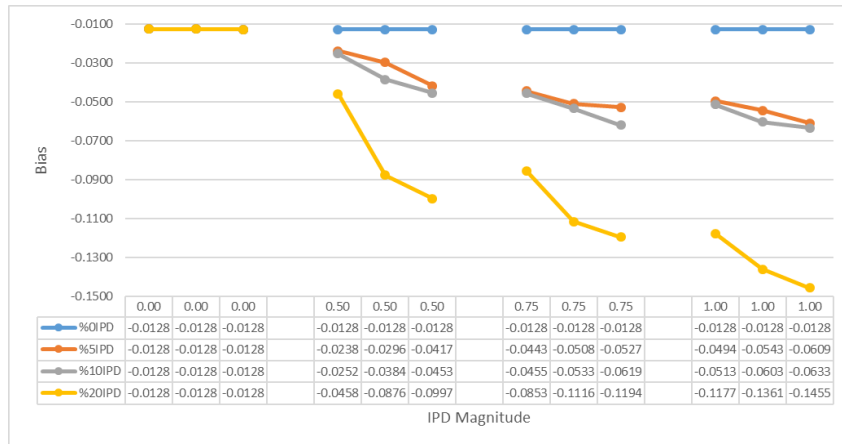
This section first discusses the findings and interpretations obtained from data for the sample size of 1000, then goes on to findings and interpretations of data with a sample size of 5000.

3.1. Findings on Comparison of Conditions with Sample Size 1000

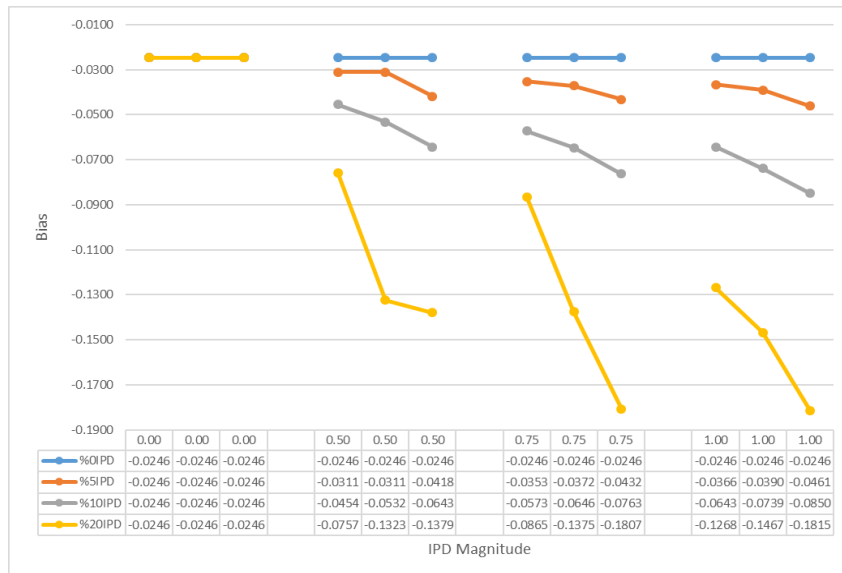
The first criterion for measurement precision, i.e., the dependent variable, is the bias values regarding ability estimations. Findings of bias values are shown in Figure 1. a, b and c. When bias values were examined for a sample size of 1000, the increase in IPD size (0.00, 0.50, 0.75, 1.00) and IPD percentage (0%, 5%, 10%, 20%) for item bank sizes of 200, 500 and 1000, resulted in a tendency of ability estimation *bias* values obtained at three time points to increase in the negative direction. Besides this, as item bank size increased, no increasing or decreasing bias tendency were observed. Negative bias values mean that individuals' estimated ability values are lower than their real ability values. Since certain items in the item bank displayed IPD in the easier direction, we would have expected individuals' estimated ability values to be higher than their real ability values; in other words, bias values should have increased in the positive direction. There may be two reasons for obtaining results in the opposite direction.

Figure 1. a, b and c. Figures denoting comparison of bias values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size $n=1000$.

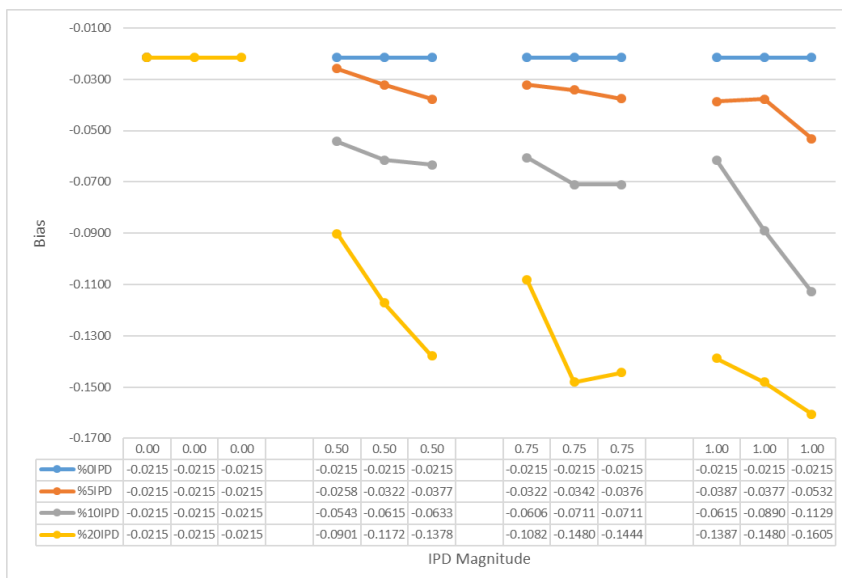
a. Bias values for item bank of 200 with sample size $n=1000$.



b. Bias values for item bank of 500 with sample size $n=1000$.



c. Bias values for item bank of 1000 with sample size $n=1000$.



Firstly, IPD in the easier direction occurred only for the item difficulty parameter. If IPD had occurred at both directions, bias estimations would have been calculated as near-zero (Aksu Dünya, 2017; Wei, 2013). Secondly, although individuals were provided with items according to their ability level, they may have answered incorrectly. Some studies in relevant literature have also come up with similar findings (Chen, 2013; Risk, 2015; Rupp & Zumbo, 2003).

On the other hand, a study by Guo and Wang (2003) that examined the impact of the parameter drift in CAT administrations on test scores showed that ability estimation bias values for item banks with IPD were not affected. This is because the study carried out measurements at two time points. Babcock and Albano (2012) also stated that taking ability measurements at two time points is insufficient to make clear inferences about how IPD affects ability estimations. In other words, in order to reveal the effects of IPD, it is necessary to take measurements at least three time points.

Three-factor ANOVA results for independent samples, as shown in Table 3, examine whether obtained differences were statistically significant according to above-mentioned bias values. In Table 3, IPD size represents the drift size of items containing IPD in the item bank (0.00 logit, 0.50 logit, 0.75 logit, 1.00 logit), IPD percentage represents the percentage of items containing IPD in the item bank (0%, 5%, 10%, 20%), and measurement factor represents the number of measurements performed with the item bank containing items with IPD (3 measurements).

Table 3. Comparison of bias values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size $n=1000$.

Item Bank Size	Source of Variation	Sum of Squares	df	Mean of Squares	F	Effect Size (η^2)
200	IPD Size	0.515	3	0.172	5366.84*	0.15
	IPD Percentage	2.031	3	0.677	21165.16*	0.58
	Measurement	0.235	2	0.118	2448.95*	0.06
	IPD Size*IPD Percentage	0.113	9	0.013	1177.58*	0.03
	IPD Size*Measurement	0.016	6	0.003	166.74*	0.01
	IPD Percentage*Measurement	0.071	6	0.012	739.89*	0.02
	IPD Size*IPD Percentage*Measurement	0.011	18	0.001	114.63*	0.01
	Error	0.456	4752	0.000		
	Total	3.448	4799			
500	IPD Size	0.196	3	0.065	2352.00*	0.03
	IPD Percentage	4.414	3	1.471	52968.00*	0.73
	Measurement	0.493	2	0.247	5916.00*	0.08
	IPD Size*IPD Percentage	0.070	9	0.008	840.00*	0.01
	IPD Size*Measurement	0.022	6	0.004	264.00*	0.01
	IPD Percentage*Measurement	0.333	6	0.056	3996.00*	0.05
	IPD Size*IPD Percentage*Measurement	0.060	18	0.003	720.00*	0.01
	Error	0.396	4752	0.000		
	Total	5.984	4799			
1000	IPD Size	0.271	3	0.090	2893.91*	0.05
	IPD Percentage	4.232	3	1.411	45192.05*	0.78
	Measurement	0.248	2	0.124	2648.31*	0.04
	IPD Size*IPD Percentage	0.047	9	0.005	501.90*	0.01
	IPD Size*Measurement	0.024	6	0.004	256.29*	0.01
	IPD Percentage*Measurement	0.058	6	0.010	619.36*	0.01
	IPD Size*IPD Percentage*Measurement	0.075	18	0.004	800.90*	0.01
	Error	0.445	4752	0.000		
	Total	5.400	4799			

* $p < .05$

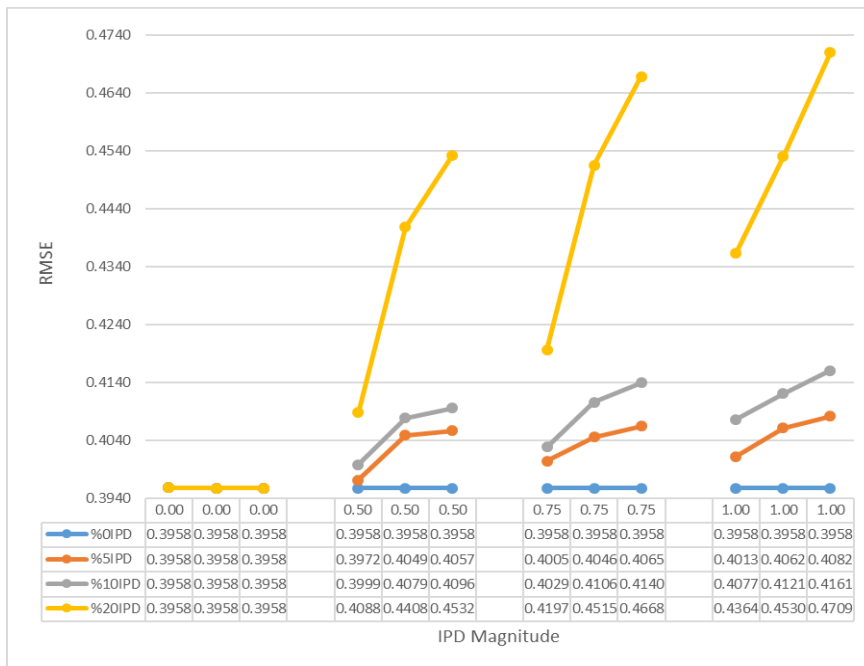
ANOVA results for independent samples regarding bias show that the main effect and effects of two-way and three-way interactions of the number of measurements, IPD size, and IPD percentage for item bank sizes of 200, 500 and 1000 items have statistically significant effects

on bias. These generally have low effect sizes (Cohen, 1988). The post-hoc analysis results also revealed differences for every level of every factor. IPD percentage is the factor with the most impact on ability estimation bias among the variables within the scope of this study. Aksu Dünya (2017) and Babcock and Albano (2012), who used the Rasch model and Abad et al. (2010), who used the 3PLM IRT model, obtained similar findings.

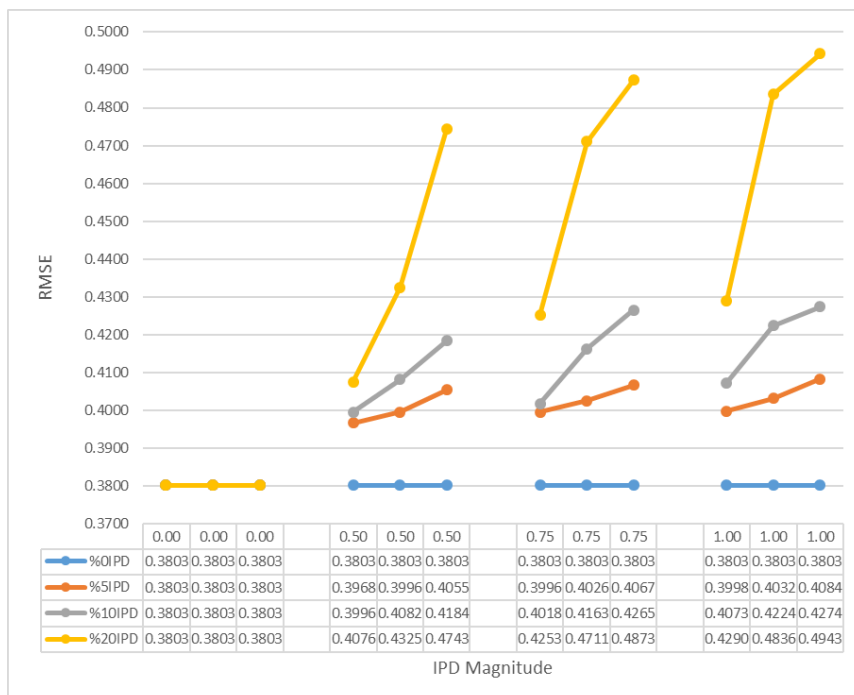
The second criterion for measurement precision, i.e., the dependent variable, is the RMSE values for ability estimations. Obtained RMSE values are shown in Figure 2. a, b and c.

Figure 2. a, b. and c. Figures denoting comparison of RMSE values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is $n=1000$.

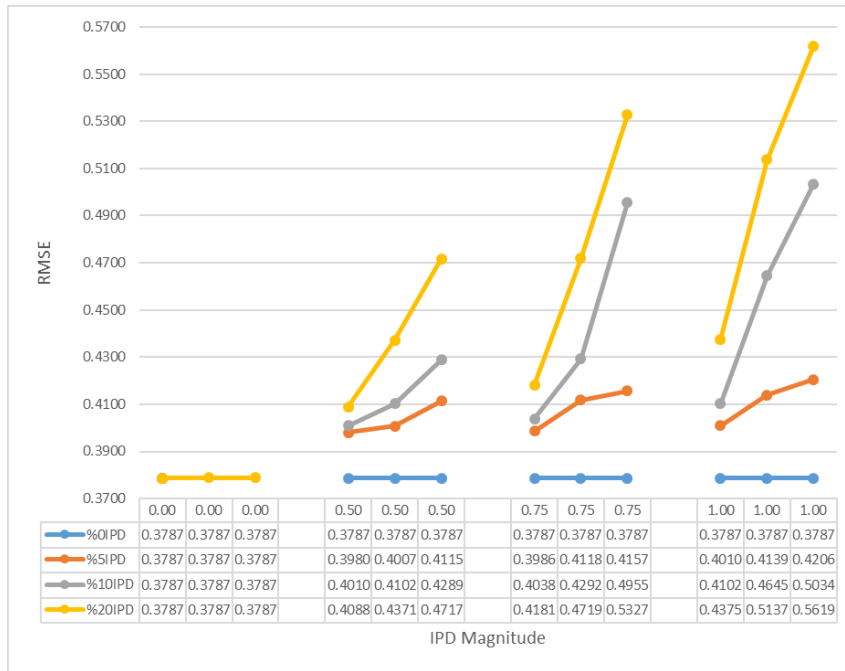
a. RMSE values for item bank of 200 with sample size $n=1000$.



b. RMSE values for Item Bank of 500 with Sample Size $n=1000$.



c. RMSE Values for Item Bank of 1000 with Sample Size n=1000.



When the RMSE values were examined for the sample size of 1000, the increase in IPD size and IPD percentage for item bank sizes of 200, 500, and 1000, resulted in a tendency of ability estimation RMSE values obtained at three time points to increase. The increase in the number of measurements in item banks containing IPD, IPD size, and IPD percentage results in more erroneous ability estimations leading to a decrease in measurement precision. The lowest values of RMSE were obtained in the baseline data set, since there was no IPD. However, RMSE values decreased as the item bank size increased for the baseline datasets. In other words, as the item bank size increases, less erroneous results regarding ability estimations were obtained in the baseline data set. Besides this, as item bank size increased for data sets with IPD, no increasing or decreasing RMSE tendency were observed. Some studies in relevant literature have also obtained similar findings (Aksu Dünya, 2017; Babcock & Albano, 2012; Chen, 2013; Risk, 2015; Wells et al., 2002). While Aksu Dünya (2017) argues that the lowest RMSE value was obtained for the baseline data set, it is stated that the increase in the percentage of items containing IPD resulted in more erroneous ability estimations. Wells et al. (2012) found that as sample size increased, RMSE values decreased, leading to more accurate estimates. However, as IPD size increased within the same sample size, RMSE values increased, leading to less precise measurements. Three-factor ANOVA results for independent samples are shown in Table 4 which examine whether obtained differences were statistically significant according to the RMSE values discussed above.

Results of a three-factor ANOVA on RMSE values for independent samples indicate that the main effect and effects of two-way and three-way interactions of the number of measurements, IPD size, and IPD percentage for item bank sizes of 200, 500 and 1000 items have statistically significant effects on RMSE. These generally possess low and high effect sizes (Cohen, 1988). The results of post-hoc analysis revealed differences for every level of every factor. IPD percentage is the factor with the most impact on ability estimation RMSE among the variables within the scope of this study. Risk (2015) also reached similar findings. A study by Babcock and Albano (2012) obtained similar findings, but argued that the factor with the most impact on RMSE values was IPD size.

Table 4. Comparison of RMSE values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size is $n=1000$.

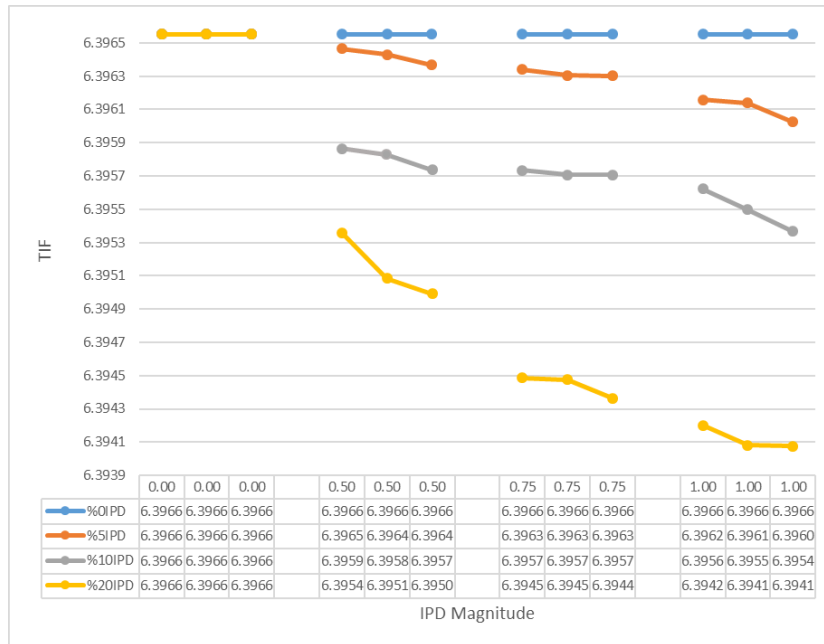
Item Bank Size	Source of Variation	Sum of Squares	df	Mean of Squares	F	Effect Size (η^2)
200	IPD Size	0.039	3	0.013	652.56*	0.02
	IPD Percentage	0.884	3	0.295	14791.44*	0.57
	Measurement	0.180	2	0.090	3011.83*	0.11
	IPD Size*IPD Percentage	0.023	9	0.003	384.85*	0.01
	IPD Size*Measurement	0.005	6	0.001	83.66*	0.01
	IPD Percentage*Measurement	0.115	6	0.019	1924.23*	0.07
	IPD Size*IPD Percentage*Measurement	0.005	18	0.000	83.66*	0.00
	Error	0.284	4752	0.000		
Total	1.535	4799				
500	IPD Size	0.104	3	0.035	1752.51*	0.03
	IPD Percentage	1.431	3	0.477	24113.87*	0.54
	Measurement	0.449	2	0.225	7566.13*	0.16
	IPD Size*IPD Percentage	0.068	9	0.008	1145.87*	0.02
	IPD Size*Measurement	0.016	6	0.003	269.62*	0.00
	IPD Percentage*Measurement	0.274	6	0.046	4617.19*	0.10
	IPD Size*IPD Percentage*Measurement	0.020	18	0.001	337.02*	0.00
	Error	0.282	4752	0.000		
Total	2.644	4799				
1000	IPD Size	0.718	3	0.239	9425.24*	0.12
	IPD Percentage	1.880	3	0.627	24678.90*	0.32
	Measurement	1.777	2	0.889	23326.81*	0.30
	IPD Size*IPD Percentage	0.253	9	0.028	3321.15*	0.04
	IPD Size*Measurement	0.205	6	0.034	2691.05*	0.03
	IPD Percentage*Measurement	0.552	6	0.092	7246.14*	0.09
	IPD Size*IPD Percentage*Measurement	0.089	18	0.005	1168.31*	0.01
	Error	0.362	4752	0.000		
Total	5.836	4799				

* $p < .05$

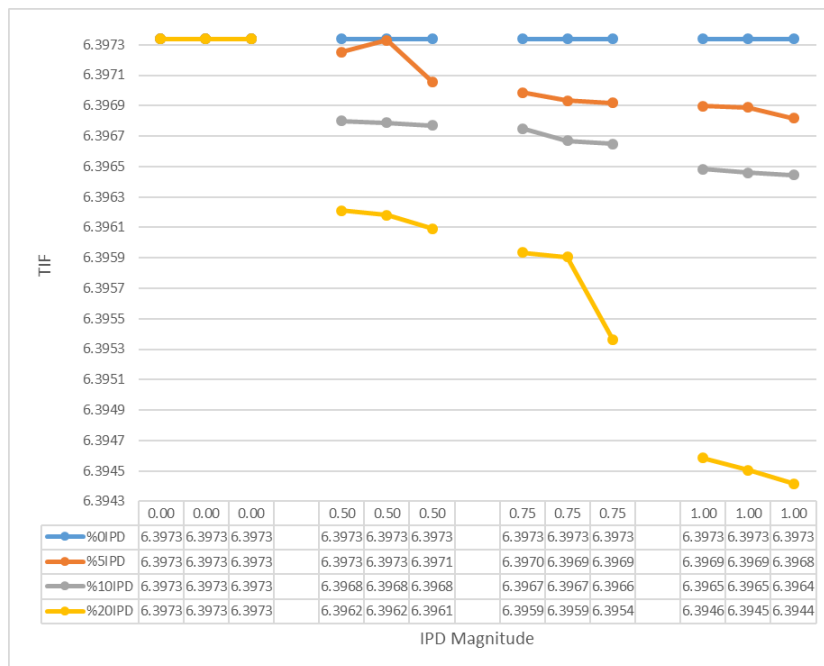
The third criterion for comparing independent variables discussed in the study is the TIF values. Findings for TIF values are shown in Figure 3. a, b and c. When TIF values were examined for a sample size of 1000, the increase in IPD size and IPD percentage for item bank sizes of 200, 500, and 1000 resulted in a tendency of ability estimation *TIF* values obtained at three time points to decrease. Therefore, the increase in the number of measurements, IPD size and IPD percentage result in a decrease in the amount of information the test provides. This tendency does not change with an increase in item bank size. Studies in the literature indicate that TIF tend to change even at low levels when IPD is present (Chan et al., 1999; Deng & Melican, 2010; Guo & Wang, 2003).

Figure 3. a, b and c. Figures denoting comparison of TIF values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is $n=1000$.

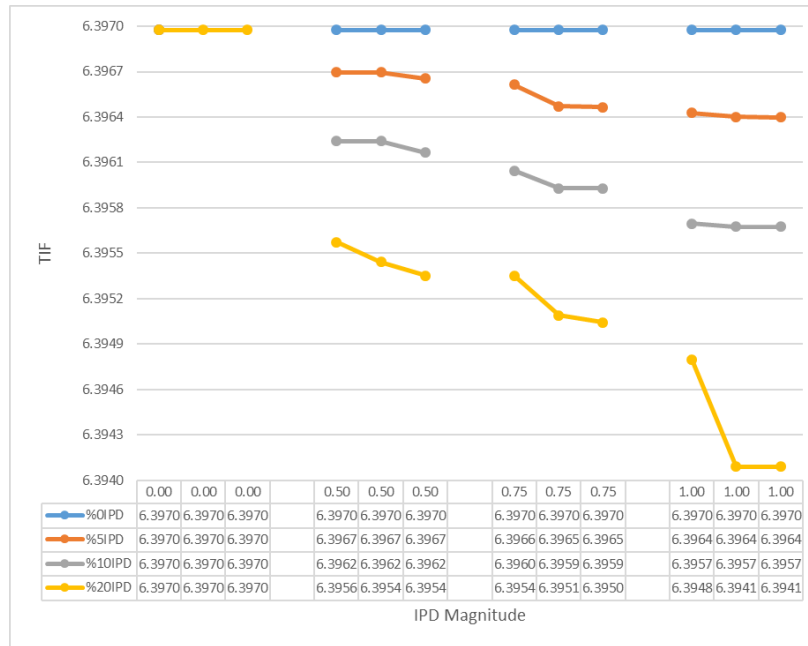
a. TIF values for item bank of 200 with sample size $n=1000$.



b. TIF values for item bank of 500 with sample size $n=1000$.



c. TIF values for item bank of 1000 with sample size n=1000.



The three-factor ANOVA results for independent samples, shown in Table 5, examine whether the differences obtained were statistically significant according to the TIF values discussed above.

Table 5. Comparison of TIF values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size n=1000.

Item Bank Size	Source of Variation	Sum of Squares	df	Mean of Squares	F	Effect Size (η^2)
200	IPD Size	19.342	3	6.447	79578.51*	0.35
	IPD Percentage	8.060	3	2.687	33161.14*	0.14
	Measurement	1.075	2	0.538	4422.86*	0.02
	IPD Size*IPD Percentage	16.112	9	1.790	66289.37*	0.30
	IPD Size*Measurement	2.148	6	0.358	8837.49*	0.03
	IPD Percentage*Measurement	2.148	6	0.358	8837.49*	0.04
	IPD Size*IPD Percentage*Measurement	4.294	18	0.239	17666.74*	0.08
	Error	1.155	4752	0.000		
Total	54.334	4799				
500	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.001	3	0.000	316.80*	0.02
	Measurement	0.012	2	0.006	3801.60	-
	IPD Size*IPD Percentage	0.000	9	0.000	0.00*	0.00
	IPD Size*Measurement	0.003	6	0.000	950.40	-
	IPD Percentage*Measurement	0.005	6	0.000	1584.00	-
	IPD Size*IPD Percentage*Measurement	0.008	18	0.000	2534.40	-
	Error	0.015	4752	0.000		
Total	0.045	4799				
1000	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.001	3	0.001	279.53*	0.00
	Measurement	0.018	2	0.009	5031.53	-
	IPD Size*IPD Percentage	0.068	9	0.007	19008.00*	0.00
	IPD Size*Measurement	0.003	6	0.000	838.59	-
	IPD Percentage*Measurement	0.014	6	0.002	3913.41	-
	IPD Size*IPD Percentage*Measurement	0.008	18	0.000	2236.24	-
	Error	0.017	4752	0.000		
Total	0.130	4799				

*p<.05

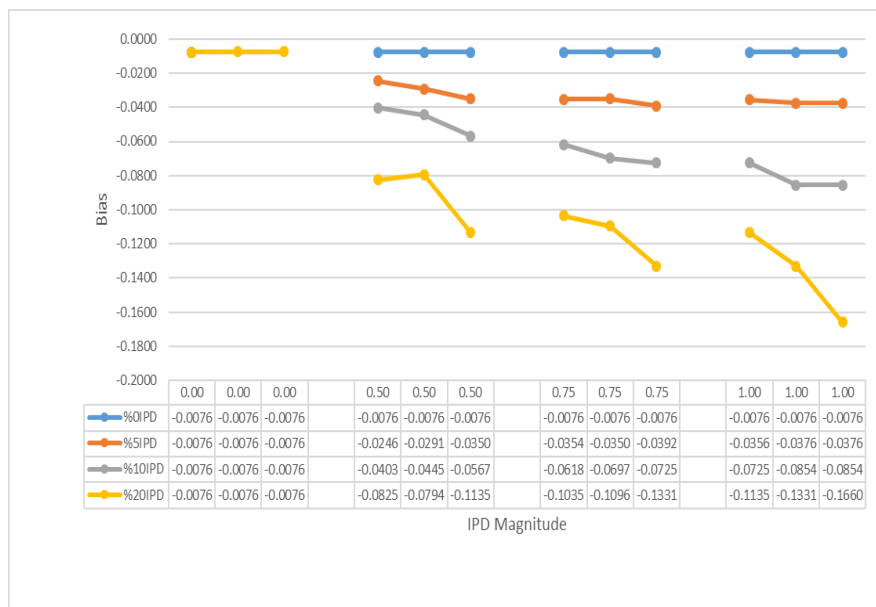
The three-factor ANOVA results for independent samples in terms of TIF values show that the main effect and the effects of the two- and three-factor interactions of the number of measurements, IPD size and IPD percentage have statistically significant effects on TIF for an item bank of 200 items in a sample of 1000. Especially for an item bank of 200 items, IPD size factors significantly affect TIF values. This is high-level effect (Cohen, 1988). Although IPD size, IPD percentage and IPD size*IPD percentage have interaction effects on item banks of 500 and 1000 items, these effects are low-level (Cohen, 1988). While some studies on the impacts of IPD on TIF (Chan et al., 1999) support this finding, some studies argue that there are no statistically significant differences (Deng & Melican, 2010; Guo and Wang, 2003).

3.2. Findings of Comparison of Conditions for Sample Size 5000

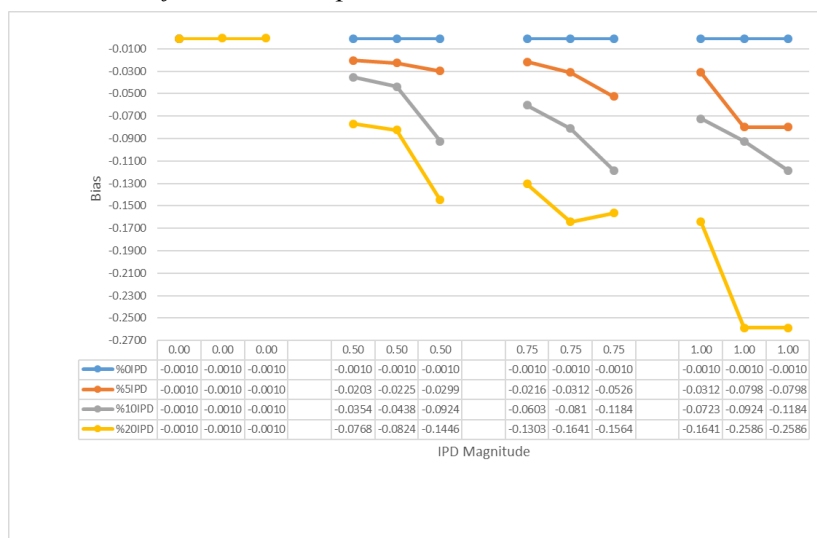
The findings for bias values, which constitute the first criterion for comparing independent variable conditions for a sample size of 5000, are shown in Figure 4. a, b and c.

Figures 4. a, b and c. Figures denoting comparison of bias values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is n=5000.

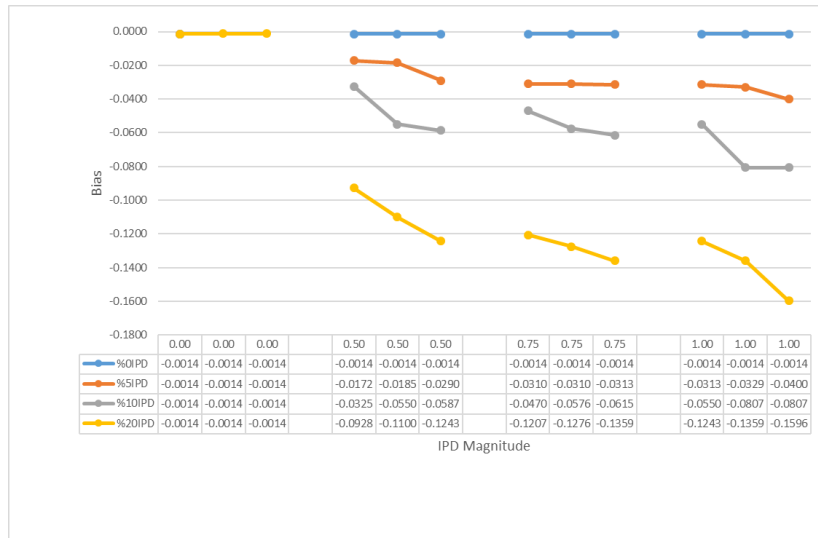
a. Bias values for item bank of 200 with sample size n=5000.



b. Bias values for item bank of 500 with sample size n=5000.



c. Bias values for item bank of 1000 with sample size n=5000.



When bias values were examined for a sample size of 5000, the increase in IPD size and IPD percentage for item bank sizes of 200, 500 and 1000 resulted in a tendency of ability estimation bias values obtained at three time points to grow in the negative direction as in 1000 sample size. Negative bias values mean that individuals' estimated ability values are lower than their real ability values. Since certain items in the item bank displayed IPD in the easier direction, we would have expected that individuals' estimated ability values to be higher than their real ability values. The reason could be either that IPD occurred only in the easier direction and only in the item difficulty parameter (Aksu Dünya, 2017; Wei, 2013), or individuals were provided items according to their ability level and may have answered them incorrectly (Chen, 2013; Risk, 2015; Rupp & Zumbo, 2003). The increase in the number of measurements, IPD size, and IPD percentage results in more biased ability estimations leading to a decrease in measurement precision. Studies in the literature indicate that IPD negatively affects bias values (Aksu Dünya, 2017; Chen, 2013; Risk, 2015; Rupp & Zumbo, 2003). IPD occurrences at and over 0.50 logit in particular significantly affect parameter estimations (Han & Wells, 2007; Wollack et al., 2005). Since this study also examined conditions with IPD at and over 0.50 logit, differences were obtained in bias values, albeit low.

Three-factor ANOVA results for independent samples, shown in Table 6, examine whether obtained differences were statistically significant according to the bias values discussed above. Three-factor ANOVA results for independent samples regarding bias show that both the main effect and effects of two-way and three-way interactions of the number of measurements, IPD size and IPD percentage for item bank sizes of 200, 500 and 1000 items have statistically significant effects on bias. These generally have low effect sizes (Cohen, 1988). The results of post-hoc analysis also revealed differences for every level of every factor. IPD percentage is the factor with the most impact on ability estimation bias among the variables within the scope of this study. Some studies in the literature have also reached similar findings (Abad et al., 2010; Babcock & Albano, 2012).

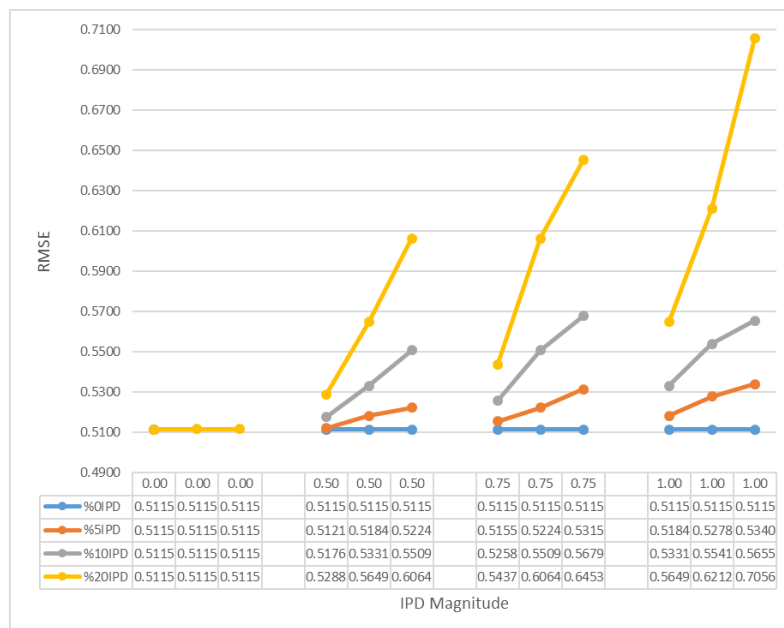
Table 6. Comparison of bias values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size n=5000.

Item Bank Size	Source of Variation	Sum of Squares	df	Mean of Squares	F	Effect Size (η^2)
200	IPD Size	0.383	3	0.128	18571.59*	0.10
	IPD Percentage	2.971	3	0.990	144063.18*	0.76
	Measurement	0.166	2	0.083	8049.31*	0.04
	IPD Size*IPD Percentage	0.117	9	0.013	5673.31*	0.03
	IPD Size*Measurement	0.011	6	0.002	533.39*	0.00
	IPD Percentage*Measurement	0.104	6	0.017	5042.94*	0.02
	IPD Size*IPD Percentage*Measurement	0.018	18	0.001	872.82*	0.00
	Error	0.098	4752	0.000		
Total		3.868	4799			
500	IPD Size	2.081	3	0.694	99888.00*	0.18
	IPD Percentage	6.748	3	2.249	323904.00*	0.60
	Measurement	1.069	2	0.535	51312.00*	0.09
	IPD Size*IPD Percentage	0.806	9	0.090	38688.00*	0.07
	IPD Size*Measurement	0.224	6	0.037	10752.00*	0.02
	IPD Percentage*Measurement	0.183	6	0.031	8784.00*	0.01
	IPD Size*IPD Percentage*Measurement	0.212	18	0.012	10176.00*	0.02
	Error	0.099	4752	0.000		
Total		11.422	4799			
1000	IPD Size	0.228	3	0.076	11169.65*	0.04
	IPD Percentage	4.403	3	1.468	215701.61*	0.88
	Measurement	0.160	2	0.080	7838.35*	0.03
	IPD Size*IPD Percentage	0.033	9	0.004	1616.66*	0.00
	IPD Size*Measurement	0.017	6	0.003	832.82*	0.00
	IPD Percentage*Measurement	0.049	6	0.008	2400.49*	0.01
	IPD Size*IPD Percentage*Measurement	0.007	18	0.000	342.93*	0.00
	Error	0.097	4752	0.000		
Total		4.994	4799			

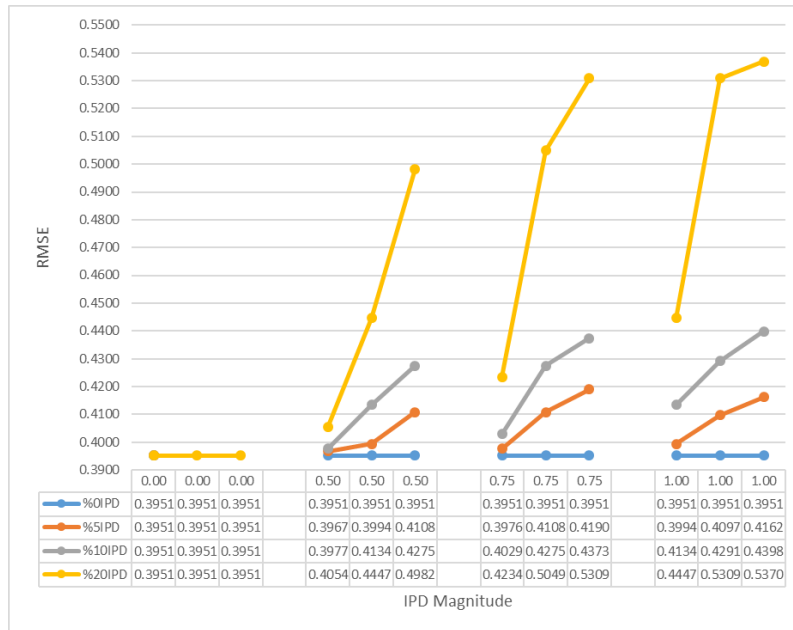
*p<.05

The findings for RMSE values, which constitute the second criterion for measurement precision where independent variable conditions are compared are shown in Figure 5. a, b and c.

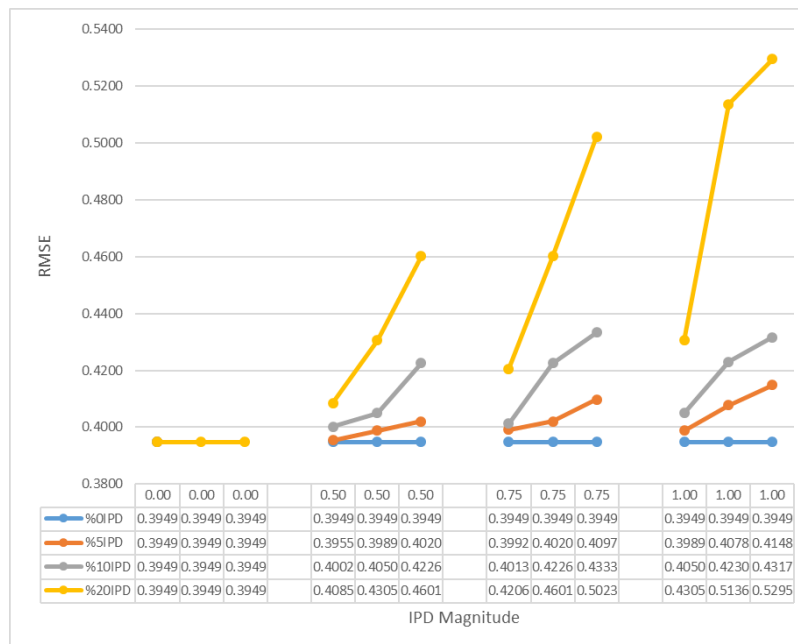
Figure 5. a, b and c. Figures denoting comparison of RMSE values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is n=5000. **a.** RMSE values for item bank of 200 with sample size n=5000.



b. RMSE values for item bank of 500 with sample size $n=5000$.



c. RMSE values for item bank of 1000 with sample size $n=5000$.



When RMSE values were examined, the increase in IPD size and IPD percentage for item bank sizes of 200, 500, and 1000 resulted in a tendency of ability estimation *RMSE* values obtained at three time points to increase. The increase in the number of measurements in IPD size, IPD percentage, and item banks containing IPD results in more erroneous ability estimations leading to a decrease in measurement precision. Some studies in the literature also show that IPD conditions increase error values (Aksu Dünya, 2017; Babcock & Albano, 2012; Chen, 2013; Risk, 2015; Wells et al., 2002).

When Figure 5. a and c were examined, it is found that the increase in item bank size of 200, 500, and 1000 items resulted in a tendency of *RMSE* values to decrease. A study by Risk (2015) used item bank sizes of 300, 500, and 1000 and observed that an increase in item bank size

resulted in a decrease in RMSE values. This signifies that an increase in item bank size results in a slight decrease in error values between real and estimated ability values.

Three-factor ANOVA results for independent samples, shown in Table 7, examine whether obtained differences were statistically significant according to above-mentioned RMSE values.

Table 7. Comparison of RMSE values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size $n=5000$.

Item Bank Size	Source of Variation	Sum of Squares	df	Mean of Squares	F	Effect Size (η^2)
200	IPD Size	0.408	3	0.136	7400.06*	0.07
	IPD Percentage	2.760	3	0.920	50059.24*	0.47
	Measurement	1.226	2	0.613	22236.46*	0.21
	IPD Size*IPD Percentage	0.267	9	0.030	4842.69*	0.04
	IPD Size*Measurement	0.044	6	0.007	798.05*	0.00
	IPD Percentage*Measurement	0.706	6	0.118	12805.01*	0.12
	IPD Size*IPD Percentage*Measurement	0.085	18	0.005	1541.68*	0.01
	Error	0.262	4752	0.000		
Total	5.758	4799				
500	IPD Size	0.160	3	0.053	11880.00*	0.03
	IPD Percentage	2.723	3	0.908	202182.75*	0.53
	Measurement	1.089	2	0.545	80858.25*	0.21
	IPD Size*IPD Percentage	0.087	9	0.010	6459.75*	0.01
	IPD Size*Measurement	0.151	6	0.025	11211.75*	0.03
	IPD Percentage*Measurement	0.602	6	0.100	44698.50*	0.11
	IPD Size*IPD Percentage*Measurement	0.206	18	0.011	15295.50*	0.04
	Error	0.064	4752	0.000		
Total	5.082	4799				
1000	IPD Size	0.299	3	0.100	22916.90*	0.08
	IPD Percentage	1.702	3	0.567	130450.06*	0.48
	Measurement	0.673	2	0.337	51582.19*	0.19
	IPD Size*IPD Percentage	0.240	9	0.027	18394.84*	0.06
	IPD Size*Measurement	0.063	6	0.011	4828.65*	0.02
	IPD Percentage*Measurement	0.372	6	0.062	28512.00*	0.10
	IPD Size*IPD Percentage*Measurement	0.064	18	0.004	4905.29*	0.02
	Error	0.062	4752	0.000		
Total	3.475	4799				

* $p < .05$

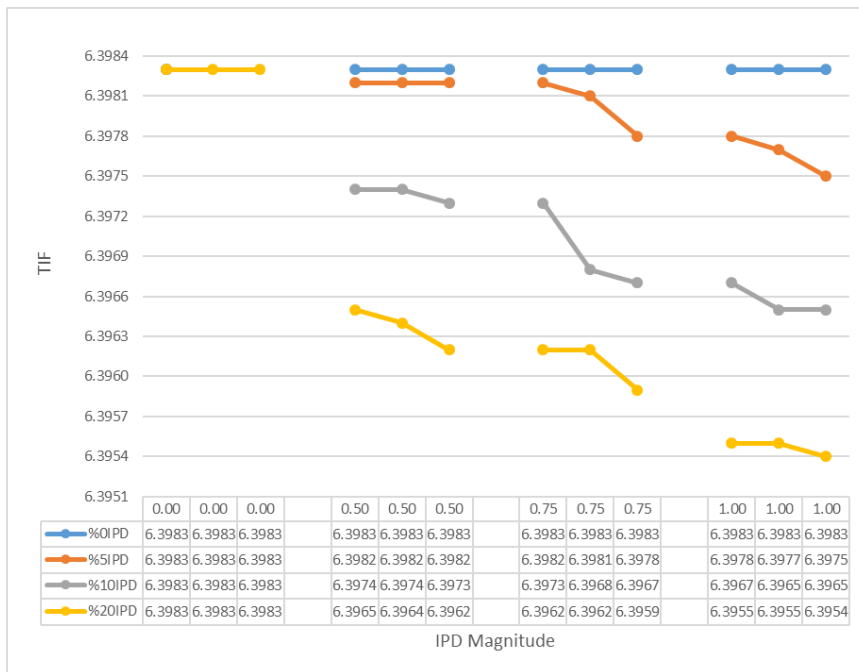
The three-factor ANOVA results for independent samples regarding RMSE values show that the main effect and the effects of the two- and three-way interactions of the number of measurements, IPD size, and IPD percentage for item bank sizes of 200, 500 and 1000 items have statistically significant effects on RMSE. These generally possess low and high effect sizes (Cohen, 1988). The results of post-hoc analysis also revealed differences for every level of every factor. IPD percentage is the factor with the most impact on ability estimation RMSE among the variables within the scope of this study. Some studies in the literature also support this finding (Aksu Dünya; 2017; Babcock & Albano, 2012; Risk, 2015). On the other hand, some studies argue that the impact of IPD on RMSE values was not statistically significant (Chen, 2013; Wells et al., 2002). For instance, Chen (2013) argued that although an increase in the percentage of items containing IPD in the item bank increased RMSE values, this increase was low-level and statistically insignificant.

The findings for TIF values, which constitute the third criterion for comparing independent variable conditions, are shown in Figure 6. a, b and c. When TIF values were examined for a sample size of 5000, the increase in IPD size and IPD percentage for item bank sizes of 200, 500, and 1000 resulted in a tendency of ability estimation TIF values obtained at three time points to decrease. The lowest TIF values were obtained for IPD size of 1.00, IPD percentage

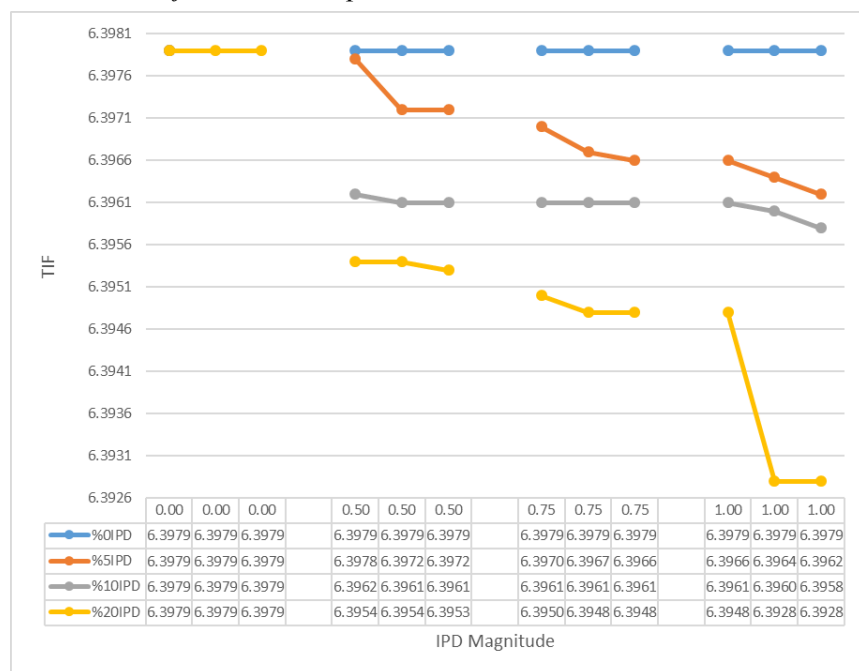
of 20, and at the third time point. Therefore, the increase in the number of measurements, IPD size, and IPD percentage results in a decrease in TIF, i.e., the amount of information the test provides for item banks of 200, 500, and 1000. This decreasing tendency does not change with an increase in item bank size. Similarly, TIF values are generally slightly higher in the 5000 sample than 1000 sample, but no increasing or decreasing trend was observed within each sample.

Figure 6. a, b and c. Figures denoting comparison of TIF values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is $n=5000$.

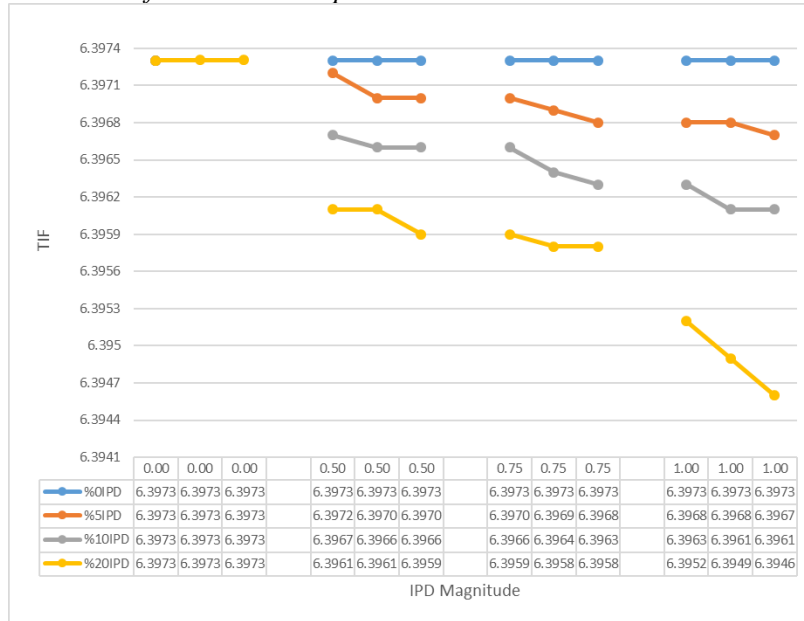
a. TIF values for item bank of 200 with sample size $n=5000$.



b. TIF values for item bank of 500 with sample size $n=5000$.



c. TIF values for item bank of 1000 with sample size n=5000.



Three-factor ANOVA results for independent samples, shown in Table 8, examine whether the differences obtained were statistically significant according to the TIF values discussed above.

Table 8. Comparison of TIF values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size n=5000.

Item Bank Size	Source of Variation	Sum of Squares	df	Mean of Squares	F	Effect Size (η^2)
200	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.032	3	0.010	38016.00*	0.45
	Measurement	0.002	2	0.001	2376.00*	0.03
	IPD Size*IPD Percentage	0.016	9	0.002	19008.00*	0.22
	IPD Size*Measurement	0.007	6	0.001	8316.00	-
	IPD Percentage*Measurement	0.003	6	0.000	3564.00	-
	IPD Size*IPD Percentage*Measurement	0.007	18	0.000	8316.00	-
	Error	0.004	4752	0.000		
	Total	0.071	4799			
500	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.002	3	0.000	2376.00*	0.01
	Measurement	0.000	2	0.000	0.00*	0.00
	IPD Size*IPD Percentage	0.000	9	0.000	0.00*	0.00
	IPD Size*Measurement	0.054	6	0.009	64152.00*	0.50
	IPD Percentage*Measurement	0.046	6	0.008	54648.00*	0.43
	IPD Size*IPD Percentage*Measurement	0.000	18	0.000	0.00*	0.00
	Error	0.004	4752	0.000		
	Total	0.106	4799			
1000	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.001	3	0.000	1584.00*	0.01
	Measurement	0.022	2	0.011	34848.00*	0.20
	IPD Size*IPD Percentage	0.067	9	0.007	106128.00*	0.63
	IPD Size*Measurement	0.001	6	0.000	1584.00	-
	IPD Percentage*Measurement	0.003	6	0.001	4752.00	-
	IPD Size*IPD Percentage*Measurement	0.010	18	0.000	15840.00	-
	Error	0.003	4752	0.000		
	Total	0.106	4799			

*p<.05

The three-factor ANOVA results for independent samples in terms of TIF values show that both the main effect and the effects of the two- and three-factor interactions of the number of measurements, IPD size, and IPD percentage have statistically significant effects on TIF for an item bank of 500 items and a sample size of 5000. However, these generally possess low effect sizes (Cohen, 1988). IPD size, IPD percentage, measurement and interaction effect of IPD size*IPD percentage for item bank sizes of 200 and 1000 on TIF values were statistically significant (Cohen, 1988). The results of the post-hoc analysis also revealed differences for every level of every factor that revealed significant differences. Therefore, the item bank containing IPD decreases the amount of information the test provides by increasing the errors. Although the TIF values for the 5000 sample size were higher than for the 1000 sample size, the samples themselves show neither an increasing nor a decreasing trend.

4. DISCUSSION and CONCLUSION

This study examines the impact of IPD on measurement precision and TIF in CAT administrations. When the results were examined in terms of measurement precision, the increase in the number of measurements, IPD size, and IPD percentage for item bank sizes of 200, 500 and 1000 items resulted in a decrease in measurement precision because items containing IPD in item bank led to drifts in ability estimations. The increase of IPD in the item bank resulted in bias values growing in the negative direction and RMSE values growing in the positive direction. The cause of positive RMSE values is the square in the RMSE formula. When compared with the baseline data set, the highest values of bias and RMSE were obtained at the third time point, with an IPD size of 1.00 and items containing an IPD percentage of 20. Measurement precision was calculated at its lowest point when conditions for IPD were at the highest point. Three-factor ANOVA for independent samples also revealed statistically significant results regarding these factors for measurement precision and indicated that the factor that affected measurement precision the most was the number of items containing IPD in the item bank. Research findings (Abad et al., 2010; Aksu Dünya, 2017; Babcock & Albano, 2012; Chan et al., 1999; McCoy, 2009; Risk, 2015; Wells et al., 2002) that examine the effects of IPD on measurement precision in CAT administrations are consistent with the finding that argues that the increase in IPD size results in a decrease in precision. While changes in IPD conditions affect measurement precision, an increase in sample size does not result in a changing pattern in either the positive or negative bias direction. The RMSE values were somewhat greater for the 5000-person sample, but no overall growing or declining trend was detected. The study has found that the factor that affected measurement precision the most was IPD percentage. While some studies contend that IPD percentage has the greatest impact on measurement precision (Babcock & Albano, 2012), others contend that IPD size (Risk, 2015), sample size, and IPD percentage (Wells et al., 2002) all influence measurement precision. Using the Rasch model, Risk (2015) examined the effect of various IPD conditions on measurement precision and discovered that the factor affecting measurement precision the most was IPD size rather than the number of items containing IPD in the item bank, but the effect was insignificant. Similarly, Wells et al. (2002) stated in their studies which used the 2PLM model, that sample size and IPD percentage were factors affecting ability estimations the most. It is worth noting that the simulated sample size, item bank size, IPD conditions and the IRT model vary in these studies. While the presence of items containing IPD in the item bank negatively affects measurement precision in CAT administrations, the factor negatively impacts the value depends on the IRT model, sample size, item bank size and IPD conditions.

When the results were examined in terms of TIF values, the increase in the number of items containing IPD in item bank, IPD size and number of measurements in CAT administrations resulted in a slight decrease in the amount of information the test provided. The highest TIF values under all conditions were obtained in the baseline data set not containing IPD, and the

lowest TIF values were obtained at the third time point with the highest rate of IPD conditions. As the number of measurements and IPD conditions increased, the amount of information provided by the test decreased. However, TIF values are generally marginally higher in the 5000-person sample than 1000-person sample, but neither an increasing nor a decreasing trend was observed within each sample. Similarly, there were no observed increasing or decreasing trend in TIF values as the item bank size changed. However, TIF values were affected by the number of measurements and IPD conditions. When the statistical significance of obtained TIF values was examined, statistically significant results were calculated mostly for the main effect and IPD size*IPD percentage factor. While some studies in the literature support the finding of the impact of IPD on TIF values (Chan et al., 1999), other studies obtained statistically insignificant differences (Deng & Melican, 2010; Guo & Wang, 2003). In a study by Guo and Wang (2003), which examined the impacts of parameter drift on CAT using real and simulated data, test characteristic curves were compared. However, since measurements were taken at two time points, very small differences were obtained in terms of TIF values, which were not significant.

In conclusion, this study has found that IPD under-examined conditions negatively affects measurement precision and TIF values. Although the IRT model and CAT administrations bring considerable advantages in ability estimations, the importance of developing tests for the item bank and reviewing items should be particularly emphasized to carry out ability estimations accurately. The chosen way of administrating tests and the models picked for use will only produce accurate results if high-quality items are available in the item bank, and these items can maintain this characteristic.

In light of the study's findings, the following recommendations can be made to researchers: This research was conducted using simulated data. Using test administrations with real data, the impact of IPD on the aforementioned factors could be examined. The examined samples in this study were generated using the normal distribution. However, since non-normally distributed extreme values are frequently encountered in real-world applications, the effects of IPD could also be examined under skewed distribution conditions. This study utilized the Rasch model, and there were no restrictions on item exposure. Consequently, the effects of IPD could also be examined by employing alternative IRT models and imposing various item exposure restrictions. This study examined the conditions under which all individuals may encounter IPD-containing items. However, only a subset of individuals may encounter IPD-containing items due to their prior test-taking experience or a change in the curriculum. Consequently, when IPD-containing items are given to a specific group of individuals in CAT applications, the effects of the condition on ability estimates could be investigated.

Acknowledgments

This paper was produced from the part of the first author's doctoral dissertation prepared under the supervision of the second and third author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Ankara University Social Sciences Sub-Ethical Committee, 22/04/2019, 05-181.

Authorship Contribution Statement

Merve Sahin Kursad: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Omay Cokluk Bokeoglu:** Investigation,

Resources, Methodology, Supervision, Writing-original draft. **Rahime Nukhet Cikrikci:** Investigation, Resources, Methodology, Supervision, Writing-original draft.

Orcid

Merve SAHIN KURSAD  <https://orcid.org/0000-0002-6591-0705>

Omay COKLUK BOKEOGLU  <https://orcid.org/0000-0002-3879-9204>

Rahime Nukhet CIKRIKCI  <https://orcid.org/0000-0003-0876-6644>

REFERENCES

- Abad, F.J., Olea, J., Aguado, D., Ponsoda, V., & Barrada, J.R. (2010). Deterioro de parámetros de los ítems en tests adaptativos informatizados: estudio con eCAT [Item parameter drift in computerized adaptive testing: Study with eCAT]. *Psicothema*, 22, 340-7.
- Aksu Dünya, B. (2017). *Item parameter drift in computer adaptive testing due to lack of content knowledge within sub-populations* [Doctoral dissertation, University of Illinois].
- Babcock, B., & Albano, A.D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, 36(7), 565-580. <https://doi.org/10.1177/0146621612455090>
- Babcock, B., & Weiss, D.J. (2012). Termination criteria in computerized adaptive test do variable-length CAT's provide efficient and effective measurement? *International Association for Computerized Adaptive Testing*, 1, 1-18. <http://dx.doi.org/10.7333%2Fjcat.v1i1.16>
- Barrada, J.R., Olea, J., Ponsoda, V., & Abad, F.J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*, 34, 438-452. <https://doi.org/10.1177/0146621610370152>
- Bergstrom, B.A., Stahl, J., & Netzky, B.A. (2001, April). *Factors that influence parameter drift* [Conference presentation] American Educational Research Association, Seattle, WA.
- Blais, J. & Raiche, G. (2002, April). Features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules, *International Objective Measurement Workshop*, New Orleans.
- Bock, D.B., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25(4), 275-285. <https://doi.org/10.1111/j.1745-3984.1988.tb00308.x>
- Burton, A., Altman, D.G., Royston, P., & Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25, 4279-4292. <https://doi.org/10.1002/sim.2673>
- Chan, K.Y., Drasgow, F., & Sawin, L.L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology*, 84(4), 610-619. <https://doi.org/10.1037/0021-9010.84.4.610>
- Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222. <https://doi.org/10.1177/01466219922031338>
- Chang, S.W., & Ansley, T.N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 40, 71-103. <https://www.jstor.org/stable/1435055>
- Chen, S.Y., Ankenmann, R.D., & Chang, H.H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, 24, 241-255. <https://doi.org/10.1177/01466210022031705>
- Chen, Q. (2013). *Remove or keep: linking items showing item parameter drift* [Unpublished Doctoral Dissertation]. Michigan State University.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum.

- Çikrikçi-Demirtaşlı, N. (1999). Psikometride yeni ufuklar: Bilgisayar ortamında bireye uyarlanmış test [New horizons in psychometrics: Individualized test in computer environment]. *Türk Psikoloji Bülteni*, 5(13), 31-36.
- Deng, H., Ansley, T., & Chang, H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement*, 47(2), 202-226. <https://doi.org/10.1111/j.1745-3984.2010.00109.x>
- Deng, H., & Melican, G. (2010, April). *An investigation of scale drift in computer adaptive test* [Conference presentation] Annual Meeting of National Council on Measurement in Education, San Diego, CA.
- Donoghue, J.R., & Isham, S.P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, 22(1), 33-51. <https://doi.org/10.1177/01466216980221002>
- Engen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, 23(3), 249-261. <https://doi.org/10.1177/01466219922031365>
- Eroğlu, M.G. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması* [Comparison of different test termination rules in terms of measurement precision and test length in computerized adaptive testing] [Unpublished Doctoral Dissertation]. Hacettepe University.
- Evans, J.J. (2010). *Comparability of examinee proficiency scores on Computer Adaptive Tests using real and simulated data* [Unpublished Doctoral dissertation]. The State University of New Jersey.
- Filho, N.H., Machado, W.L., & Damasio, B.F. (2014). Effects of statistical models and items difficulties on making trait-level inferences: A simulation study. *Psicologia Reflexão e Crítica*, 27(4). <https://doi.org/10.1590/1678-7153.201427407>
- Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, 20(4), 369-377. <https://doi.org/10.1111/j.1745-3984.1983.tb00214.x>
- Guo, F., & Wang, L. (2003, April). *Online calibration and scale stability of a CAT program* [Conference presentation] The Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Hagge, S., Woo, A., & Dickison, P. (2011, October). *Impact of item drift on candidate ability estimation* [Conference presentation] The Annual Conference of the International Association for Computerized Adaptive Testing, Pacific Grove, CA.
- Han, K.T., & Guo, F. (2011). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (R-11-02). Graduate Management Admission Council Research Report.
- Hatfield, J.P., & Nhoyvanisvong, A. (2005, April). *Parameter drift in a high-stakes computer adaptive licensure examination: An analysis of anchor items* [Conference presentation] The Annual Meeting of the American Educational Research Association, Montreal, Canada.
- Huang, C., & Shyu, C. (2003, April). *The impact of item parameter drift on equating* [Conference presentation] National Council on Measurement in Education, Chicago, IL.
- Jiang, G., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing*, 9(4), 283-309. <https://doi.org/10.1080/15305050903351901>
- Jones, P.E., & Smith, R.W. (2006, April) *Item parameter drift in certification exams and its impact on pass-fail decision making* [Conference presentation] National Council of Measurement in Education, San Francisco, CA.

- Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability* [Unpublished Doctoral Dissertation] Middle East Technical University.
- Kaptan, F. (1993). *Yetenek kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kağıt-kalem testi uygulamasının karşılaştırılması [Comparison of adaptive (individualized) test application and traditional paper-pencil test application in ability estimation]* [Unpublished Doctoral Dissertation]. Hacettepe University.
- Keller, A.L. (2000). *Ability estimation procedures in computerized adaptive testing* (Technical Report). American Institute of Certified Public Accountants-AICPA Research Consortium-Examination Teams.
- Kezer, F. (2013). *Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması [Comparison of computerized adaptive testing strategies]* [Unpublished Doctoral Dissertation]. Ankara University.
- Kim, S.H., & Cohen, A.S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, 29, 51–66. <https://www.jstor.org/stable/1434776>
- Kingsbury, G.G., & Wise, S.L. (2011). Creating a K-12 adaptive test: Examining the stability of item parameter estimates and measurement scales. *Journal of Applied Testing Technology*, 12.
- Kingsbury, G.G., & Zara, A.R. (2009). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2(4), 359-375. https://doi.org/10.1207/s15324818ame0204_6
- Köse, İ.A. & Başaran, İ. (2021). 2 parametrelili lojistik modelde normal dağılım ihlalinin madde parametre kestirimine etkisinin incelenmesi [Investigation of the effect of different ability distributions on item parameter estimation under two-parameter logistics model]. *Journal of Digital Measurement and Evaluation Research*, 1(1), 01-21. <https://doi.org/10.29329/dmer.2021.285.1>
- Li, X. (2008). An investigation of the item parameter drift in the examination for the certificate of proficiency in English (ECPE). *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 6, 1–28.
- Linda, T. (1996, April). *A comparison of the traditional maximum information method and the global information method in CAT item selection* [Conference presentation] National Council on Measurement in Education, New York.
- Linden, W.J., & Glas, G.A.W. (2002). *Computerized adaptive testing: Theory and practice*. Kluwer Academic Publishers.
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates Publishers.
- McCoy, K.M. (2009). *The impact of item parameter drift on examinee ability measures in a computer adaptive environment* [Unpublished Doctoral Dissertation]. University of Illinois.
- McDonald, P.L. (2002). *Computer adaptive test for measuring personality factors using item response theory* [Unpublished Doctoral Dissertation]. The University Western of Ontario.
- Meng, H., Steinkamp, S., & Matthews-Lopez, J. (2010). *An investigation of item parameter drift in computer adaptive testing* [Conference presentation] The Annual Meeting of the National Council on Measurement in Education, Denver, CO.
- Meyers, J., Miller, G.E., & Way, W.D. (2009, April). *Item position and item difficulty change in an IRT based common item equating design* [Conference presentation] The American Educational Research Association, San Francisco, CA.
- Nydick, S.W. (2015). An R package for simulating IRT-based computerized adaptive tests.
- Patton, J.M., Cheng, Y., Yuan, K.H., & Diao (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, 37(1), 24–40. <https://doi.org/10.1177/0146621612461727>

- Ranganathan, K., & Foster, I. (2003). Simulation studies of computation and data scheduling algorithms for data grids. *Journal of Grid Computing*, 1, 53-62. <https://doi.org/10.1023/A:1024035627870>
- Reckase, M.D. (2011). Computerized adaptive assessment (CAA): The way forward. In *The road ahead for state assessments, policy analysis for California education and Rennie Center for Education Research & Policy* (pp.1-11). Rennie Center for Education Research & Policy.
- Risk, N.M. (2015). *The impact of item parameter drift in computer adaptive testing (CAT)* [Unpublished Doctoral Dissertation]. University of Illinois.
- Rudner, L.M., & Guo, F. (2011). Computer adaptive testing for small scale programs and instructional systems. *Graduate Management Council (GMAC)*, 11(01), 6-10.
- Rupp, A.A., & Zumbo, B.D. (2003). *Bias coefficients for lack of invariance in unidimensional IRT models*. Vancouver: University of British Columbia.
- Rupp, A.A., & Zumbo, B.D. (2004). A note on how to quantify and report whether item parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement*, 64, 588-599. <https://doi.org/10.1177/0013164403261051>
- Schulz, W., & Fraillon, J. (2009, September). The analysis of measurement equivalence in international studies using the rasch model [Conference presentation] The European Conference on Educational Research (ECER), Vienna.
- Scullard, M.G. (2007). *Application of item response theory based computerized adaptive testing to the strong interest inventory* [Unpublished Doctoral Dissertation]. University of Minnesota.
- Segall, D.O. (2004). Computerized adaptive testing. In K. Kempf-Lenard (Ed.), *The Encyclopedia of social measurement*. Academic Press.
- Song, T., & Arce-Ferrer, A. (2009, April). *Comparing IPD detection approaches in common-item nonequivalent group equating design* [Conference presentation] The Annual Meeting of the National Council on Measurement, San Diego, CA.
- Stahl, J.A., & Muckle, T. (2007, April). *Investigating displacement in the Winsteps Rasch calibration application* [Conference presentation] The Annual Meeting of the American Educational Research Association, Chicago, IL.
- Sulak, S. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında kullanılan madde seçme yöntemlerinin karşılaştırılması [Comparison of item selection methods in computerized adaptive testing]* [Unpublished Doctoral Dissertation]. Hacettepe University.
- Svetina, D., Crawford, A.V., Levy, R., Green, S.B., Scott, L., Thompson, M., Gorin, J.S., Fay, D., & Kunze, K.L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling*, 55(4), 335-360.
- Şahin, A. (2012). *Madde tepki kuramında test uzunluğu ve örneklem büyüklüğünün model veri uyumu, madde parametreleri ve standart hata değerlerine etkisinin incelenmesi [An investigation on the effects of test length and sample size in item response theory on model-data fit, item parameters and standard error values]* [Unpublished Doctoral Dissertation]. Hacettepe University.
- Veldkamp, B.P., & Linden van der, W. (2006) Designing item pool for computerized adaptive testing. In *Designing Item Pools* (pp.149-166). University of Twente.
- Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, 12(1), 15-20. <https://doi.org/10.1111/j.1745-3992.1993.tb00519.x>
- Wainer, H., Dorans, N.J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (2010). *Computerized adaptive testing: A primer*. Lawrence Erlbaum Associates Publishers.

- Wang, T. (1997, March). *Essential unbiased EAP estimates in computerized adaptive testing* [Conference presentation] The American Educational Association, Chicago, IL.
- Wang, H-P., Kuo, B-C., Tsai, Y-H., & Liao, C-H. (2012). A Cerf-Based computerized testing system for Chinese proficiency. *TOJET: The Turkish Journal of Educational Technology*, 11(4), 1–12.
- Weiss, D.J., & Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4), 361–375. <http://www.jstor.org/stable/1434587>
- Wells, C.S., Subkoviak, M.J., & Serlin, R.C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26(1), 77-87. <https://doi.org/10.1177/0146621602261005>
- Wise, S.L., & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, 21(2000), 135-155.
- Witt, E.A., Stahl, J.A., Bergstrom, B.A., & Muckle, T. (2003, April). *Impact of item drift with nonnormal distributions* [Conference presentation] The Annual Meeting of the American Educational Research Association, Chicago, IL.
- Wollack, J.A., Sung, H.J., & Kang, T. (2005) *Longitudinal effects of item parameter drift* [Conference presentation] The Annual Meeting of the National Council on Measurement in Education, Montreal, CA.
- Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement*, 37(1), 3-23. <https://doi.org/10.1177/0146621612455687>
- Yi, Q., Wang, T., & Ban, J.C. (2001). Effects of scale transformation and test termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement*, 38, 267-292. <https://doi.org/10.1111/j.17453984.2001.tb01127.x>

The Effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests

Basak Erdem-Kara^{1,*}, Nuri Dogan²

¹Anadolu University, Faculty of Education, Department of Educational Sciences, Eskisehir, Türkiye

²Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

ARTICLE HISTORY

Received: Apr. 19, 2022

Revised: Aug. 08, 2022

Accepted: Aug. 19, 2022

Keywords:

Computer adaptive test,
Multi-stage test,
Differential item
functioning.

Abstract: Recently, adaptive test approaches have become a viable alternative to traditional fixed-item tests. The main advantage of adaptive tests is that they reach desired measurement precision with fewer items. However, fewer items mean that each item has a more significant effect on ability estimation and therefore those tests are open to more consequential results from any flaw in an item. So, any items indicating differential item functioning (DIF) may play an important role in examinees' test scores. This study, therefore, aimed to investigate the effect of DIF items on the performance of computer adaptive and multi-stage tests. For this purpose, different test designs were tested under different test lengths and ratios of DIF items using Monte Carlo simulation. As a result, it was seen that computer adaptive test (CAT) designs had the best measurement precision over all conditions. When multi-stage test (MST) panel designs were compared, it was found that the 1-3-3 design had higher measurement precision in most of the conditions; however, the findings were not enough to say that 1-3-3 design performed better than the 1-2-4 design. Furthermore, CAT was found to be the least affected design by the increase of ratio of DIF items. MST designs were affected by that increment especially in the 10-item length test.

1. INTRODUCTION

Traditional linear tests that have been the milestone of educational assessment since the 1900's are generally administered using paper-pencil and have been a popular way to measure examinees' knowledge, skills, and abilities (Weiss & Kingsbury, 1984; Yan et al., 2014). However, especially over the past 40 years, computer-based tests have gained popularity over linear tests thanks to great advances in computer technology, thereby becoming a viable alternative to those paper-pencil tests (Keng, 2008; Luecht & Sireci, 2011; Magis et al., 2017; Yan et al., 2014). According to Yan et al. (2014) computer-based tests can be classified into three main groups; namely, linear, adaptive, or multi-stage.

Computer-based linear tests are the computerized version of traditional linear tests. As in linear tests, all individuals answer the same items in these tests, and the test length is fixed (Magis et al., 2017; Sarı, 2016; Yan et al., 2014). On the other hand, the primary purpose of computer adaptive tests (CAT) is to select items from the item pool so as to match the ability level of the

*CONTACT: Basak ERDEM KARA ✉ basakerdem@anadolu.edu.tr 📍 Anadolu University, Faculty of Education, Department of Educational Sciences, Eskisehir, Türkiye

individual and to ensure that test is neither too easy nor too difficult for the individual (Thai, 2015; Yan et al., 2014, Zheng & Chang, 2014). In the process, an item is administered and answered and the individual's ability level (θ) at that point is estimated according to the answer. Depending on that estimated θ , the next item is chosen from the pool and administered. Until the stopping criteria are met, the process goes on (Tay, 2015; Weiss & Kingsbury, 1984). Individuals only face items convenient to their ability levels and do not spend time with items which are too easy or too difficult for them. Thus, the main advantage of CAT over linear tests is that they reach the desired measurement precision with fewer items (Wainer, 2000; Wang, 2013; Wang, 2017).

The other type of computer-based tests which has become popular, especially in recent years, is multi-stage test (MST), which combines the many advantages of linear tests and CAT while minimizing their disadvantages (Hendrickson, 2007; Magis et al., 2017). MST which can be considered as a variation of adaptive testing differs from CAT in test adaptation level. While test adaptation occurs at item level in CAT, adaptation occurs at item set (module) level in multi-stage testing (Hendrickson, 2007; Yan, 2010). In MST, a set of items which is named as the module is administered to the examinee, examinees' ability is estimated based on his/her responses to that module, and s/he is routed to the next module at the next stage (Hendrickson, 2007; Wang, Lin, Chang, & Douglas, 2016). In MST, each module can be assembled so as to have desired contextual and statistical specifications; thus, test developers have more control over the construction of the desired test form when compared to CAT. Although, MST has less adaptation points than those of CAT, they provide more efficient test assembly and more controlled content balancing. Furthermore, MST allows some item review and change previous answers within each module. However, going back to previous stages and reviewing items in previous module/s are not allowed in MST. (Hendrickson, 2007; Sarı & Huggins-Manley, 2017; Wainer, 2000; Wang, 2017). In addition to their advantages, MST also has some disadvantages such as requiring more items to get the same measurement precision with CAT (Berger, Verschoor, Eggen & Moser, 2019). Besides, since MST modules are designed so as to be at optimum difficulty only at target ability levels (e.g., three levels at low, medium, and high proficiency), final ability estimations may not be as accurate as CAT designs (Rome, 2017).

The increase in computer-based testing application has brought some problems especially in test fairness issue (Chu & Lai, 2013; Gierl, Lai & Li, 2013; Zwick, 2010). Test fairness and equity issues are related with items presenting some bias towards a specific group of students. Non-bias items only measure ability of individuals that is intended to be measured without being affected by unrelated factors such as gender, socio-economic status, etc. On the other hand, bias items are affected by those factors which are not related with the characteristic which is intended to be measured. Because test results are used in critical decision-making situations that may affect individuals' future, test fairness becomes even more significant (Camilli & Shepard, 1994; Crocker & Algina, 1986; Hambleton & Swaminathan, 1991). Differential item functioning analyses are one of the most popular methods used to get information on the bias. Potentially problematic items are identified with DIF analyses and expert opinions are obtained on whether those items are really problematic or not (Zumbo, 1999).

1.1. DIF and Adaptive Testing

The quality of adaptive testing applications largely depends on item pool quality (Han & Guo, 2011). Therefore, large item pools should be developed for those applications and each item in that pool should be checked in order to ensure that they satisfy the main fairness and equity issues (Gierl, Lai & Li, 2013). However, even if the item writing process is well planned and carefully designed, it is not easy to avoid the effects of DIF completely. Many factors which are not related with an item such as computer familiarity, testing environment, physical impairments, etc. may cause DIF (Birdsall, 2011). Independent of the item content, the context

in which the item is presented, for instance, item order, may also affect item parameters and may become a source for DIF (National Research Council, 1999). Besides, although items in the pool have no DIF initially, some may become DIF items over time. As a result of repeatedly usage of items over time, they may become known for other individuals prior to their administration. Even if this is not the case, the interaction between the item and test taker may change because of several reasons, which is known as item parameter drift. Therefore, the changing interaction between an item and a test taker may cause different item characteristics than initially calibrated item characteristics (Aksu-Dunya, 2017; Han & Guo, 2011). Parameter drift on items could be defined as a kind of DIF since items behave differently in groups which are involved in different testing applications (Aksu-Dunya, 2017; Babcock & Albano, 2012). Item parameter drift is a serious threat to validity and fairness (Han & Guo, 2011). DIF analyses may be more important for adaptive testing applications than they are for linear tests. Since the number of items administered in adaptive tests is fewer than in linear tests, each item has greater effect on final ability estimation. Therefore, any flaw in an item may cause more consequential results (Zwick, 2010; Gierl et al., 2013; Zwick & Bridgeman, 2014). So, DIF items may play an important role in examinees' test scores. Besides, performing the test application via computer may reveal some possible sources of DIF such as computer familiarity, anxiety, and environment that are not found in traditional tests (Zwick, 2010). These factors have increased the importance of DIF analyses in computer adaptive tests. Steinberg et al. (2000) stated that adaptive tests may be more sensitive to the effects of DIF on validity than linear tests. In addition, the presence of bias may affect the order of administration of the items because the next item/module in CAT and MST is determined according to the answers to the previous item/module (Zwick, 2010). It is important to note that concentration of biased items on certain modules for the MST may also pose a problem.

1.2. Purpose of the Study

Despite the importance of the existence of DIF items in adaptive testing is known, DIF studies in adaptive tests are limited to a few studies in the literature (Chu & Lai, 2013; Gierl et al., 2013; Lei, Chen & Yu, 2006; Piromsombat, 2014). Besides, those studies were limited to the investigation and comparison of DIF detection methods under different conditions (Chu & Lai, 2013; Gierl et al., 2013; Lei, Chen & Yu, 2006) and the investigation of the effect of DIF items on ability estimation on CAT (Piromsombat, 2014). No studies were found in the literature focusing on comparison of CAT and MST approaches in case of the presence of DIF items in the test. The current study aims to investigate the performances of two adaptive testing approaches, CAT and MST, in case of the presence of DIF items under different conditions. Therefore, the results of the study are likely to contribute to the literature focusing on DIF in adaptive testing applications. To this end an answer for the following research question is sought in the context of this research:

- *How does the test performance of CAT and MST change in case of the presence of DIF items on the test under different test lengths (10-20-30-40 item), test designs (CAT, 1-3-3 MST and 1-2-4 MST), and ratio of DIF items (10%, 20% and 30%)?*

2. METHOD

Within the scope of the research, it is aimed to examine the effect of the inclusion of items that have differential item functioning (DIF) in the test on the effectiveness of CAT and MST under different conditions. The data used in the research were generated by the simulation method and different test designs were compared under different conditions in a controlled manner. The related study is a Monte Carlo simulation study in which the data are simulated. Simulation data were preferred because it was difficult to meet all the conditions discussed in the study simultaneously in real data.

2.1. Research Design

In this study, test performances of three different adaptive test designs (CAT, 1-3-3 MST and 1-2-4 MST) were compared in case the test consists of DIF items. Those MST designs were some of the most popular ones. Two-stage test designs have only one adaptation point which may make them open for routing errors more (Yan, et al., 2014; Zenisky et. al, 2010). On the other hand, it was stated that more than three stages add little to the accuracy of ability estimations and increase the complexity of test designs (Yan et al., 2014). In general, maximum four modules in one stage and three stages were thought to be enough (Armstrong et al., 2004; Zenisky et al., 2010). The preferred test designs in this study were among the most popular ones used in the literature.

The manipulated factors were test length with four levels and ratio of DIF items in the test with three levels. Three different test designs (CAT, MST 1-2-4 and MST 1-3-3) were compared under three different DIF item ratio and four different test length conditions. All manipulated conditions were fully crossed within each of three test designs, which resulted in 36 conditions (Three Test Designs, Four Test Length and Three DIF Item Ratio). For each condition, 30 replications were performed and the whole simulation processes were performed by using R programming language (R Core Team, 2018). Detailed information on simulation processes is given as follows:

2.2. Data Generation

Five thousand examinees were randomly generated based on standard normal distribution and the same theta values were used for all test designs. Generated theta values were restricted to be generated between -3 and 3 in order to eliminate the effect of outliers. Besides, an item pool of 600 items was generated using the three-parameter logistic model. Discrimination, difficulty, and guessing parameters were randomly sampled from Uniform (0.5, 2.0), N (0, 1) and Uniform (0, 0.25) distributions, respectively. Difficulty parameters were restricted to be in the range of [-3, 3]. Descriptive statistics related to item pool are given in [Table 1](#).

Table 1. Descriptive statistics of item pool.

	a parameter	b parameter	c parameters
N	600	600	600
Mean	1.268	-0.097	0.125
Standard Deviation	0.442	1.22	0.074
Minimum	0.501	-2.967	0.0002
Maximum	1.999	2.988	0.249

As can be seen in [Table 1](#), the discrimination values (a parameter) had a minimum value of 0.501 and a maximum of 1.999 with a mean of 1.268. The item pool had items with a wide range of discrimination. Item difficulties ranged from -2.967 to 2.988 with a mean of -0.097 indicating that the item pool had items with a wide difficulty range in the specified range of [-3, 3]. Guessing parameter ranged from 0.0002 to 0.249 with a mean of 0.125. When the test information function of that item pool was examined, it was seen that items in the pool gave high information especially around the point where the ability level was 0 and covered the [-3, 3] ability range as intended.

2.2.1. Generation of DIF items

Item pool was developed to have 200 items on each difficulty level and 600 items in total. After that, 20% of the items on each difficulty level were randomly selected and rendered into DIF items. In order to make those items indicate DIF, +1 constant was added to the initial b

parameters and that value was considered as the focal group b parameter. For DIF items, the difference between b parameters of focal and reference groups was set as +1 ($b_{\text{focal}} - b_{\text{reference}} = 1$) and those items always worked in favor of the reference group, which means that all of them had uniform DIF (U-DIF). As a result, there were 40 U-DIF and 160 Non-DIF items in each level and 120 U-DIF and 480 Non-DIF items in total.

2.3. CAT and MST Simulations

After generation of item parameters, theta values and formation of DIF items, CAT, and MST environments were constructed. For CAT and MST simulations conducted on the same item pool and the same theta values, the commonly manipulated variable is the test length (10-20-30-40) and the rate of DIF items in the test (10% - 20% - 30%). Besides, panel designs (1-2-4 and 1-3-3) were manipulated within MST simulation. RMSE, bias, and correlation values were calculated and averaged across 30 replications and the performance of simulations was interpreted based on those values. In order to make better comparisons; maximum item exposure rates, IRT model, ability estimation method, and item/module selection method were fixed for both CAT and MST. Maximum Fisher Information (MFI) method was used in item selection and Expected A Posteriori (EAP) estimation method was used in ability estimation for both CAT and MST. MFI method is preferred since it provides the selection of the item that provides the maximum information each time. Although this method is quite popular, it has the disadvantage that items with high discrimination levels are chosen more because they provide more information and choosing those same items over and over leads to item exposure problem (Hambleton, Jac ve Pieters, 2000; van der Linden & Pashley, 2010; Wang, 2017); hence, this method should be used carefully. Controlling the item exposure can be an effective method to prevent this situation. As another method, Hambleton et al. (2000) suggested that the item be chosen randomly among items that provide maximum information at the relevant skill level. In our study, both of those methods were implemented. The maximum item use rate was fixed as 0.25 for CAT and four separate parallel panels were created for MST to ensure that the maximum item use rate was 0.25. Besides, 'randomesque' method was used and instead of choosing the most informative item, items were randomly selected from among the most informative ones at that ability level. As the ability estimation method, two most common methods are maximum likelihood (MLE) and Bayesian methods. However, MLE can be problematic since it cannot be estimated for individuals who answer all items correctly or incorrectly. This is particularly problematic for the early stages of computer adaptive tests and is not recommended when the test length is short. The use of MLE is not recommended until a true or false answer is received (Hambleton & Swaminathan, 1991; Wang, 2017). Bayesian methods are more consistent for short-length tests. The combination that is generally suggested for item selection and ability estimation in adaptive tests is the EAP estimation in ability estimation together with the maximum information method in item selection (van der Linden, 2008; van der Linden & Pashley, 2010). That is why, MFI with EAP combination was preferred in this study.

All simulation processes were carried out with the help of the catR and mstR packages that are conjugate of each other. Detailed explanations of simulations are given as follows:

2.3.1. CAT simulations

CAT environment was created via catR package and 12 different conditions in total (4 test lengths x 3 ratios of items with DIF), including four different test lengths (10-20-30-40) and three different ratios of items with the DIF (10 % - 20% - 30%) were examined as specified earlier. As a starting rule, the initial ability level was set to 0 and this value was used for each condition. According to this rule, the initial ability levels of individuals were accepted as '0' (zero) and the first item that the individual would encounter was determined accordingly.

2.3.2. MST simulations

To create MST environment, xxIRT and mstR packages were used. In MST simulation, 24 different conditions in total, including four different test lengths, three different ratios of items with DIF, and two other test designs (1-2-4 and 1-3-3) were examined. In the 1-3-3 MST design, a single module was used in the first stage, while there were three modules each in the second and third stages. For that 1-3-3 design, which included 7 modules in total, a single module common to all individuals was created at the first stage, and the difficulty level of this module was determined as medium. The three modules in the second and third stages had three different difficulty levels (easy, medium, and hard). Each individual answered three modules, each one from a different stage, in total. Similarly, in the 1-2-4 panel design, individuals responded to a total of three modules. Individuals who answered a single module in the first stage were directed to one of the two different modules in the second stage according to the ability estimations obtained from the first module. After completing this second stage, they were directed to one of the four modules in the third stage, considering the abilities estimated at the end of the second stage. The number of items in the modules and the number of items required to form a panel differed according to test lengths and panel design and are presented in detail in Table 2.

Table 2. Number of items in modules and panels.

Panel Design		Test Length			
		10	20	30	40
1-3-3	Module length	3-3-4	6-7-7	10-10-10	13-13-14
	Number of items used in panel	24	48	70	94
1-2-4	Module length	3-3-4	6-7-7	10-10-10	13-13-14
	Number of items used in panel	25	48	70	95

Since the modules in a panel are at different ability levels, the number of items used while creating the related panel is more than the test length, e.g., in the 1-3-3 design, in the condition that the test length was 40, individuals answered a total of 40 items, 13 each in the first two stages, and 14 in the last stage. However, while 13 items were needed in the first stage, 39 and 42 items were needed in the second and third stages for the modules at three different levels, respectively. As a result, a total of 94 items were used. For both panel designs, 10%, 20%, and 30% of the items in the modules in the second stage were selected among the items with DIF, e.g., in the case where the test length was 10, under the condition that the rate of items with DIF is 20%, 8 of the items were selected among the items that did not show DIF and 2 of them showed DIF. It was ensured that the selected 2 DIF items were included in the modules of the second stage.

Within the scope of the study, four different panels were created, so that the maximum panel, module, and item exposure became comparable with the CAT. Four different panels were obtained through an open source "mixed integer linear programming solver" (lp_solve 5.5) included in the xxIRT package, and it was ensured that the items used in one panel were not included in the other panels. "Bottom-up" method was used in the creation of the panels. In this method, firstly, four different parallel forms were created for each module. In order to ensure that the modules were parallel, information function targets were determined at the module level and the modules were structured to meet those targets. The items in the modules were chosen to provide maximum information at the specified skill levels. After the construction of four parallel forms for each module, those modules were assigned to the panels randomly and parallel panels were obtained. Thanks to the parallelism of the constructed modules, these modules could be used alternately between the panels (Yan et al., 2014).

2.4. Data Analysis

In the analysis of data, Root Mean Square Error (RMSE), bias, and correlation (ρ) values between estimated and true ability parameters were used to evaluate the results obtained from CAT and MST. $\hat{\theta}_j$ represents the estimated ability parameter, θ_j represents the true ability parameter, and N represents the total number of individuals. RMSE and bias values were calculated with the help of the following formulas:

$$RMSE = \sqrt{\frac{\sum_{j=1}^N (\hat{\theta}_j - \theta_j)^2}{N}} \quad Bias = \frac{\sum_{j=1}^N |\hat{\theta}_j - \theta_j|}{N} \tag{1}$$

The correlation value was obtained by the following formula, with the standard error values of the estimated ($\sigma_{\hat{\theta}_j}$) and true (σ_{θ_j}) ability parameters.

$$\rho_{\hat{\theta}_j, \theta_j} = \frac{cov(\hat{\theta}_j, \theta_j)}{\sigma_{\hat{\theta}_j} \sigma_{\theta_j}} \tag{2}$$

RMSE, bias, and correlation values were calculated for each of the 30 replications and interpretations were made based on the average of those values. Based on the calculated values, evaluations were made as to which of the two MST and one CAT application gave higher measurement accuracy than the others under different conditions. After those evaluations, whether the differences between the test designs (CAT, MST 1-3-3 and MST 1-2-4) reached a significant level were examined by ANOVA analysis. Post-Hoc analyses were made for the design groups that differed significantly from each other and the results were interpreted.

3. RESULTS

In this section, results of CAT and MST simulations are presented in detail. Findings are evaluated for each condition under different DIF item ratio.

3.1. Results of the Condition that the Ratio of DIF Items is 10%

In this part, ratio of DIF items in the test was fixed at 10% and the performance of three test design was examined under different test lengths. RMSE, bias and correlation values are presented in Table 3.

Table 3. RMSE, bias and correlation values of test designs under test length and DIF item ratio.

DIF Item Ratio	Test Length	RMSE			Bias			Correlation		
		CAT	MST (1-3-3)	MST (1-2-4)	CAT	MST (1-3-3)	MST (1-2-4)	CAT	MST (1-3-3)	MST (1-2-4)
10%	10 items	0.269	0.382	0.389	0.212	0.299	0.307	0.963	0.924	0.921
	20 items	0.192	0.282	0.306	0.151	0.221	0.240	0.982	0.963	0.953
	30 items	0.164	0.246	0.253	0.130	0.193	0.200	0.987	0.974	0.969
	40 items	0.155	0.236	0.238	0.122	0.184	0.187	0.989	0.977	0.974
20%	10 items	0.268	0.408	0.400	0.211	0.318	0.312	0.963	0.913	0.916
	20 items	0.192	0.284	0.320	0.151	0.223	0.252	0.982	0.962	0.949
	30 items	0.164	0.252	0.270	0.130	0.197	0.213	0.987	0.971	0.964
	40 items	0.153	0.242	0.241	0.121	0.190	0.191	0.989	0.975	0.973
30%	10 items	0.269	0.448	0.451	0.212	0.348	0.352	0.963	0.894	0.892
	20 items	0.194	0.301	0.303	0.153	0.237	0.238	0.981	0.955	0.954
	30 items	0.166	0.276	0.269	0.131	0.218	0.212	0.987	0.962	0.965
	40 items	0.153	0.246	0.259	0.121	0.194	0.204	0.989	0.970	0.967

As indicated in [Table 3](#), RMSE values range from [0.155, 0.269] for CAT design, [0.236, 0.382] for MST 1-3-3 design, and [0.238, 0.389] for MST 1-2-4 design. The bias values ranged between [0.122 - 0.212] for the CAT design, [0.184-0.299] for the MST 1-3-3 design, and [0.187-0.307] for the MST 1-2-4 design. It was seen that CAT application had the lowest and MST 1-2-4 application had the highest RMSE and bias values for all test lengths. However, in MST 1-3-3 and 1-2-4 designs, those values seemed to be quite close to each other throughout all test lengths. In addition, it was observed that as the number of items increased, the bias values decreased and the difference between the designs decreased. Finally, when the correlation values were examined, it was observed that those values varied between the range of [0.963-0.989] for CAT, [0.924-0.977] for the MST 1-3-3 design and [0.921- 0.974] for the MST 1-2-4 designs. Looking at the correlation values in [Table 3](#), it was determined that the design with the highest correlation value throughout all test lengths was CAT and the design with the lowest correlation value was the MST 1-2-4 design. Correlation values increased as the number of items increased for all test designs and got closer to each other.

3.2. Results of the Condition that the Ratio of DIF Items is 20%

In this part, ratio of DIF items in the test was fixed at 20% and the performance of three test design was examined under different test lengths. As indicated in [Table 3](#), RMSE values range from [0.153, 0.268] for CAT, [0.242, 0.408] for MST 1-3-3, and [0.241, 0.400] for MST 1-2-4 design. The lowest RMSE value for all test lengths was obtained in CAT, while the highest RMSE value was found in the MST 1-3-3 design for the 10 and 40-item tests, and the MST 1-2-4 design for the 20 and 30-item tests. When the bias values were examined, it was seen that the CAT design had values in the range of [0.121, 0.211], the MST 1-3-3 design was in the range of [0.190, 0.318], and the MST 1-2-4 design was in the range of [0.191, 0.312]. As is seen in [Table 3](#), the lowest bias values belong to CAT whereas the highest bias is in the MST 1-3-3 design for 10 items, and in the MST 1-2-4 design for other test lengths. When the test designs were compared in terms of correlation, the design with the highest correlation across all test lengths was CAT [0.963, 0.989], the lowest correlation was in the MST 1-3-3 design for the 10-item test, and the MST 1-2-4 design for the other test lengths. As the number of items increased for all designs, the correlation values increased and got closer to each other.

3.3. Results of the Condition that the Ratio of DIF Items is 30%

Finally, values were examined for the condition that DIF item ratio was fixed at 30%. As indicated in [Table 3](#), considering the RMSE values, the lowest RMSE value for all test lengths was obtained in the CAT design. The RMSE values for the MST 1-3-3 and 1-2-4 designs appear to be quite close to each other for all test lengths. When the bias values were examined, it was observed that the lowest bias values were calculated in the CAT design at all test lengths and MST designs gave very close results to each other. When the test designs were compared in terms of correlation values, the design with the highest correlation value across all test lengths is the CAT design [0.963-0.989] ([Table 3](#)). As the test length increased, the bias value for all designs decreased and correlation values increased as the number of items increased for all designs.

3.4. ANOVA Analysis

After the interpretation of RMSE, bias and correlation values, separate one-way ANOVA tests were conducted in order to observe whether those values differ significantly between test designs. Three separate one-way ANOVA analyses were performed for each DIF item ratio (10%, 20% and 30%), in which the RMSE, bias, and correlation values were taken as the dependent variable and the test design as the independent variable, and the findings were analyzed separately for each test length. While the assumption of normal distribution was provided in the analyses, the assumption of homogeneity of variances was violated in some

cases. In cases where that assumption was violated, the Welch test was used and in other cases the data in the ANOVA table (Table 3) were interpreted. ANOVA results are given in detail for each DIF item ratio condition as follows:

3.4.1. Ratio of DIF items is 10%

As a result of ANOVA analysis, it was seen that RMSE, bias, and correlation values significantly differed between test designs at each test length ($p < .05$). According to the results of the Post-Hoc comparison, the difference in RMSE, bias, and correlation values reached a significant level among all designs at all test lengths. In short, the lowest values for RMSE and bias were obtained in CAT design at all test lengths and those values differed significantly from the values of the MST designs. The highest RMSE and bias values were observed in the MST 1-2-4 design at all test lengths and differed significantly from others. For the correlation, the highest values were obtained in CAT and the lowest values were obtained in the MST 1-2-4 design along all test lengths. The difference in correlation values between the designs was significant over all test lengths. When all the results were considered together, it was concluded that the CAT design that had the lowest RMSE, bias, and the highest correlation values provided the highest measurement precision. On the other hand, the MST 1-2-4 design, which had the highest RMSE and bias and the lowest correlation values, was the design with the lowest measurement precision.

3.4.2. Ratio of DIF items is 20%

As a result of ANOVA analyses made for the condition that the DIF item ratio was 20%, RMSE, bias, and correlation value differences among designs were found to be significant for all cases. Therefore, it was concluded that CAT was the test that provided the highest measurement accuracy among the three designs. The design with the highest RMSE was MST 1-3-3 for the 10 and 40-item tests, and MST 1-2-4 for the 20 and 30-item tests. The difference between the MST designs reached a significant level in the 10, 20, and 30-item tests. The highest values of the bias were in the MST 1-3-3 design for 10 items, and in the MST 1-2-4 design for the other test lengths. The lowest correlation was obtained from the MST 1-3-3 design in the 10-item test and the MST 1-2-4 design in the other test lengths. When all the results were considered together, it was concluded that the CAT design with the lowest RMSE and bias and the highest correlation values provided the highest measurement precision. Besides, the lowest measurement precision was obtained in the MST 1-3-3 design in the 10-item test and in the MST 1-2-4 design for other test lengths.

3.4.3. Ratio of DIF items is 30%

It was seen that RMSE, bias, and correlation values significantly differed between test designs at each test length ($p < .05$) as in the previous DIF item ratios. When the Post-Hoc comparison results were analyzed in terms of the RMSE variable, there was a significant difference between all designs for the 10, 30, and 40-item tests; however, in the 20-item test, it was seen that the mean difference of .001 between the MST 1-3-3 and MST 1-2-4 designs could not reach a significant level. When the mean differences of the bias values of the designs were examined, it was concluded that all three designs differed significantly from each other for all test lengths. Finally, for the correlation values, Post-Hoc results were examined and it was concluded that there was a significant difference between all designs for all test lengths. Therefore, the highest measurement accuracy was obtained for the CAT design as this measurement precision was maintained over all test lengths and it was significantly higher than the precision of other designs. It can be seen that the values of the MST designs were very close to each other. The lowest measurement precision for the 30-item test was observed in the MST 1-3-3 design, and for the other conditions it was in the MST 1-2-4 design.

Finally, in order to descriptively see the effects of the increase in the rate of items with DIF in the test on RMSE, bias and correlation values, graphs were formed and presented in Figure 1. Looking at the RMSE graph, RMSE values for CAT were quite close to each other at different DIF ratios; however, it was observed that an increase in the DIF ratio increased the RMSE value in the MST designs, especially in the 10-item test where the test length was the lowest. For MST designs, the effect of the increase in DIF ratio on the RMSE decreased as the number of items increased. Similarly, the bias values were close to each other at different DMF ratios for CAT. In the MST designs, the increase in the DIF ratio in the 10-item test affected the bias values considerably, and this effect decreased as the test length increased. Looking at the correlation graph in Figure 1c, similar comments can be made to the comments made for RMSE and bias. It was found that the correlation values decreased as the DIF ratio increased for the CAT designs.

It was determined that the increase in DIF item ratio indicated the most serious effect for the 10-item test. For CAT, the increase in the DIF item ratio did not have a great effect. Those findings showed that CAT was the least affected test design by the increase in the ratio of items with DIF in the test. Two MST designs generally indicated parallel findings which were especially affected by the change in DIF item ratio in the 10-item test, and this effect decreased as the test length increased.

Figure 1. Change of RMSE, bias, and correlation values with the increase of DIF item ratio.

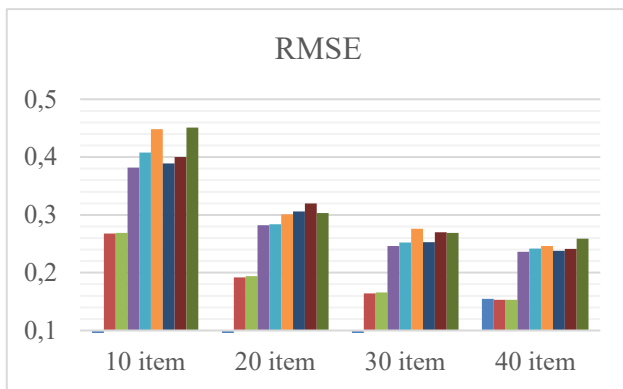


Figure 1a. RMSE values.

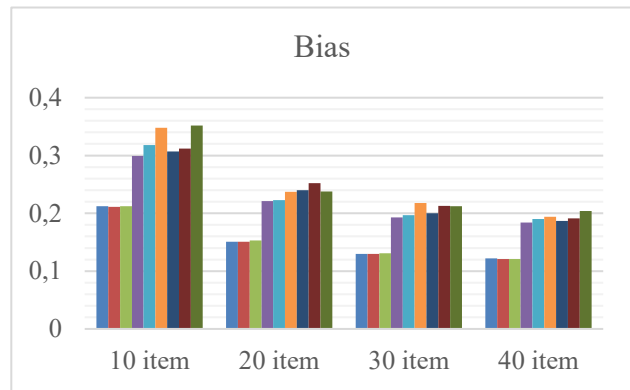


Figure 1b. Bias values.

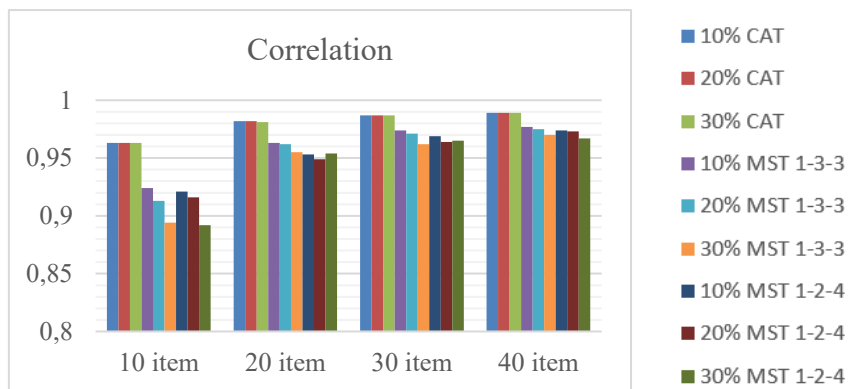


Figure 1c. Correlation values.

4. DISCUSSION and CONCLUSION

Within the scope of this study, it was aimed to examine the effect of the inclusion of items that have differential item functioning (DIF) in the test on the effectiveness of computer adaptive test (CAT) and multi-stage test (MST) under different conditions. For this purpose, data were generated by the simulation method and the performances of different test designs (CAT, MST 1-3-3 and MST 1-2-4) were compared under different test lengths and DIF-item ratios.

In order to evaluate test performances, RMSE, bias, and correlation values were considered together. When the obtained results were analyzed in terms of RMSE and bias, it was seen that the CAT design had the lowest values for all conditions. When the MST 1-3-3 and 1-2-4 designs were compared, a general interpretation couldn't be made. While the RMSE value of the 1-2-4 design was significantly higher than that of the 1-3-3 design throughout all test lengths in the condition that the DIF rate was 10%. When it was 20%, the RMSE of MST 1-3-3 was higher in the 10 and 40-item test length, and it was higher for the 30-item test when the ratio was 30%. Similar to the RMSE, in the condition that the DIF item rate was 10%, while the bias value of the 1-2-4 design was significantly higher than that of the 1-3-3 design throughout all test lengths; the 1-3-3 design gave higher bias values in the 10 and 40-item test when it was 20%, and in the 30-item test when it was 30%. Finally, the findings obtained from the correlation values indicate that CAT had the highest correlation value in all conditions. The lowest correlation values were obtained for 1-3-3 design in 20% DIF-10 items, 30% DIF-30 item conditions, and 1-2-4 design in all other conditions.

In addition, when we looked at how the increase in the DIF ratio affected the performance of the test designs, it was observed that the CAT gave similar results in terms of RMSE, bias and correlation, regardless of the ratio of items with DIF (Figure 1). However, the same was not the case for MST designs. The increase in the DIF ratio in the MST designs generally led to an increase in the RMSE and bias values and a decrease in the correlation values. In particular, the 10-item tests were more affected by the increase in the DIF item ratio than that in the CAT, and this effect decreased as the number of items increased.

When the information given above is interpreted, it can be concluded that CAT provided better measurement accuracy compared to the other two MST designs under all test length and DIF item ratio conditions. In addition, the design that was least affected by the increase in the ratio of items with DIF was CAT. Therefore, it can be interpreted that CAT could reduce the effect of DIF more than other designs. When the two MST designs were compared, it was seen that the 1-3-3 design offered higher measurement accuracy in most conditions. However, those findings were not sufficient to say that the 1-3-3 design outperformed the 1-2-4 design.

The main finding from this study is that the CAT was the design that minimized the effect of DIF throughout all test lengths. The finding that CAT can regulate the effect of DIF is in line with the findings obtained from the study of Piromsombat (2014). Piromsombat examined the effect of DIF items in the test on ability estimation on CAT and revealed that CAT can modulate the effect of DIF if it comes early in the test, especially when the DIF level is moderate. In other cases, CAT reduced the effect of DIF. Besides, that the number of adaptation points in CAT is higher than that in MST can result in higher CAT measurement accuracy (Sarı, 2016; Thai, 2015). For example, while the 1-3-3 panel design has only two adaptation points, regardless of the number of items, there are 19 adaptation points in a 20-item CAT. This finding may be the result of this fact. Since CAT has more adaptation points than MST designs have, CAT may control the DIF effect in a better way. Another finding obtained from this specific study is that the effect of increase in DIF item ratio on CAT performance is lower compared to the effect on MST designs. MST designs were highly affected by the increase in DIF item rate, especially when the number of items was 10, which is also thought to be relevant with the number of adaptation points. As stated before, CAT has more adaptation points than MST has, regulating

the DIF effect better. Therefore, it is expected that CAT offers better measurement accuracy compared to MST designs and is less affected by the DIF item ratio in the presence of an item with DIF in the test. Since no other studies examining the effect of DIF items on adaptive tests have been found in the literature, the discussion on this finding has been limited.

Apart from the DIF effect, studies that CAT and MST designs were compared in the literature were also examined. Kim and Plake (1993) examined measurement precision of CAT and MST in terms of first stage module length (10, 15 and 20 items), total test length (40, 45 and 50 items), number of second stage modules (6, 7, 8 modules), and item difficulty distribution in the first stage module. It has been revealed that CAT gives better results in terms of measurement accuracy than MST does. In the study conducted by Patsula (1999), the accuracy of the ability estimations obtained from different CAT designs, paper-and-pencil tests, and MST designs (number of stages, number of modules in each stage, and number of items in each module) were compared and it was determined that CAT produced the most accurate ability estimation and that the increase in the number of modules in each stage affected the measurement precision and effectiveness. In another study, Sari (2016) investigated the precision of the results obtained from CAT and MST, while the number of content areas varied in tests of different lengths. The main finding of the study was that CAT gave better results than the other two MST designs for all conditions and the two MST designs offered comparable results. In addition, Tay (2015) stated that CAT has more adaptation points than those of MST, therefore they are more effective designs. The common result obtained from the studies in the literature is that CAT gives better results than MST does in different studies and under different conditions. This inference based on those studies shows parallelism with the finding that the CAT performance obtained as a result of the study is higher than the MST performance.

The last finding to state, not related with DIF again, was that when available findings were examined, it was seen that the RMSE and bias values decreased and the correlation values increased as the test length increased for all designs. Therefore, it can be concluded that increasing the test length increases the measurement accuracy. Similar to this finding, Sari (2016) also revealed in his study that increasing the test length resulted in a decrease in the RMSE and bias value and an increase in correlation for both CAT and MST. Another finding obtained as a result of the research was that regarding the comparison of MST designs among themselves, the 1-3-3 design offered high measurement accuracy in a larger number of conditions, but the available findings were not sufficient to say that the 1-3-3 design outperforms the 1-2-4 design. There is no study in the literature comparing those two designs. Findings from different studies are needed to make a discussion about the relevant finding. Based upon the results of our particular study, some recommendations for practitioners are stated as follows. Firstly, it has been seen that CAT gives better results compared to MST for situations where items with DIF are present in the test. In cases with similar conditions to this study, the use of CAT may be recommended. Secondly, MST designs were more affected by items with DIF than those with CAT. Both MST designs used could not regulate the effect of DMF, and the measurement accuracy was more negatively affected compared to that of CAT. If MST is to be used, DIF analyses must be performed. Lastly, especially when the test is 10-items length, the increase in the DIF rate negatively affects the measurement accuracy of the MST. In those cases, the use of MST should not be preferred or should be used very carefully. When RMSE, bias, and correlation values were carefully examined, it can be said that values were getting closer with the increment of test length and they were very close especially after 30 items for all designs. Therefore, a test with at least 30 items can be recommended to use in cases where the presence of DIF is suspected. Those findings were thought to make a significant contribution to the literature since there were no studies found in the literature focusing on comparing CAT and MST approaches in case of the presence of DIF items in the test.

4.1. Further Research

The data set used in the research is limited to simulation data and the item pool used in the study is limited to the item parameters determined by the researcher. It can be recommended to work with real data set in future research. An item pool can also be created with different item parameter distributions and values and the study can be repeated. Besides, only dichotomously scored (1-0) items were taken into account within the scope of the study. Similar studies can be done with polytomously scored items. On the other hand, items were produced to show only uniform DIF when generating items with DIF. Similar studies can be done by adding items indicating non-uniform DIF. The study can be replicated by changing the effect size of the generated DIF items. In addition, since only fixed length was used as test termination rule in this study, the research can be repeated by using different test termination rules. Another limitation of the study is that only two designs were used for the MST. Therefore, the study can be repeated with different MST designs.

Acknowledgments

This research article was produced from the doctoral dissertation of first author under the supervision of second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Başak Erdem Kara: Investigation, Methodology, Visualization, Software, Formal Analysis, and Writing-original draft. **Nuri Dogan:** Investigation, Methodology, Supervision, and Validation.

Orcid

Basak ERDEM KARA  <https://orcid.org/0000-0003-3066-2892>

Nuri DOGAN  <https://orcid.org/0000-0001-6274-2016>

REFERENCES

- Aksu-Dunya, B. (2017). *Item parameter drift in computer adaptive testing due to lack of content knowledge within sub-populations* (Publication No. 10708515) [Doctoral Dissertation, University of Illinois]. ProQuest Dissertations & Theses.
- Armstrong, R.D., Jones, D.H., Koppel, N.B., & Pashley, P.J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement, 28*(3), 147–164. <https://doi.org/10.1177/0146621604263652>
- Babcock, B., & Albano, A.D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement, 36*(7), 565-580. <https://dx.doi.org/10.1177/0146621612455090>
- Berger, S., Verschoor, A.J., Eggen, T.J.H.M., & Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontiers in Education, 4*(1), 1–18. <https://doi.org/10.3389/feduc.2019.00001>
- Birdsall, M. (2011). *Implementing computer adaptive testing to improve achievement opportunities*. Office of Qualifications and Examinations Regulation Report. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/606023/0411_MichaelBirdsall_implementing-computer-testing-Final_April_2011_With_Copyright.pdf

- Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items* (4th ed.). Sage Publications, Inc.
- Chu, M.W., & Lai, H. (2013). Detecting biased items using CATSIB to increase fairness in computer adaptive tests. *Alberta Journal of Educational Research*, 59(4), 630–643. <https://doi.org/10.11575/ajer.v59i4.55750>
- Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth Group/Thomson Learning.
- Gierl, M.J., Lai, H., & Li, J. (2013). Identifying differential item functioning in multi-stage computer adaptive testing. *Educational Research and Evaluation*, 19(2-3), 188–203. <https://www.tandfonline.com/doi/full/10.1080/13803611.2013.767622>
- Hambleton, R.K., & Swaminathan, H. (1991). *Item response theory: Principles and applications*. Springer.
- Hambleton, R.K., Jac, N.Z., & Pieters, J.P.M. (2000). Computerized adaptive testing: Theory, applications and standards. In R.K. Hambleton, & J.N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (4th ed., pp. 341–366). Springer.
- Han, K.T., & Guo, F. (2011). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (Report No. RR-11-02). Graduate Management Admission Council (GMAC) Research Reports. https://www.gmac.com/~media/Files/gmac/Research/research-report-series/rr1102_itemcalibration.pdf
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, Summer 2007, 44-52. <https://doi.org/10.1111/j.1745-3992.2007.00093.x>
- Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Publication No. 3315089) [Doctoral Dissertation, University of Texas]. ProQuest Dissertations & Theses.
- Lei, P.W., Chen, S.Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement*, 43(3), 245-264. <http://dx.doi.org/10.1111/j.1745-3984.2006.00015.x>
- Luecht, R.M., & Sireci, S.G. (2011). *A review of models for computer-based testing* (Report No. 2011-12). College Board Research Report. <https://files.eric.ed.gov/fulltext/ED562580.pdf>
- Magis, D., Yan, D., & von-Davies, A. (Eds.). (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer.
- National Research Council (1999). *Designing mathematics or science curriculum programs: A guide for using mathematics and science education standards*. National Academies Press. <https://www.nap.edu/catalog/9658.html>
- Piromsombat, C. (2014). *Differential item functioning in computerized adaptive testing: Can cat self-adjust enough?* (Publication No. 3620715) [Doctoral Dissertation, University of Minnesota]. ProQuest Dissertations & Theses.
- Sari, H.I. (2016). *Examining content control in adaptive tests: Computerized adaptive testing vs. Computerized multistage testing* (Publication No. 403003) [Doctoral Dissertation, University of Florida]. The Council of Higher Education National Thesis Center.
- Sari, H.I., & Huggins-Manley, A.C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory and Practice*, 17, 1759-1781. <http://doi:10.12738/estp.2017.5.0484>
- Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2. ed., p. 185–229). Routledge.

- Tay, P.H. (2015). *On-the-fly assembled multistage adaptive testing* (Publication No. 3740572). [Doctoral Dissertation, University of Illinois]. ProQuest Dissertations & Theses.
- van der Linden, W.J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, 33(1), 5-20. <https://doi.org/10.3102/1076998607302626>
- van der Linden, W.J., & Pashley, P.J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing*. Springer.
- Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., p. 1–22). Lawrence Erlbaum Associates.
- Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* (Publication No. 10273809). [Doctoral Dissertation, Michigan State University]. ProQuest Dissertations & Theses.
- Wang, S., Haiyan, L., Chang, H.H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, 53(1), 45–62. <https://doi.org/10.1111/jedm.12100>
- Wang, X. (2013). *An investigation on computer-adaptive multistage testing panels for multidimensional assessment* (Publication No. 3609605). [Doctoral Dissertation, University of North Carolina]. ProQuest Dissertations & Theses.
- Weiss, D.J., & Kingsbury, G.G. (1984). Application of computer adaptive testing to educational problems. *Journal of Educational Measurement*, 21 (4), 361-375. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>
- Yan, D. (2010). *Investigation of optimal design and scoring for adaptive multi-stage testing: A tree-based regression approach* (Publication No. 3452799). [Master Thesis, Fordham University]. ProQuest Dissertations & Theses.
- Yan, D., von-Davies, A.A., & Lewis, C. (2014). Overview of computerized multistage tests. In D. Yan, A.A. von-Davies, & C. Lewis (Eds.), *Computerized multistage testing* (p. 3–20). CRC Press; Taylor & Francis Group.
- Zheng, Y., & Chang, H.H. (2014). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, 39 (2), 104-118. <https://doi.org/10.1177/0146621614544519>
- Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Headquarters of National Defense.
- Zwick, R. (2010). The investigation of differential item functioning in adaptive tests. In W.J. van der Linden and C.A.W. Glas (Eds.), *Elements of adaptive testing*. Springer
- Zwick, R., & Bridgeman, B. (2014). Evaluating validity, fairness, and differential item functioning in multistage testing. In D. Yan, A.A. von-Davies, & C. Lewis (Eds.), *Computerized multistage testing*. CRC Press; Taylor&Francis Group.

A Comparison of type I error and power rates in procedures used determining test dimensionality

Gul Guler^{1,*}, Rahime Nukhet Cikrikci¹

¹Istanbul Aydın University, Faculty of Education, Department of Elementary Education, Türkiye

ARTICLE HISTORY

Received: Jan. 18, 2022

Revised: June 30, 2022

Accepted: Aug. 14, 2022

Keywords:

Dimensionality,
Construct validity,
DIMTEST T statistic,
DETECT,
Nonlinear factor analysis.

Abstract: The purpose of this study was to investigate the Type I Error findings and power rates of the methods used to determine dimensionality in unidimensional and bidimensional psychological constructs for various conditions (characteristic of the distribution, sample size, length of the test, and interdimensional correlation) and to examine the joint effect of the conditions (effect of the interaction of conditions) as well as the main effect of each condition. The simulative data were generated for the study using the SAS program. Within the scope of the study, the data were analyzed using the DIMTEST T statistic and the Dimensionality DETECT IDN index, which is one of the non-parametric methods. The Nonlinear Factor Analysis (NOHARM) method was preferred from among parametric methods. As a result of the study, it was noted that the most consistent results in making the unidimensionality decisions belong to the Nonlinear Factor Analysis method showing standard normal distribution according to the shape of the distribution. When the power study results were examined, it was noted that the DIMTEST T statistic gave more accurate results in conditions with large samples, consisting of data with standard normal distribution. On the other hand, while results of the DETECT IDN index and Nonlinear factor analysis were more internally consistent, it was noted that in conditions where the sample size was 1000 and above, the DIMTEST T statistic also made the right decisions in determining dimensionality.

1. INTRODUCTION

In the process of test and scale development in education and psychology, dimensionality is frequently used in validity studies. Dimensionality is the relationship between the items in a test and the implicit feature that the test is thought to measure (Svetina, 2011). Dimensionality is related to the number of skills or psychological constructs that a test or item set measures. The dimensionality determination process is an important issue to consider, regardless of whether the measurement model is unidimensional or multidimensional (Embretson & Reise, 2000). A test has a theoretical structure and is prepared for a specific purpose. The underlying structure of the test should be examined and verified. In this context, construct validity studies are important in terms of the technical features of instruments in education and psychology and

*CONTACT: Gul Guler ✉ gulyuce2010@gmail.com 📍 Istanbul Aydın University, Faculty of Education, Department of Elementary Education, Türkiye

are one of the necessary steps in assessing the dimensionality of tests and scales. A feature to be measured may be associated with more than one implicit feature by nature. When we look at the tests used in education and psychology, it is seen that most of them measure more than one latent feature. For example, while a science test was developed to measure science process skills, it could also measure reading comprehension. For this reason, it is useful to know whether the structure to be measured is one-dimensional or multidimensional. Considering the purpose of creating and applying the test, this situation will affect the validity of decisions made about individuals based on test scores. Determining the dimensionality of the items in a test is extremely important as it will also shape the statistical analysis of the data (Svetina, 2011; Zhang, 2008).

In case a measurement procedure is treated as unidimensional while being in fact multidimensional, the interpretation of test scores, and thus the validity of measurement processes would be misleading (Göçer Şahin, 2016; Touron et al., 2012). Determination of dimensionality, in addition to the determination of the extent to which unidimensionality is neglected and revealing the power of tests with Type I or significant in terms of the validity of decisions made as a result of the tests applied. When a test is unidimensional, that is, when the H_0 hypothesis is true, accepting the H_1 hypothesis with a statistical decision, meaning that the test is multidimensional, causes a Type I Error. Accepting the H_0 hypothesis while a test is multidimensional, in other words, saying it is unidimensional causes a Type II Error. In addition, deciding that a test is statistically multidimensional while it is actually multidimensional displays the power of the test. Thus, it is considered that testing of unidimensionality is required since the determination of all these situations is directly related to the validity of the decisions.

When studies in the literature are assessed, dimensionality determination methods are generally separated as parametric and non-parametric methods (Abswoude et al., 2004; Mroch & Bolt, 2006; Özbek, 2012; Reinchenberg, 2013; Svetina, 2011; Svetina & Levy, 2014). Conditions such as small samples, low numbers of items, and a high degree of interdimensional correlation revealed the need to study and use non-parametric methods and comparison conditions in addition to parametric methods. The purpose of this study is to investigate the Type I Error and power rates of the methods used to determine dimensionality in unidimensional and two-dimensional psychological constructs depending on sample size, characteristics of the distribution, test length, and interdimensional correlation conditions while comparing the main effect of each condition in addition to joint effects of conditions (effect of the interaction of conditions). In line with this general purpose, answers were sought to the following questions:

1. How do *Type I Error rates* obtained from unidimensional data change where the length of the test, characteristics of distribution, and sample size are manipulated, according to various dimensionality determination methods, in tests scored dichotomously?
2. How do *power rates of the test*, obtained from bidimensional data change where test length, interdimensional correlation degree, distributions and sample size are manipulated according to various dimensionality determination methods, in tests scored dichotomously?
3. What are the Type I Error rates and the power rates of the test using standard, normal and skewed data according to various dimensionality determination methods in tests scored dichotomously?

The most significant reason for choosing the DIMTEST T statistic in this study was the fact that it was a testing method that worked well in large samples and large item pools, and it was effective in displaying even small secondary features (Svetina, 2011). The reason for preferring Nonlinear Factor Analysis was that its results could be interpreted easily, it worked well in small samples, and it was based on factor analytical approaches. In addition, the fact that all methods were accessible for free supported the preference (Svetina & Levy, 2014; Touron et

al., 2012). While factor analysis is generally preferred in unidimensional studies, many studies stated that examining unidimensionality with factor analysis alone is not sufficient and recommended other methods (Finch & Monahan, 2008; Hattie et al., 1996; Ledesma & Valero-Mora, 2007; Özbek, 2012; Reichenberg 2013; Svetina, 2011; Svetina & Levy, 2014; Touron et al., 2012; Yen, 2007). Despite this argument, in many national or international studies factor analysis is used and considered sufficient in the examination of unidimensionality. However, factor analysis requires the assumption of a multivariate normal distribution which might not be achieved in social sciences frequently.

Applying factor analysis to prove unidimensionality – due to the nature of test and scale development – or not using any methods and calculating test scores over the test totals to arrive at decisions about individuals taking achievement tests at national or international test centers are limiting factors in terms of the validity of the decisions.

The fact that achievement tests used by national or international test centers that use factor analysis only or do not use any methods to accept unidimensionality and calculate test scores over the total test to arrive at decisions about individuals – due to the nature of test and scale development – is a limiting factor in terms of the validity of the decisions. If there is a violation of unidimensionality, the multidimensional structure must be determined with correct methods and indices, and it should be investigated for construct validity studies. Another important point in the process of determining unidimensionality is the requirement for test developers to investigate the effect of sample size on determining unidimensionality considering the difficulties experienced in data collection processes in our country.

2. METHOD

2.1. Data Production Study

In this study, simulation data were used to respond to the research questions. Simulation models should be based on realistic parameters (Davey et al., 1997; as cited in Göçer Şahin, 2016). In addition, simulation studies are meaningful when they are similar to real situations. Since it is difficult to meet all the conditions stated in this study in real data at the same time, it was decided to use simulation data. The data of this study were produced using the SAS software. The data were generated in a 2-parameter logistic and compensatory model for power analysis, in accordance with a dichotomous bidimensional structure. For the Type I Error study, unidimensional dichotomous data was generated in the 2-parameter logistic model. Variables, number of conditions, and condition values are presented in [Table 1](#):

Table 1. *Variables and their conditions used in data production.*

Study	Variables	Number of Conditions	Condition Values
Type I Error study	Properties of Distribution	2	Normal, Skewed
	Sample Size	6	200, 300, 500, 1000, 2000, and 3000
	Test Length	3	10, 20, 30
Power Analysis	Properties of Distribution	2	Normal, Skewed
	Sample Size	6	200, 300, 500, 1000, 2000, and 3000
	Test Length	3	10, 20, 30
	Interdimensional correlation	4	0.25, 0.50, 0.75, 0.90
Number of Replications		100	

When Table 1 is examined, considering the manipulated variables for Type I Error study, $2*6*3=36$ conditions and for power analysis $2*6*3*4=144$ conditions were generated, and 100 replications were performed for each condition. Before the data was produced, discrimination parameters of the items were defined considering the research design. The multidimensionality of the test was determined according to the discrimination parameters. Accordingly, an item that loads on both dimensions must have two discrimination coefficients. If the item predominantly loads on both dimensions, it is defined as complex; while if it loads dominantly on one dimension and loads little on the other, it is defined as approximately simple, and if it loads dominantly on one dimension and none on the other, it is defined as a simple item. For example, in this study, the first five items of a 10-item test predominantly belong to the first dimension and a small amount to the second dimension while the other five items are arranged in a way that loads predominantly on the second dimension and to a small extent on the first dimension. Thus, a multidimensional test was developed, which predominantly loaded on two different dimensions. While producing the item parameters, ITEM-GENv2 software developed by Ackerman (1994) was used. In this software, parameters are generated by entering only the file name, test length, item angles, the range of the intersection parameter, and the range of the MDISC parameter. Accordingly, items that load on the first dimension make angles with the x-axis that vary between 5° and 20° while items that load on the second dimension make angles that vary between 70° and 88° (Ackerman et al., 2003).

MDISC is the discrimination parameter of multidimensional Item Response Theory (IRT) and corresponds to the item discrimination in unidimensional IRT. Since there is more than one dimension at this point, there is a distinctiveness for each dimension. Item discrimination (MDISC) is represented by a vector $(\alpha_1, \alpha_2, \alpha_3 \dots \alpha_k)$. The vector length is expressed as:

$$MDISC = \sqrt{\sum_{n=1}^k \alpha^2_{ik}} \quad (1)$$

The vector length terms as the common item discrimination (Göçer Şahin, 2016). It could be argued that as the length increases, the discrimination of the item also increases. The α_{ik} in the formula above represents the distinctiveness values of each dimension. The MDISC value here can also be interpreted as distinctiveness in unidimensional IRT.

In addition to the vector length, it is useful to know the vector direction and its distance from the origin. The vector direction is expressed with:

$$\alpha_i = \arccos\left(\frac{\alpha_{i1}}{MDISC}\right) \quad (2)$$

The α_i is the angle that the item vector makes with the θ_1 axis. Thus, an angle of 45° means that the item measures both abilities well. If the angle is greater than 45° , it means that the second dimension is measured better than the first dimension. However, if it is less than 45° , it means that this item primarily measures θ_1 ability, meaning, the first dimension is measured better than the second dimension (Göçer Şahin, 2016; Sünbül, 2011).

In unidimensional IRT, the D parameter is the b parameter's equivalent in Multidimensional Item Response Theory (MIRT) and that expresses the distance of the item vector from the starting point and gives information about the item difficulty (Reckase, 2009). This parameter is calculated as:

$$D = \frac{-d_i}{MDISC} \quad (3)$$

The d_i in the formula is described as an intercept term. A negative sign of the item is interpreted as being easy while a positive sign is interpreted as being difficult.

In this study, the range of the MDISC parameter for multidimensional items was entered as 0.8 and 1.8. The study of Ackerman (1994) was taken into account in determining this range. In the condition that the number of simple items is 10 and the structure is bidimensional, the structure of the item, parameters, and the angles of the items with the axes are presented in Table 2 as an example.

Table 2. Item parameters in data generation.

Dimensions	Items	a_{j1}	a_{j2}	b	MDISC	D	Angle
1	1	1.265	.111	-.579	1.27	.46	5.00
	2	1.074	.126	.422	1.08	-.39	6.67
	3	1.671	.245	-.109	1.69	.06	8.33
	4	1.312	.231	-.533	1.33	.40	10.00
	5	.980	.202	-.233	1.00	.23	11.67
	6	.937	.222	-.123	.96	.13	13.33
	7	.903	.242	-.726	.93	.78	15.00
	8	1.164	.349	.415	1.22	-.34	16.67
	9	1.076	.356	.074	1.13	-.07	18.33
	10	.765	.278	-.147	.81	.18	20.00
2	11	.434	1.194	-.579	1.27	.46	70.00
	12	.334	1.029	.422	1.08	-.39	72.00
	13	.465	1.623	-.109	1.69	.06	74.00
	14	.322	1.293	-.533	1.33	.40	76.00
	15	.208	.979	-.233	1.00	.23	78.00
	16	.167	.948	-.123	.96	.13	80.00
	17	.130	.926	-.726	.93	.78	82.00
	18	.127	1.209	.415	1.22	-.34	84.00
	19	.079	1.130	.074	1.13	-.07	86.00
	20	.028	.813	-.147	.81	.18	88.00

2.2. Data Analysis

Both parametric and non-parametric methods were used to compare the performances of various methods in the assessment of unidimensionality. In the scope of this study, the DIMTEST T statistic and Dimensionality DETECT IDN index were used among non-parametric methods. Among parametric methods, Nonlinear Factor Analysis (NOHARM) method was used. The data were analyzed in the following steps:

In the first stage, unidimensional and multidimensional data were generated respectively for testing Type I Errors and power rates. In addition to Stout et al. (1996), Forelich and Habing (2008) studied AT and PT partitioning for the DIMTEST T statistic and (a) it was noted that AT items should be homogeneous in terms of dimensionality, meaning, in terms of geometric representation the angle at which the AT items are located should be relatively narrow. (b) Θ_{AT} and Θ_{PT} should be as different as possible, in other words, in terms of geometric representation the angles between Θ_{AT} and Θ_{PT} should be as large as possible. (c) There must be at least four items in AT while the PT must have at least half of the items in the test. In this study, for the DIMTEST T statistic, AT and PT items were fixed for all conditions, with half of the items in

the AT subtest and the other half in the PT subtest. In the cases where the DIMTEST T statistic was greater than the critical value of 1.96, the H_0 hypothesis was rejected.

Dimensionality DETECT IDN index and Nonlinear Factor Analysis methods were used in their default options. Dimensionality DETECT IDN index value of 1 or higher indicates high multidimensionality, while a value between 0.4 and 1 indicates moderate multidimensionality, and a value between 0.2 and 0.4 indicates unidimensionality. In a simulation study by Kim (1994) it was noted that if the Dimensionality DETECT IDN index was less than 0.10, the data could be considered unidimensional. In the same study, it was noted that a value between 0.10 and 0.50 would indicate multidimensionality which was a low probability, a value between 0.51 and 1 would indicate moderate multidimensionality, and a value over 1 would indicate strong multidimensionality (Ackerman & Walker, 2003). 100 replications were performed for all analyses. For each condition of the DIMTEST T statistic and the Dimensionality DETECT IDN index, 4 different result tables were obtained including the reliability coefficients, theta values, the DIMTEST T statistic and the Dimensionality DETECT IDN index. T statistic and p-significance values were reported for the DIMTEST T statistic.

Among parametric methods, nonlinear factor analysis (NOHARM) was applied, and reliability coefficients, theta values and NOHARM result tables were obtained. Two indices, Tanaka Goodness of Fit Index (TIGF) and RMSR, were used to interpret the outputs of the NOHARM program. A TIGF value of ≥ 0.95 and an RMSR value of ≤ 0.05 were evidence of a good fit of the model (Hooper et al., 2008; Hu & Bentler, 1999). In the final step, unidimensionality rejection rates for all outcomes were reported for each condition.

3. FINDINGS

In this section the rates of rejection of unidimensionality as a result of the effect of all conditions and the joint effect of the interaction of the conditions are presented. According to the results of DIMTEST T statistics in Table 3, it was considered that the test length was more inconsistent in making the decision of unidimensionality when the test length was 10 items than when the test length was 20 and 30 items. In addition, in the cases where the test length was 20 and 30, it was considered that it gave more consistent results regardless of the sample size. According to the results of DIMTEST T statistics, regardless of the sample size, as the length of the test increased in unidimensional data, the rate of rejection of unidimensionality generally decreased, in other words, the rate of Type I Error decreased. Another remarkable point in the results of DIMTEST T statistics was that as the sample size increased, the test length produced accurate results for unidimensional data with standard normal distribution, especially in the cases where the test length was 20 and 30 items. It gave more accurate results, especially with a sample size of 300 and above. It could be argued that this finding supports the studies of Finch and Habing (2007) and Finch and Monahan (2008).

When the DETECT IDN index results were examined, the Type I Error rate generally increased as the sample size decreased for the data showing standard normal distribution. Especially when the sample size was 200, 300 and 500, it was noted that the rate of Type I Error was high. However, it could be argued that it gave more inconsistent results when the length of the test was 10 items. In the study conducted by Roussos and Özbek (2006), it was stated that the DETECT IDN index exhibited statistical bias, especially when the test length was 10 or less and the data was unidimensional. Accordingly, the researchers recommended against using DETECT for test lengths of less than 20 items. Although this study coincided with the study of Roussos and Özbek (2006), an important finding was that the sample size should be increased in order to use the DETECT method.

Table 3. DIMTEST T Statistic, Dimensionality DETECT IDN Index, and Type I Error Rates for Nonlinear Factor Analysis in the data showing normal distribution according to various sample sizes and different numbers of items.

		DIMTEST T Statistic	DETECT IDN INDEC	RMSR	TIGF
Sample Size	Number of Items	Rejection rate	Rejection rate	Rejection rate	Rejection rate
200	10	0.04	0.71	0.00	0.00
	20	0.00	0.62	0.00	0.00
	30	0.03	0.53	0.00	0.00
300	10	0.06	0.64	0.00	0.00
	20	0.00	0.57	0.00	0.00
	30	0.00	0.38	0.00	0.00
500	10	0.02	0.59	0.00	0.00
	20	0.00	0.50	0.00	0.00
	30	0.00	0.38	0.00	0.00
1000	10	0.05	0.47	0.00	0.00
	20	0.00	0.43	0.00	0.00
	30	0.00	0.31	0.00	0.00
2000	10	0.18	0.39	0.00	0.00
	20	0.00	0.30	0.00	0.00
	30	0.01	0.20	0.00	0.00
3000	10	0.10	0.38	0.00	0.00
	20	0.00	0.28	0.00	0.00
	30	0.01	0.19	0.00	0.00

Note. N (0,1): Standard Normal Distribution, number of replications: 100, software used for Dimensionality T Statistic: DIMTEST, software used for DETECT IDN index: DETECT, software used for Nonlinear Factor Analysis and Achieved Indexes: NOHARM- RMSR and TIGF

When the RMSR and Tanaka Goodness of Fit Indices were obtained as a result of nonlinear factor analysis that is one of the parametric dimensionality determination methods examined, the Tanaka Goodness of Fit Index (TIGF) value was ≥ 0.95 for unidimensional data with standard normal distribution, regardless of the sample size and the length of the test. However, the RMSR value of ≤ 0.05 in all results proved that the fitness of the model was well. This finding seems to overlap with the study findings of Seo and Sünbül (2012). However, the study by Gessaroli and De Champlain (1996) also showed consistency with conditions where the test length was 15, 30, and 45 items. The DIMTEST T statistic, DETECT IDN index, and Type I Error rates for nonlinear factor analysis in the condition that the test scores were skewed, the sample size was 200, 300, 500, 1000, 2000, and 3000 and the test length was 10, 20 and 30 items are summarized in Table 4.

Table 4. DIMTEST T Statistic, Dimensionality DETECT IDN Index, and Type I Error Rate for Nonlinear Factor Analysis in skewed data for various sample sizes and number of items.

Sample Size	Number of Items	DIMTEST T STATISTIC	DETECT IDN INDEX	NOHARM RMSR	NOHARM TIGF
		Rejection rate	Rejection rate	Rejection rate	Rejection rate
200	10	0.04	0.40	0.00	0.00
	20	0.00	0.30	0.00	0.00
	30	0.00	0.31	0.00	0.00
300	10	0.05	0.39	0.00	0.00
	20	0.03	0.21	0.00	0.00
	30	0.00	0.29	0.00	0.00
500	10	0.04	0.29	0.00	0.00
	20	0.02	0.16	0.00	0.00
	30	0.01	0.18	0.00	0.00
1000	10	0.09	0.27	0.00	0.00
	20	0.00	0.03	0.00	0.00
	30	0.02	0.08	0.00	0.00
2000	10	0.16	0.17	0.00	0.00
	20	0.01	0.01	0.00	0.00
	30	0.00	0.00	0.00	0.00
3000	10	0.18	0.06	0.00	0.00
	20	0.01	0.00	0.00	0.00
	30	0.01	0.00	0.00	0.00

Note. (1.75, 3.75) Skewed Distribution, number of replications:100, software used for Dimensionality T Statistic: DIMTEST, software used for DETECT IDN index: DETECT, Software used for Nonlinear Factor Analysis and Indexes: NOHARM-RMSR and TIGF

According to Table 4, the Type I Error rate was particularly higher in small samples and in the cases when test length was short, and the distribution was skewed. Although it was noted that the DIMTEST T statistic gave more accurate results than DETECT IDN index, it was found that the error rate was higher in the DIMTEST T statistic results when the test length was 10 items compared to other test lengths. However, in all conditions where the test length was 20 and 30 items, it was noted that the DIMTEST T statistic gave very accurate results. When the nonlinear factor analysis (NOHARM) results were examined, it showed a rejection rate of 0.00 for unidimensional data with skewed distribution, regardless of the sample size and the test length. The findings of the third group of the study were in conditions where the data had standard normal distribution, the sample sizes were 200, 300, 500, 1000, 2000 and 3000, and the test length was 10, 20, and 30 items and there was an interdimensional correlation with 0.25, 0.50, 0.75, and 0.90. The power rates of the test for DIMTEST T statistic, the Dimensionality DETECT IDN index, and the Nonlinear Factor Analysis (NOHARM) results are summarized in Table 5:

Table 5. Power rates for DIMTEST T Statistic, Dimensionality DETECT IDN index and Nonlinear Factor Analysis in data with standard normal distribution according to various sample sizes, different numbers of items, and different interdimensional correlations.

N~(0,1)		Interdimensional Correlation															
		0.25				0.50				0.75				0.90			
		Rejection Rate				Rejection Rate				Rejection Rate				Rejection Rate			
Sample Size	Number of Items	DIMTEST T STATISTIC	Dimensionality DETECT IDN INDEX	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT IDN INDEX	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT IDN INDEX	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT IDN INDEX	RMSR	TIGF
200	10	0.91	0.86	1.0 0	1.00	0.64	0.92	1.00	1.00	0.37	0.92	1.00	1.00	0.14	0.96	1.00	1.00
	20	1.00	1.00	1.0 0	1.00	0.87	1.00	1.00	1.00	0.29	1.00	1.00	1.00	0.08	1.00	1.00	1.00
	30	1.00	1.00	1.0 0	1.00	0.95	1.00	1.00	1.00	0.52	1.00	1.00	1.00	0.13	1.00	1.00	1.00
300	10	0.99	0.74	1.0 0	1.00	0.95	0.85	1.00	1.00	0.55	0.91	1.00	1.00	0.26	0.97	1.00	1.00
	20	0.99	1.00	1.0 0	1.00	0.95	1.00	1.00	1.00	0.66	1.00	1.00	1.00	0.23	1.00	1.00	1.00
	30	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	0.86	1.00	1.00	1.00	0.27	1.00	1.00	1.00
500	10	1.00	1.00	1.0 0	1.00	0.97	1.00	1.00	1.00	0.69	0.99	1.00	1.00	0.30	1.00	1.00	1.00
	20	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	0.73	1.00	1.00	1.00
	30	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.47	1.00	1.00	1.00
1000	10	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	1.00	1.00	1.00
	20	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.56	1.00	1.00	1.00
	30	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.83	1.00	1.00	1.00
2000	10	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00
	20	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00
	30	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
3000	10	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	20	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	98	1.00	1.00	1.00
	30	1.00	1.00	1.0 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. N (0,1): Standard Normal Distribution, number of replications:100, software used for Dimensionality T Statistic: DIMTEST, software used for DETECT IDN index: DETECT, Software used for Nonlinear Factor Analysis and Achieved Indexes: NOHARM- RMSR and TIGF

According to the results of the DIMTEST T statistics, in the data showing standard normal distribution, in the case of an interdimensional correlation of 0.25 and with a sample size of 500 and above, no matter what the length of the test was, the unidimensionality in bidimensional data showed standard normal distribution for all conditions while the rejection rate was found to be 1.00. The rejection rate for unidimensionality was found to be 1.00 in all conditions, except for the condition where the interdimensional correlation was 0.50, the sample size was 500 and the test length was 10 items. In addition, for the two conditions (200 and 300) where

the sample size was less than 500, the rejection rate of unidimensionality was lower than in the cases with larger sample sizes. As a result, it could be argued that the DIMTEST T statistic gave more accurate results in conditions with large samples. This finding is consistent with the studies of Finch and Habing (2007), Finch and Monahan (2008), and Özbek Baştuğ (2012). Especially in the cases where the sample size was less than 300, the error rate of DIMTEST T statistics increased significantly. According to the results of DIMTEST T statistics, as the interdimensional correlation value increased, the unidimensionality rejection rate in bidimensional data decreased. In other words, the power of the test decreased. In the cases where the interdimensional correlation was low, the rejection rate of unidimensionality was 1.00 for the DIMTEST T statistic regardless of the sample size and the test length. In other words, the data was accepted to be bidimensional and the power of the test was high. It was noted that the DIMTEST T statistic was significantly affected by the interdimensional correlation for the multidimensionality decision. However, in the cases when the sample size was 3000 and the test length was 30, regardless of the correlation value between dimensions, a rejection rate of 1.00 was achieved for unidimensionality. In other words, an excellent decision was made for multidimensionality. In the study conducted by Zhang (2008), it was stated that in the condition of low interdimensional correlation, short tests produced better results than long tests. However, in this study, when the interdimensional correlation was very low, the results of DIMTEST T statistics gave an excellent performance in terms of test power as the test length increased. Although the result of this study was inconsistent with the study of Zhang (2008), it seemed to overlap with the studies by Alexandra et al. (2004), Seo and Sünbül (2012) and Özbek Baştuğ (2012).

When the results of the dimensionality DETECT IDN index for the power of the test were examined, in the case of bidimensional data with standard normal distribution, with a sample size of 500 and above, the correlation value between dimensions and the test length displayed a rejection rate of 1.00 for all conditions except one. It could be argued that the Dimensionality DETECT IDN statistic worked well in rejecting unidimensionality and accepting bidimensionality in cases with bidimensional data where the sample size was 500 and above. This finding was consistent with the findings of the study by Svetina (2011) and the studies of Roussos and Özbek (2006). In the data with standard normal distribution, when the RMSR and Tanaka Goodness of Fit Index values were examined following nonlinear factor analysis (NOHARM) as a parametric method for test power, it was observed that for bidimensional data with standard normal distribution, interdimensional correlation displayed a rejection rate of 1.00 for unidimensionality, regardless of sample size and test length. In other words, the null hypothesis that the test was unidimensional in all circumstances was correctly rejected. When the relevant literature was reviewed, it was stated in the study conducted by Kaya and Kelecioğlu (2016) that the results of nonlinear factor analysis were more consistent in determining multidimensionality in samples of 50 or more. Contrary to this study, studies by Özbek Baştuğ (2012) and Seo and Sünbül (2012) found that nonlinear factor analysis (NOHARM) was not a powerful statistical method for determining multidimensionality. However, Svetina (2011) stated that statistics based on nonlinear factor analysis (NOHARM) results in determining dimensionality in data suitable for non-compensatory multidimensional IRT models showed a stronger performance compared to Dimensionality DETECT IDN index.

As a result, it was noted that the dimensionality DETECT IDN index and nonlinear factor analysis (NOHARM) results gave more accurate decisions than the DIMTEST T statistic under all conditions in the data with standard normal distribution. It could be argued that the DIMTEST T statistic gave more accurate decisions in conditions where the interdimensional correlation was low, and the sample size was large. In addition, it could be argued that the DIMTEST T statistic worked better in samples of 2000 and above in the cases where the interdimensional correlation was high.

The findings for the 4th group of the study are presented in Table 6. Accordingly, the DIMTEST T Statistic, Dimensionality DETECT IDN index and Nonlinear Factor Analysis (NOHARM) results were compared in terms of test power ratios in the data with skewed distribution, where the sample size was 200, 300, 500, 1000, 2000, and 3000, the test length was 10, 20, and 30 items, and the degree of interdimensional correlation was 0.25, 0.50, 0.75, and 0.90.

Table 6. Power rates for DIMTEST T Statistic, Dimensionality DETECT IDN index and Nonlinear Factor Analysis in data with skewed distribution according to various sample sizes, different numbers of items, and different interdimensional correlation values.

		Interdimensional Correlation															
		0.25				0.50				0.75				0.90			
Sample Size	Number of Items	DIMTEST T STATISTIC	Dimensionality DETECT IDN INDEX	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT IDN INDEX	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT IDN INDEX	RMSR	TIGF	DIMTEST T STATISTIC	Dimensionality DETECT IDN INDEX	RMSR	TIGF
		200	10	0.84	0.76	1.00	1.00	0.69	0.79	1.00	1.00	0.24	0.86	1.00	1.00	0.14	0.85
	20	0.96	1.00	1.00	1.00	0.89	1.00	1.00	1.00	0.36	0.99	1.00	1.00	0.08	0.99	1.00	1.00
	30	0.90	1.00	1.00	1.00	0.82	1.00	1.00	1.00	0.61	1.00	1.00	1.00	0.13	1.00	1.00	1.00
300	10	0.98	0.67	1.00	1.00	0.91	0.71	1.00	1.00	0.61	0.67	1.00	1.00	0.26	0.77	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.23	0.99	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00	1.00	1.00	0.27	1.00	1.00	1.00
500	10	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.77	1.00	1.00	1.00	0.73	0.99	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.92	1.00	1.00	1.00	0.68	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.40	1.00	1.00	1.00
1000	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	0.68	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.56	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	0.82	1.00	1.00	1.00
2000	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00
3000	10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00

According to Table 6, in the data with skewed distribution when the sample size increased and the number of items in the test increased and the power ratios for DIMTEST T Statistic, Dimensionality DETECT IDN index and Nonlinear Factor Analysis were analyzed according to different interdimensional correlation values, it was noted that all conditions in which nonlinear factor analysis and Dimensionality DETECT IDN index were used gave more accurate decisions than the DIMTEST T statistics. However, according to the DIMTEST T statistic, it could be argued that more accurate decisions were made in conditions when interdimensional correlation was low. In addition, in the cases when the interdimensional correlation was high, it was noted that the DIMTEST T statistic worked better in samples of 1000 and above.

In the conditions where the sample size was 200 and 300 and the test length was 10 items, it was noted that the rate of correct decision-making decreased in the results of the Dimensionality DETECT IDN index and the DIMTEST T statistics, regardless of the interdimensional correlation. Although the correct decision rate of DETECT IDN index and the DIMTEST T statistic increased as the test length increased, it could be argued that the correct decision rate of the DIMTEST T statistic decreased as the sample size decreased. It could be argued that nonlinear factor analysis worked better than the Dimensionality DETECT index and DIMTEST T statistics in the process of determining dimensionality with skewed data.

4. DISCUSSION and CONCLUSION

When the DIMTEST T statistic, Dimensionality DETECT IDN index, and Type I Error rates for Nonlinear Factor Analysis were examined in data with standard normal distribution, according to various sample sizes and different item numbers, the Nonlinear Factor Analysis (NOHARM) results were the most consistent in making the unidimensionality decision. In addition, although results of the DIMTEST T statistics were argued to be more consistent, it was thought that the use of DIMTEST T statistics in determining dimensionality in short tests would not be appropriate. In addition, it could be argued that the DETECT IDN index would be more appropriate to use with large samples and large test lengths. The DETECT IDN index should not be used in the dimensionality determination process, especially in short tests. When the DIMTEST T statistic, Dimensionality DETECT IDN index, and Nonlinear Factor Analysis (NOHARM) Type I Error rates were examined according to various sample sizes and different numbers of items with the data showing skewed distribution, it was observed that the results of DETECT IDN index were more consistent with the data showing skewed distribution compared to the data showing standard normal distribution. The results of DIMTEST T statistics and Nonlinear Factor Analysis were found to be more accurate in making the unidimensionality decision.

When the power rates for the DIMTEST T Statistics, dimensionality DETECT IDN index and Nonlinear Factor Analysis were examined according to various sample sizes, different numbers of items and different interdimensional correlation values in the data with standard normal distribution, it could be argued that the DIMTEST T statistic gave more accurate results in conditions with large samples. Especially in the cases when the sample size was less than 300, the error rate of DIMTEST T statistics increased significantly. At the same time, it could be argued that the DIMTEST T statistic was affected by the interdimensional correlation for the multidimensionality decision. In data with standard normal distribution, the results of the dimensionality DETECT IDN index and nonlinear factor analysis (NOHARM) seemed to make more accurate decisions than the DIMTEST T statistic under all conditions. DIMTEST T statistic, on the other hand, was found to make more accurate decisions in conditions with low interdimensional correlation and high sample sizes.

It could be argued that dimensionality determination methods gave less consistent results when the test length was less than 10 items with skewed distribution. On the other hand, although it was seen that the results of DETECT IDN index and Nonlinear factor analysis had higher inner consistency, it could be argued that the DIMTEST T statistic gave the right decisions in determining dimensionality when the sample size was 1000 and above.

As in every study, this study also had some limitations. The conditions discussed in this study were limited to sample size (200, 300, 500, 1000, 2000, and 3000), interdimensional correlation (0.25, 0.50, 0.75, 0.90), test length (10, 20, 30 items), and different ability distributions (standard normal distribution and skewed distribution). A similar study could be repeated with smaller samples and conditions with a larger test length. In addition, the research could be repeated by adding other variables. Based on the results of the DIMTEST T statistic used in

this study together with DETECT IDN index and nonlinear factor analysis and considering item pools and large samples of the large-scale tests used in the exams administered by the Student Selection and Placement Center (ÖSYM) or the Ministry of National Education (MEB), use of nonlinear factor analysis, the DIMTEST T statistic, and DETECT IDN index were found suitable to determine their dimensionality. In addition, nonlinear factor analysis seems to be a more accurate decision, especially instead of DETECT IDN index and the DIMTEST T statistic, in determining the dimensionality of short exams applied in the school environment.

In this study, 2PL and compensatory models were used. In future studies, together with 3PL models, the results can be examined using non-compensatory models, especially for tests containing items where one dimension does not compensate for the other dimension. In this study, test cases that were scored 1-0 were created. Considering the scale development and scale adaptation studies in education and psychology in future studies, the effectiveness of the same methods can be investigated in tests with multiple scores.

A similar study can be conducted by increasing the number of dimensions. The efficiency of the methods can also be tested on real data in the same study. The structure of the test discussed in this study is fixed and the test is semi-mixed. A similar study can be conducted with a different structure by varying the number of simple or complex items and different test structures can be used to test the effect of the test structure. Different item parameter sets can affect the performance of methods. Thus, in order to make the findings more generalizable, it could be useful to compare the present results with results based on a different set of item parameters. Considering the answers not given in the test items used in the exams held in our country, the efficiency of the methods can be tested by manipulating the amount of missing data in another study. While creating the skewed distribution in this study, skewness and kurtosis values (1.75, 3.75) in Fleisman's (1978) study were taken into account. Data set could be created considering the different deviations from the standard normal distribution, and the Type I Error and power study could be assessed for the dimensionality determination process.

In this study, Nonlinear Factor Analysis (NOHARM) from among parametric dimensionality determination methods and the DIMTEST T statistic from among non-parametric methods and Dimensionality DETECT IDN Index were used. In a different study, performances of other parametric and non-parametric methods in dimensionality determination can be tested. Among the parametric and non-parametric methods selected for the scope of this study, indices such as RMSR, Tanaka Goodness of Fit Index, and DETECT IDN index were used. In a different study, the Type I Error and power study can be assessed using other indices such as the approximate chi-square ($\chi^2_{G/D}$) statistic index obtained using the same methods. One of the important results of this study is that authors should consider the strengths and weaknesses of the methods in terms of the characteristics of the data while deciding or choosing the methods for determining dimensionality. Considering the difficulties in data collection processes, especially in the field of social sciences in our country, studies should be conducted using recommended methods in order not to reach inconsistent results due to the effect of sample size. Finally, for authors that would like to conduct a determination of dimensionality studies in the cases where research has not yet proven the superiority of one method over another, the application of multidimensionality methods may be useful if authors would like to have a comprehensive understanding of structure and dimensionality of the data before moving on to the scores obtained from the tests.

Acknowledgments

This paper was produced from the part of the first author's doctoral dissertation prepared under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Gul Guler: Investigation, Software, Methodology, Formal Analysis, Visualization, Resources, and Writing the original draft. **Rahime Nukhet Cikrikci:** Software, Methodology, Supervision, and Validation.

Orcid

Gul Guler  <https://orcid.org/0000-0001-8626-4901>

Rahime Nukhet Cikrikci  <https://orcid.org/0000-0001-8853-4733>

REFERENCES

- Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7(4), 255-278. https://doi.org/10.1207/s15324818ame0704_1
- Ackerman, T.A., Gierl, M.J., & Walker, C.M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37-51. <https://doi.org/10.1111/j.1745-3992.2003.tb00136.x>
- Davey, T., Nering M.L., & Thompson, T. (1997). Realistic simulation of item response data. ACT Research Report Series, 97-4. <https://files.eric.ed.gov/fulltext/ED414297.pdf>
- Embretson, S.E., & Reise, S. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARMbased statistics for testing unidimensionality. *Applied Psychological Measurement*, 31, 292-307. <https://doi.org/10.1177/0146621606294490>
- Fleisman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532. <https://doi.org/10.1007/BF02293811>
- Froelich, A.G., & Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. *Applied Psychological Measurement*, 32, 138-155. <https://doi.org/10.1177/0146621607300421>
- Gessaroli, M.E., & De Champlain, A.F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement*, 33, 157-179. <https://doi.org/10.1111/j.1745-3984.1996.tb00487.x>
- Göçer Şahin, S. (2016). *Yarı karışık yapılu çok boyutlu yapıların tek boyutlu olarak ele alınması durumunda kestirilen parametrelerin incelenmesi [Examining parameter estimation when treating semi-mixed multidimensional constructs as unidimensional]* [Unpublished doctoral dissertation]. Hacettepe University.
- Hattie, J. (1985). Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(8), 139 – 145. <http://dx.doi.org/10.1177/014662168500900204>
- Hattie, J., Krakowski, K., Rogers, H.J., & Swaminathan, H. (1996). An assessment of Stout's index of essential dimensionality. *Applied Psychological Measurement*, 20, 1-14. <https://doi.org/10.1177/014662169602000101>
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods* 6(1), 53-60. <https://doi.org/10.21427/D7CF7R>

- Hu, L-T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: *Conventional criteria versus new alternatives*. *Structural Equation Modeling*, 6, 1-55. <https://doi.org/10.1080/10705519909540118>
- Kaya, K.Ö., & Kelecioğlu, H. (2016). The effect of sample size on parametric and nonparametric factor analytical methods. *Educational Sciences: Theory & Practice*. 16(1), 153-171. <http://dx.doi.org/10.12738/estp.2016.1.0220>
- Kim, H.R. (1994). New techniques for dimensionality assessment of standardized test data. [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign, Department of Statistics.
- Ledasma, R.D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out Parallel Analysis. *Practical Assessment, Research & Evaluation*, 12 (2). <https://doi.org/10.7275/wjnc-nm63>
- Mroch, A.A., & Bolt, D.M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education*, 19 (1), 67-91. https://doi.org/10.1207/s15324818ame1901_4
- Nandakumar, R., & Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, 18(1), 41-68. <https://psycnet.apa.org/doi/10.2307/1165182>
- Özbek Baştuğ, Ö.Y. (2012). Assessment of Dimensionality in Social Science Subtest. *Educational Sciences: Theory & Practice*. 12(1), Winter: 382-385.
- Reichenberg, R.E. (2013). *A comparison of DIMTEST and generalized dimensionality discrepancy approaches to assessing dimensionality in item response theory* [M.S. dissertation, Arizona State University, Arizona]. <https://doi.org/10.3102%2F10769986018001041>
- Reckase, M.D. (2009). *Multidimensional item response theory*. Springer Dordrecht Heidelberg.
- Roussos, L.A., & Özbek, O.Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, 43, 215-243. <https://doi.org/10.1111/j.1745-3984.2006.00014.x>
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617. <https://doi.org/10.1007/BF02294821>
- Stout, W., Habing, B., Douglas, J., Kim, H.R., Roussos, L., & Zhang J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 19, 331-354. <https://doi.org/10.1177%2F014662169602000403>
- Sünbül, Ö. (2011). *Çeşitli boyutluluk özelliklerine sahip yapılarda, madde parametrelerinin değişmezliğinin klasik test teorisi, tek boyutlu madde tepki kuramı ve çok boyutlu madde tepki kuramı çerçevesinde incelenmesi* [Examining item parameter invariance for several dimensionality types by using classical test theory, unidimensional item response theory and multidimensional item response theory] [Unpublished doctoral dissertation]. Mersin University.
- Sünbül, Ö., & Seo, M. (2012). *Performance of test statistics for verifying unidimensionality*, [Conference presentation abstract]. 2012 Annual Meeting, April 12-16, Vancouver, British Columbia, CANADA
- Svetina, D. (2011). Assessing dimensionality in complex data structures: A performance comparison of DETECT and NOHARM procedures [Unpublished doctoral dissertation]. Arizona State University.
- Svetina, D., & Levy, R. (2014). A framework for dimensionality assessment for multidimensional item response models. *Educational Assessment*, 19(1), 35-57. <https://doi.org/10.1080/10627197.2014.869450>

- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, 27, 159-203. <https://doi.org/10.1177/0146621603027003001>
- Touron, J., Lizasoain, L., & Joaristi, L. (2012). Assessing the unidimensionality of the School and College Ability Test (SCAT, Spanish version) using non-parametric methods based on item response theory. *High Ability Studies*. 23(2), 183-202. <https://doi.org/10.1080/13598139.2012.735401>
- Zhang, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimension. *The Journal of Experimental Education*, 77 (2), 147-166. <https://doi.org/10.3200/JEXE.77.2.147-166>

Evaluation of impact factors of articles in scientific open access journals in Türkiye

Orhan Alav^{1,*}

¹Süleyman Demirel University, Knowledge Center, Isparta Türkiye.

ARTICLE HISTORY

Received: Feb. 22, 2022

Revised: July 19, 2022

Accepted: Aug. 25, 2022

Keywords:

SOBIAD Index,

Impact factor,

Open access journals.

Abstract: In this study, the phonographic view of the acceleration of scientific publishing in Türkiye has been revealed with TÜBİTAK/DergiPark data and the values of the measurements of the impact factors of scientific journals have been sampled with the SOBIAD Index data. SOBIAD Index dataset was used in the study. Using the "full count" research method, the data were analyzed by providing access to the entire mass, which is the research population, based on purpose-oriented descriptive analysis. In the calculation of the impact factors of the articles in the SOBIAD index, multiple parameters such as the total number of citations of the articles in the journal, citation comparison (percentage) and area-weighted citation impact, new metric joint values and the similarity criteria in the content evaluation were determined. In the study, the measurement and evaluation standards of international impact factor measuring institutions (WOS-SSCI, Google Scholar, Eigenfactor Metrix and Elsevier/Scopus Index) were also used. According to the results of the research, while the average value of the impact factors of scientific journals in Türkiye is 0.19, this is seen as 6,19 in WOS-SSCI. With the research, the examination of the impact factors of scientific journals and articles in Türkiye was presented as an original review through the SOBIAD index sample. In order to increase the quality and impact factor of journal/article in academic publishing in Türkiye, qualified growth is required rather than quantitative growth.

1. INTRODUCTION

Article sharing and use of open access journals directly or indirectly affect article authors, journal publishers, researchers and information centers in terms of productivity. The sharing of scientific information articles by open access journals provides the sharing of information=commodity resource, which is intellectual capital. In this context, the number of citations to scientific publications is one of the most important criteria used to measure the scientific, intellectual, economic and social impact of a publication. The organization of knowledge is not independent of its production. In addition to the bibliometric measurements based on the indicators of the scientific journals in Türkiye and the articles published in these journals, bibliometric network analyzes were also carried out. It has become important for open access journals to publish scientific articles produced by scientists and to measure the impact values of the articles in terms of value and value creation. Bibliometric measurement is one of the important criteria used to measure not only the number of citations to scientific publications,

*CONTACT: Orhan ALAV ✉ orhanalav@sdu.edu.tr 📍 Süleyman Demirel University, Knowledge Center, Isparta Türkiye.

e-ISSN: 2148-7456 /© IJATE 2022

but also the scientific and intellectual effects of authors. The average number of citations to a country's scientific publications is interpreted as an indicator of that country's scientific wealth (Tonta & Akbulut, 2021, p.389). In the modern world, knowledge has become a valuable commodity. Knowledge has a direct impact on the creation and provision of new knowledge, values, technologies, resources and employment. In this context, unrestricted open information has become important for corporate and legal identities that produce scientific knowledge, information users, and those who transform information into value and product. On the other hand, access to open information is mostly found in open access journals, and unhindered access can be provided. Therefore, with unhindered easy access to open access journals, the use of the journal, readability of the articles and citation levels have become important. In this study, it was evaluated how the accessibility of open access scientific journals in Türkiye, bibliometric measurements of the effects of citing articles to other scientific studies, and the rational and objective evaluation of the findings influence the developing open access publishing journals, open access platforms, authors and information users in the context of interaction. With our research study, the citation rates of the results of 871 open access scientific journals in the SOBIAD Index (SOBIAD Index, 2020a) were examined by subjecting them to a resource-based and productivity-oriented research. As a result, the knowledge and technology that develop at the global level have radically changed the publishing of academic journals. This change has developed in favor of publishers, writers and information users in terms of efficiency and production. According to the developing change, electronic format open access publishing, which is a new publication model, has forced scientific journals and publishers to open access publishing. Soon, the passwords of open access publishing will be provided by controlling the content licenses of scientific resources. This study reveals the measurement of impact factors, the use of source values, validity and operability of scientific journals that make open access publishing in Türkiye. We believe that the development of scientific publishing open access journals in Türkiye will contribute to scientific writers, information users and future scientific studies on similar topics. This study is the first to evaluate the measurement values of the impact factors of open access scientific journals in Türkiye through the SOBIAD Index sample. When the literature is examined, it is seen that many instruments are used to measure the impact factors. In the study, the population and the sample of the research consist of the same data set. In the study, all data values were accessed by using the full count method. With the study, the impact values of national-level journals in Türkiye were subjected to multiple regression of the SOBIAD Index data set to reach valid and reliable measurement values. In assessment and evaluation, the assessment and evaluation methods of WOS, Scopus, Eigenfactor Metrix and Google Scholar were also examined and referenced. We believe that this study will contribute to scientific studies on similar subjects after it.

2.1. Literature Review of the Research

Regarding the research topic, national and international scientific studies were examined. Among the prominent publications in the national literature, a limited number of studies such as Tonta and Akbulut's (2021) study titled "Factors Increasing the Citation Effect of Articles from Türkiye Published in International Journals", TÜBA, Türkiye Science Report, Türkiye Scientific and Technological Research Council-TUBITAK, ULAKBIM/Cahit Arf Bilgi Center (2021), Türkiye Scientific Publication Performance Reports: Journal Performance Indicators of Scientific Publications from Türkiye in WOS, by Alptekin Durmuşoğlu (2017) "A Study on Data Mining: Türkiye-Addressed Publications", by Al (2008), "Türkiye's Scientific Publication Policy: A Bibliometric Approach Based on Citation Indexes", and by Mecbure Aslan (2021), A Study on Work-Life Balance: "Bibliometric Analysis of Graduate Theses" were examined. In the international content of the literature, the impact factor calculations of the Web of Science (WOS) and the reports on the subject (Clarivate Journal Citation Reports: Reference Guide, 2011) were examined. These studies were followed by the metrics measurement studies of

Elsevier Science & Scopus Index (Elsevier, 2020) and Elsevier Science Index: Measuring a Journals Impact, (Elsevier, 2021a) and Google Scholar Index (Google, 2021). Also, the “Leiden Manifesto” for research articles and bibliometric research scales was reviewed (Hicks, & Wouters, 2015).

2. METHOD

Bibliometrics is a statistical analysis of existing studies and is used for quantitative analysis of articles in a particular field (Aslan, 2021, p.30). SOBIAD data set was used in the preparation of the bibliometric data set. The method of this research is based on the "counting method in bibliometrics" (Gauffriau, 2021, p.233). With this method, the basic elements of the bibliometric indicator shown by the findings in the examination and the factor analysis were made with the "full count" method in order to determine the bibliometric research findings. The "full counting" method was used in the study (Tutar & Erdem; 2020, p.245-295). A counting indicator functions as one of the essential elements of a bibliometric (Gauffriau, 2021, p.233). The "exact counting" method was used in the study (Tutar & Erdem; 2020, p.245-295). This method indicator functions as one of the basic elements of a bibliometric information (Gauffriau, 2021, p.233). Therefore, in this methodology, the research population and sample consist of the entire data set. Due to the limited research data, the research sample group consists of the entire research population. Qualitative and quantitative data tools were used together in the research. In the data collection process, the last 2 years' quantitative data of the SOBIAD Index data set and the qualitative data obtained from authorized persons constituted the data and process of the research. The data findings are presented together with the statistical results supported by visual graphics.

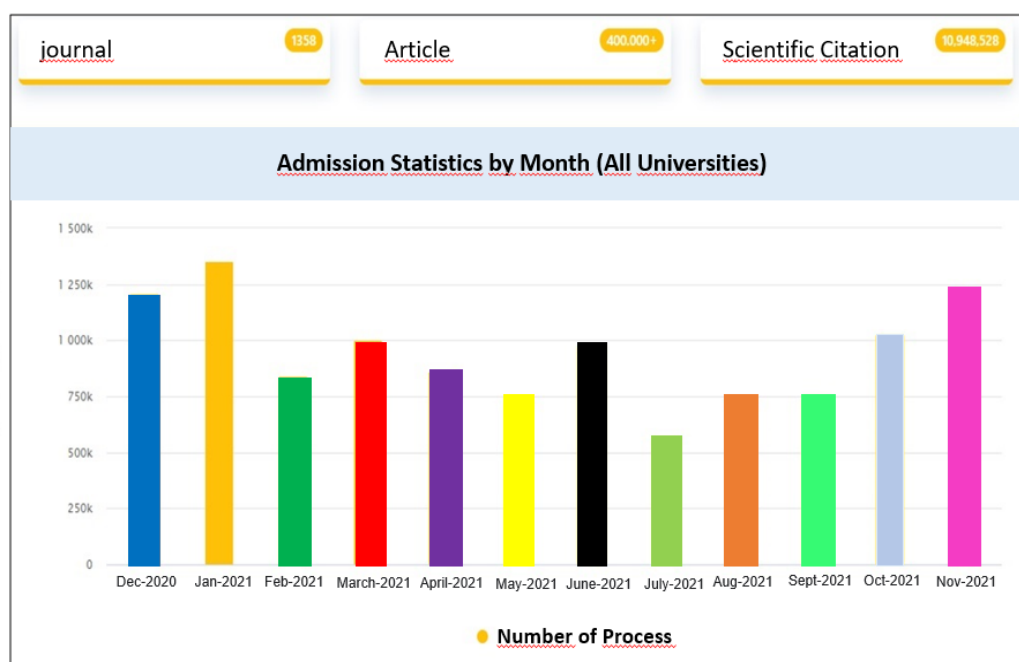
In the global world, information has become a valuable commodity. In this context, the importance of open access scientific journals is increasing day by day. Many publics, private institutions and legal entity owners in Türkiye publish scientific journals electronically. In this context, the data set of 871 scientific journals in SOBIAD Index, which indexes scientific journals in Türkiye, constitutes our research population (SOBIAD Index, 2020b). SOBIAD Index dataset is publicly available on the official website of the organization as open access. At the beginning of the research process, the general manager of the SOBIAD Index directory was informed that such research would be conducted, and the legal permission was obtained from the relevant institution for the research. In our research study, the journals in the SOBIAD index directory constitute the population and sample of the research. In addition, the impact factors of these journals and the content interactions of the citation numbers, the research methods specified in the subsections of this section, the research model, the research population and sample, data sources and data collection tool and data analysis are defined in the study. The findings of the research are indicated with graphs consisting of statistical values that blend quantitative data with factor distributions. In the conclusion part of the study, the findings were interpreted by expressing them as qualitative data supported with quantitative data. With citation analysis, scientific information about the literature flow is revealed by providing citation networks, informetric laws and productivity analysis. (Todeschini & Baccini, 2016, pp.19-20).

2.1. SOBIAD Index

SOBIAD Index is a private Turkish patented TR Index (TR Dizin) company (SOBIAD Index, 2021c). This firm examines the journals that contain the scientific articles with academic content in the fields of society, health and science, and measures the citation and citation impact values of the articles published in these journals. According to 2020 data, there are 1355 journals, 400000+ articles, 10949523 citations in SOBIAD Index (SOBIAD Index, 2021d). In the research, the impact values and citation results of 871 open access scientific journals

according to the 2019 data in the SOBIAD Index were examined by subjecting them to a resource-based and productivity-oriented research. The bibliographies of the electronic journals that have been published for at least 4 years in the field of Social Sciences and at least 3 years in the fields of Science and Health Sciences were searched in the SOBIAD Index; as a result of this searching, the citations made by the author/authors in their works were revealed. The data of the study consist of the data set of the SOBIAD index. Statistical (factor) analyzes and the resulting statistical values are detailed in the relevant section of the study (2.5. Distribution Data of Impact Factor of SOBIAD Index).

Figure 1. SOBIAD Index User Statistics Data.



Source: SOBIAD Index Statistics (2021), URL: <https://atif.SOBIAD.com/index.jsp?modul=istatistik>

In Figure 1, access by universities and users to the journals in the SOBIAD index is indicated along with the monthly entry statistics. And thus, users can access the data set in the SOBIAD index and have open access to journals, articles and scientific citations. Table 1 shows the number of transactions made by the users and the content information.

Table 1. SOBIAD User Transaction Frequency Rates (Last 1 Year).

Pocess	Number
11396826	Article detail view
4554611	View profile download
855907	Citation search
166448	Search by journal name
537927	View profile
62764	Search by title
66396	Search by author name
20820	Search in full text
16910	Search by keyword
13770	Audion listening

Source: SOBIAD Index Statistics (2021), URL: <https://atif.sobiad.com/index.jsp?modul=istatistik>

The purpose of SOBIAD Index is to reveal the citations made by the authors to other articles and books in the articles published in academic and scientific journals. For this purpose, it is aimed to determine the impact value of a journal registered in the database among other journals. It is possible to access the full texts and abstracts of registered journals with the SOBIAD Index. In addition, another aim is to statistically reveal the impact value of a journal with an impact value compared to another journal. With the SOBIAD Index, current articles can be viewed instantly and access to the abstracts and full texts of the articles can be provided upon request. Journal links will be active in the title of scanned journals. Thus, access to Türkiye-based academic and scientific journals will be faster and easier. SOBIAD Index Directory accepts corporate membership/subscription. Therefore, it appeals to both single-user and multi-user audiences. Journals and scanned articles in this index include publications in both Turkish and other languages, mostly found in Türkiye-based electronic journals. SOBIAD index is a citation index that performs citation search and bibliometric data analysis specialized in science, health and social sciences (SOBIAD Index, 2021e).

2.2. Journal Publishing and Indexing in Türkiye

In the academic world, the value of the scientific journal is evaluated in proportion to the index and the average number of citations of the articles accepted by the journal (Flint, 2021). The institutional curatorship of academic publishing in Türkiye is managed by the DergiPark Academic unit, which is run by ULAKBİM affiliated to TUBITAK (DergiPark, 2021a). This institution is a public institution that helps the journals to survive and to make quality publications at high standards by providing infrastructure support such as policy, standard, network and software to scientific journals published by public, private and legal persons that produce scientific output value in Türkiye. The purpose of the DergiPark Academic unit to provide these services is to ensure the development of academic *periodicals* in Türkiye in accordance with quality and standards, to increase the visibility and use of national academic journals all over the world, to ensure widespread and advanced use of a system that enables the management of journals in an electronic environment, to provide measurable clean data for the Türkiye TR national citation index (DergiPark, 2021b), DergiPark Academic unit office is not an institution that performs index operations. The public national citation index in Türkiye is carried out by the TR Index office, a sub-unit of ULAKBİM, which operates under TUBITAK (TR Index, 2021a). TR Index indexes the articles in scientific journals published between 1961 and 2021. Türkiye TR Index Office evaluates the article indexing processes by considering the publication ethical values (TR Index, 2021b) and standards determined by the institution (TR Index, 2021c). In Türkiye, the indexing process of academic journals that publish scientific publications other than "TR Index Institution" is carried out by SOBIAD Index company (SOBIAD Index, 2020c). Scanning model was used in the research. Monitoring and sectioning approaches were applied to the data set. Temporal developments and changes of the research sample were determined.

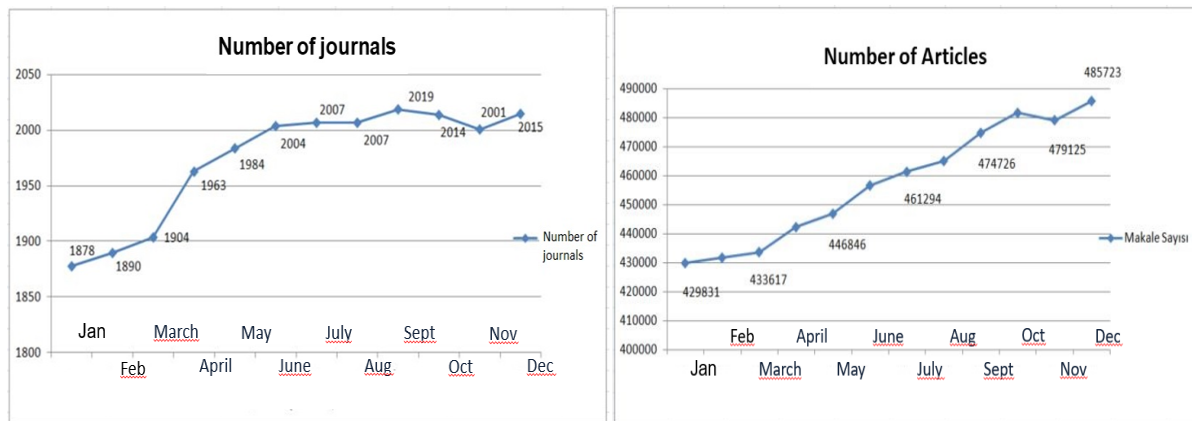
The research population and sample of the study consists of the data set that includes 871 scientific journals in the SOBIAD Index. The data set of the SOBIAD Index was used to evaluate the impact factors of the open access journals reached during the research process. SPSS and MAXQDA 2020 program were used as data collection tools. Descriptive statistics, quantitative content analysis and bibliometric analysis were used together in the analysis of the research data. The findings of the study were subjected to multiple regression in which qualitative and quantitative data were interpreted together and the resulting findings were revealed together with the statistical distribution data.

2.3. The Numbers of Scientific Journals, Publishers and Produced Articles in Türkiye

According to the 2022 official data of DergiPark Akademik (Figure 2), the number of scientific articles produced in Türkiye in a year is 578.128, the number of journals is 2.544, the number of publishers is 1.058, and the number of researchers is 496.544 (DergiPark, 2022).

When we examine the international and national research reports, it is seen that although there has been a measurable increase in scientific studies in Türkiye in recent years, it is still behind these developed countries when compared to OECD and European Union countries (Akçiğit & Tok, 2021, p.16). According to the scientific publication performance report data, Türkiye's journal impact factor averages for the years 2011-2015 are stated as 4.2. The data of this report consists of the impact factor values of the articles with Türkiye extension published in SSCI journals (TR Index, 2021d). Since an official result for the citation averages of the journals in the TR Index and DergiPark platform, which publishes scientific publications in Türkiye with Türkiye extension, has not yet been specified, the data in this field could not be reached.

Figure 2. DergiPark, Türkiye Journal and Article Development Chart (2020).



Source: DergiPark, URL: <https://dergipark.org.tr/tr/pub/page/about>

According to the 1996-2016 WOS data in the study of Tonta and Akbulut (2021) titled "Factors Increasing the Citation Effect of Articles Published in International Journals with a Turkish Address", the impact factor averages of scientific journals with Türkiye extension were expressed as 1.6 (Tonta & Akbulut; 2021, p.390). It can be predicted that these data are lower in scientific journals published at the national level in Türkiye. If TR Index journal impact factor data were available, it would be possible to have information about the national impact factors of journals and articles across Türkiye and to compare with the SOBIAD Index journal and article impact factor values, which is the subject of the research. However, the data gap in this area can be considered among the limitations of the research. In the contemporary world, bibliometric studies are considered as a part of publication policy (Al, 2020, p.14).

2.4. Evaluation of SOBIAD Index Impact Factors

871 journals in the SOBIAD Index constitute the data set of the research. In the research, the "full count" research method was used, and with this method, all information in the data set in the population (N) was accessed, and the information was obtained by examining the variable values (mean, ratio, variant and total values, etc.). In terms of the importance of the research (SOBIAD Index), the data set was examined by considering the whole population with the full count method rather than the sample (n) value, and the result value findings were determined. Therefore, an error that may arise from the estimation in the full count is minimized. In the research, 2019 Impact Factor of the journals scanned based on the information in the SOBIAD Index database is a numerical data about the citation status of the articles published in the journals scanned in the database. The calculation of the SOBIAD Index impact factor is done

as follows: The basic calculation logic in calculating the impact factor of a journal for any year is provided by dividing the number of citations for the previous year of the year to be calculated by the number of publications of two previous years (SOBIAD Index, 2020f). To illustrate, the rate of the citations in 2019 of the articles published in 2017 and 2018 gives the journal's 2019 impact factor (Karamustafaoğlu, 2007, p.2).

$$A = \frac{\text{Number of Citations 2019}}{\text{Number of Publications 2017-2018}}$$

The contents such as reader letters, translated articles, news published in the journals in the form of articles are not included in the calculation of the impact factor of the journals in the SOBIAD Index. In the impact factor calculation logic of the SOBIAD Index, the research metric measurement criteria of the Web of Science's impact factor calculation methodology (WOS, 2021) and the Elsevier Science-Science Direct SCOPUS Index (Elsevier, 2020) and components are similar. In the calculation, the measurements and standards of "Eigenfactor Metrix" were also taken into consideration (Eigenfactor Metrix, 2021a). In the evaluation of the citations made by authoritative publishers and journals with high impact factor, the measurement values of this metric have been considered in the score calculations of the articles that are in demand for citations (Eigenfactor Metrix, 2021b). Under normal circumstances, there is a direct interaction with the citation numbers of the article published in the journal, the journal impact factors and citations of other cited articles (SOBIAD Index, 2020g). However, the presence of more than one component in the evaluation of the citation and score effect of the articles published in the journals in the SOBIAD Index was reflected as small values in the calculation of the journal impact factors. Therefore, in this measurement, the citation score of any article published in the journals included in the relevant index may not be directly reflected in the impact value of the relevant journal. There are multiple reasons for this situation. The content effects of multiple components are also decisive in calculating the impact factors of the journals in the SOBIAD Index. These criteria are the institutional structure of the journal, being subject to international open access agreements and open access policies, national and international participant content of journal science and referee boards, the number of local and foreign authors, publication language of the journal, local, national and international content dimension of the journal, home page contents, publication periods of the journal, institutional or legal personality of the journal, local (regional), national and international dimension of the journal, the thematic nature of the journal and the contribution of scientific publications to local development, plagiarism status and levels of the articles in the journal, transparency of the journal and the commercial structure of the journal. In addition, the evaluation of the journal's referee practices (open refereeing, blind refereeing, peer refereeing, etc.), the objectivity and consistency of the journal editor and / or editorial working groups, the social media and social media interactions of the journals, how user-friendly the journal homepages are and journal publisher and / or publisher information are the other criteria. The advisory board of 5 people took part in the calculation of the impact factors of the journals in the SOBIAD Index. This committee includes independent faculty members selected from universities, metric software specialist engineers, librarians and index managers. In the measurement of the impact factors of the journals, this committee examined the presence of the above-mentioned components in the journals and undertook the task of making small scores and adding them to the metric. Similar metric components in the calculation of journal impact factor are also similar in organizations that measure international impact factor. Within the scope of the research, all 871 journals representing both the research population and the research sample were analyzed quantitatively and qualitatively, based on the information obtained from the data set, according to the 2019 SOBIAD Index data. In Türkiye, 678 journals out of 871 journals in the SOBIAD

Index were exposed to an impact factor in the range of 1.117 & 0.000. Since 192 journals in the research dataset were not exposed to the impact factor, the data were not evaluated. In order to ensure the reliability of the research and to reach all of the data, the "full count" sampling method was used in data collection. It is aimed to reach all units of the main mass, which is the research population, and to examine the entire population with the full count method (Tutar & Erdem, 2020, p.242-295). With this method, it is possible to reach all the elements of the population. Due to the physical form (graphic) constraints of the research, all of these journals could not be included in the graphics specified in the research, and the selected journals were included in the sample. This situation can be shown as a limitation of the research. The fact that the research is a current due diligence and compilation study for national and international literature shows the original aspect of the research. It is thought that the research will shed light on the future scientific studies.

2.5. Distribution Data of Impact Factor of SOBIAD Index

It is important to measure the impact factors of scientific journals in the world and in Türkiye based on multiple components because scientific articles have turned into the most valuable commodity that contributes to social life and economy. In this context, the impact factors of scientific journals constitute the scientific exchange rate of the journal, article, author/s and countries. In addition to scientific values, the impact factors of journals are important as trust values in every field (Law & Leung, 2019, p.734-742). In this study, the metric components in the analysis of the research data, the total number of citations (citations) of the articles in the journal according to the user date range as well as the Scopus Index measurement (Elsevier, 2021b) values, citation comparison (percentage) and area-weighted citation effect, new metric joint values in metric measurements and the number of views were considered in the evaluation of the content. The graphical contents of the data set of SOBIAD Index consist of the following data.

2.5.1. Initial scatter chart

According to SOBIAD database 2019 data, out of 871 scientific journals publishing in Türkiye, there are 3 journals with the impact factors between 1.171 and 1.000 in the national literature (SOBIAD Index, 2020i). These data also show that they remain low in terms of impact factor efficiency (see Table 2).

Table 2. *Impact Factor Initial Scatter Chart (Between 1.117-1.000).*

The Journals	Number of citations	Impact factor
Journal for the Education of Gifted Young Scientists	41	1.171
Journal of Banking and Financial Studies (BAFAD)	16	1.067
Online Journal of Technology Addiction & Cyberbullying	18	1.059

2.5.2. Second scatter chart

According to the information in the SOBIAD Index dataset, there are 19 journals in the national literature with an impact factor between 1.000 and 0.700. Since it is not possible to include all the journals in this field due to the limitations of the study, 10 journals selected from the impact factor range specified in Table 3 are indicated in the scatter chart.

Table 3. *Impact Factor Second Scatter Chart* (between 1.000 - 0.700).

The Journals	Number of citations	Impact factor
Journal of Hasan Ali Yucel Faculty of Education	31	0.939
Journal of Applied Social Sciences	13	0.929
Journal of Education and Science	127	0.882
Journal of Turkish Librarianship	11	0.846
The Turkish Journal on Addictions (ADDICTA)	36	0.837
E-Kafkas Journal of Educational Research	25	0.833
Bartın University Journal of the Faculty of Economics and Administrative Sciences	37	0.804
Journal of Dicle University Ziya Gokalp Faculty of Education	41	0.804
Journal of 100. Yil University Faculty of Education	93	0.802
Journal of Travel and Hotel Management	61	0.792

2.5.3. Third scatter chart

According to the information in the SOBIAD Index dataset, there are 64 journals in the national literature with an impact factor between 0.700 and 0.500. Since it is not possible to include all these journals in terms of the limitations of the study, 12 journals selected within the impact factor range specified in [Table 4](#) are indicated in the scatter chart.

Table 4. *Impact Factor Third Scatter Chart* (0.700 - 0.500).

The Journals	Number of citations	Impact factor
Journal of Geography	13	0.722
Journal of Bayburt Faculty of Education	48	0.716
SDU International Journal of Educational Studies	16	0.696
International Journal of Active Learning	9	0.692
Western Anatolian Journal of Educational Sciences	13	0.684
Ihlara Journal of Educational Research	19	0.679
Journal of Hacettepe University Faculty of Education	81	0.675
Süleyman Demirel University Visionary Journal	41	0.672
Tourism Academic Journal	34	0.642
Journal of Accounting and Finance	75	0.641
Ege Journal of Education	38	0.594
Journal of Ahi Evran University Kirsehir Education Faculty	131	0.590

2.5.4. Fourth scatter chart

According to the information in the SOBIAD Index dataset, there are 287 journals in the national literature with an impact factor between 0.500 and 0.350. In terms of the limitations of the study, but not all these journals can be included, 10 journals selected within the impact factor range specified in [Table 5](#) are indicated in the scatter chart.

Table 5. *Impact Factor Fourth Distribution Chart* (between 0.500 - 0.350).

The Journals	Number of citations	Impact factor
Turkish Online Journal of Qualitative Inquiry	17	0.486
International Review of Economics and Management	14	0.483
Ankara University Faculty of Educational Sciences Journal of Special Education	27	0.482
Gaziantep University Journal of Sport Sciences	34	0.479
Eskişehir Osmangazi University Turkish World Appli- cation and Research Center Education Journal	10	0.476
Management and Economics: Journal of Celal Bayar Un. Faculty of Economics and Administrative Sciences	51	0.472
Journal of Mother Tongue Education	63	0.47
Istanbul University Journal of Sport Sciencesv	15	0.469
Bilgi Journal of Social Sciences	14	0.467
Harran Education Magazine	7	0.467

2.5.5. Fifth scatter chart

According to the information in the SOBIAD Index dataset, there are 287 journals in the national literature with an impact factor between 0.350 and 0.120. In terms of the limitations of the study, but not all of these journals can be included, 12 journals selected within the impact factor range specified in Table 6 are indicated in the distribution chart.

Table 6. *Impact Factor Fifth Scatter Chart* (0.350 - 0.120).

The Journals	Number of citations	Impact factor
Boundless Journal of Education and Research	9	0.300
Journal of Sociology Studies	11	0.297
LAU Journal of Social Sciences	8	0.296
Journal of Academic Research and Studies	18	0.295
Journal of Human and Social Sciences Research	119	0.289
International Journal of Progressive Education	32	0.288
Mediterranean Journal of Educational Research	18	0.198
Journal of Uludağ University Faculty of Education	16	0.198
KTU Social Sciences Institute Journal of Social Sciences	9	0.196
Journal of Discourse Philology	8	0.195
Journal of Erzincan University Institute of Social Sciences	17	0.193
Marmara University Journal of Economic and Administrative Sciences	10	0.192

2.5.6. Sixth scatter chart

According to the information in the SOBIAD Index dataset, there are 224 journals in the national literature with an impact factor between 0.120 and 0.000. In terms of the limitations of the study, but not all these journals can be included, 13 journals selected within the impact factor range specified in Table 7 are indicated in the scatter chart.

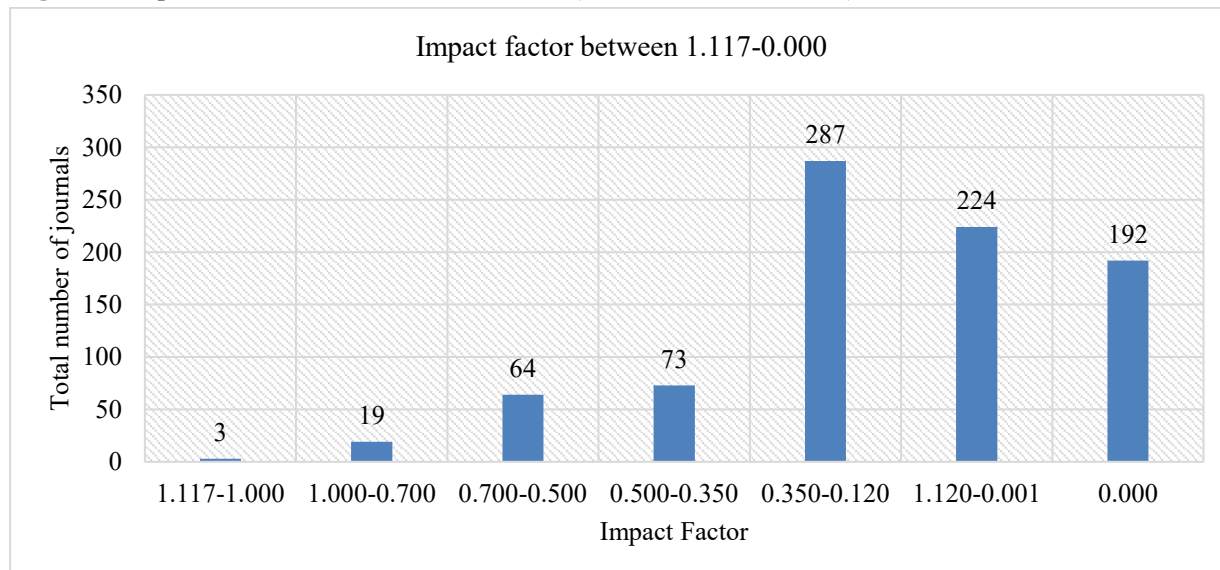
Table 7. *Impact Factor Sixth Scatter Chart (0.120 - 0.000).*

The Journals	Number of citations	Impact factor
Turkish History Education Journal	6	0.105
Journal of Artuklu Academy	3	0.103
Journal of N. Erbakan University Faculty of Theology	2	0.095
Journal of BEU Faculty of Theology	3	0.094
Journal of Cukurova Researches	4	0.093
Journal of Contemporary Turkish History Studies	6	0.085
International Journal of Sport Culture and Science	8	0.084
Journal of Information and Document Studies	1	0.083
Journal of Belgi	2	0.061
Eurasian Journal of International Studies	11	0.057
Journal of Dokuz Eylül University Faculty of Letters	1	0.033
Journal of Management and Economics Research	2	0.012
Selcuk University Journal of Turkic Studies	1	0.010

2.5.7. Seventh scatter chart

According to the information in the SOBIAD Index dataset, the total number of journals with the impact factor between 1.117 and 0.000 is indicated in the national literature. In this context, there are 3 journals with the impact factor between 1.117-1.000, 19 journals between 1.000-0.700, 64 journals between 0.700-0.500, 73 journals between 0.500-0.350, 287 journals between 0.350-0.120, 224 journals between 1.120-0.001 and 224 journals between 1.200-0.001. On the other hand, there are totally 192 journals without impact factors (Figure 3).

Figure 3. *Impact Factor Seventh Scatter Chart (between 1.117 – 0.000).*



3. DISCUSSION and CONCLUSION

The citations and impact factors of the scientific studies of the countries are similar to the scientific exchange rates of that country, just like the value of the national currency and the cross exchange rates. The higher the citation values of the articles in scientific journals published in a country and the impact factor of the journals, the more valuable the prestige of the country in science and its place among the world's nations. Journal impact factor is often used for multi-faceted interactions such as knowing about the scientific quality of individual research articles and individual journals, evaluating journals, articles and authors, and citation influence on other scientific studies. Journal impact factor is a quantitative measure based on

the ratio of annual citations in a particular journal to the total citations in that journal in the previous 2 years, and is not a mandatory measure for evaluating research quality. The research includes examining both the citation-based impact factor values and components of scientific journals from Türkiye in the SOBIAD Index. In the review, the SOBIAD Index measurement criteria were compared with the TR Index criteria and the impact values of the internationally valid indexes (WOS-SSCI) and the Elsevier-Scopus Index, and it was tried to add richness to the research. The content and impact values of national scientific journals in Türkiye can be evaluated as a compilation study based on the SOBIAD Index data sample.

Within the scope of the research, the SOBIAD Index data set consists of 871 journals representing the population of the research, based on the applied "full count" method. According to the research data and findings, 678 of 871 journals representing the population were exposed to the impact value; on the other hand, 192 journals in the data set were not included in the data evaluation process because they were not exposed to the impact value (n/a). In the study, the impact values of the journals exposed to the impact factor in the SOBIAD Index were evaluated between 1.117 and 0.000. To ensure the reliability of the research and to reach all of the data, the "full count" sampling method was used in data collection. It is aimed to reach all units of the main mass, which is the research population, and to examine the entire population with the full counting method. With this method, it is possible to reach all the elements of the population.

Within the scope of the research, a total of 871 journals were reached. Due to space, time and financial constraints in the research, all of these journals could not be included in the graphics specified in the research, and selected journals were included. This situation can be shown as a limitation of the research. On the other hand, the fact that the research is a due diligence and compilation study in the national literature shows the original aspect of the research, and it is thought that the research will shed light on future studies.

When we evaluate the research on the scale of Türkiye, it is seen that multiple components are effective in the interaction of the impact factor data of 871 journals in the SOBIAD Index. In Web of Science and SOBIAD Index journal impact factor measurements, impact factor values were started with the number 1. The values above the number 1 (>1) were considered plus increasing values, and the values below the number 1 (<1) were considered as decreasing values, and the values zero and below zero (0) were considered as unoperated (n/a).

When the results obtained in the study are compared with the WOS-SSCI (6.19) data (Clarivate Journal Citation, 2011) it is seen that the average impact factor data of SOBIAD (0.19) is well below the average value of the journal impact factor of WOS-SSCI (SOBIAD Index, 2020b). With the research, the impact value of scientific journals in Türkiye, article citation impact values, the standards used by the journals, open access policies, editorial boards, international interactions, local, national and international contributions, reliability, indexes and the determining components of the total contribution values are also discussed. In the evaluation of the study content, measuring the quality of the journal may reveal a subjective value, rather, evaluating the journal impact factors and the values of the impact factors of the article citations, the author/s who cited the article, the publication(s) cited and the prestige of the journals in which they are published can be a more objective, better and more efficient measure (Habibzadeh & Yadollahie, 2008, p.171). According to the research findings, the articles are cited in cases where the impact factor of the journal is less than a value of <1 [Ex: Journal of History, Culture and Art Studies, impact factor (0.024), number of article citations: 10], in these cases, the journal content analysis can be compared to the one stated above. We can state the following reality through the SOBIAD Index data sample, that the mean value of the impact factor 0.19 which constitutes the scientific exchange rates of scientific journals in Türkiye, has been determined to be at an extremely low level. The impact factor calculations of SOBIAD

Index, as in the multiple components used in the impact factor calculation of the WOS data include journal citation indicators, cited document rates, original documents, citation time intervals, document content values, non-open access documents, percentage values of all open access documents, rates of publications on the gold and green path, national and international collaborations and interactions of publications, sphericity values of documents, journal impact factor (JIF) quarter, half and full time frame measures, JIF rank values, citation effects, global base areas of the documents, categorical domains and normalized citation effects, percentage value distributions of the documents across all data, citations within all elements of the documents /reference percentage values, fixed base percentage values of the documents, lifetimes of citations, article/document impact values, urgency use indexes of the documents, eigen-factor values, impact values of first and last author's works, proportions of hybrid documents, published countries and average citation values of countries, publishing house data and publishers' impact values, additional categories-commission reports, JIF percentage values and JIF ranking, and so on. The numerical indicator values of the components and their current presence status could not provide strong content support to the study data, and the research was conducted on limited data. This situation limited the research.

It has been determined that the "impact factor value measurement values of the journals indexed by TR Index could not be reached or were not made, so the journals were only included in the system with the TR Index acceptance criteria and commission decisions (TR Index, 2021c). Therefore, the impact factor of any journal published in TR Index and DergiPark Academic has not been revealed as an official result. In this context, it is a major shortcoming that journal impact factors are not measured in these institutions where content support is given to national journal publications such as TR Index and DergiPark.

We believe that using and applying similar criteria and data mining methods (Durmuşoğlu, 2017, p.1118-1120) on a world scale, as in WOS and Scopus Indexes, will be beneficial in measuring the impact factors of scientific journals and articles in Türkiye. With the research, it is seen that SOBIAD Index private company is the only institution where we can get official real results about journal impact factors in Türkiye. It has been determined that the impact factor value averages 0.19 of the SOBIAD Index and the journals included in the research are lower than those of OECD and European Union countries WOS 6.19 (Clarivate-Journal Citation, 2021, May 04). In this case, it is beneficial to evaluate Türkiye by the Council of Higher Education (YÖK), universities, TÜBİTAK, publishing houses, publisher editors and information producers/authors, and to produce and implement more rational and valid real policies. As stated in the science report of TÜBA, in order for Türkiye to be among the leading countries in producing science and technology (in scientific publishing), it is necessary to identify the failing parties and to intervene in these points with the right policies (Akçiğit & Tok, 2020, p.11). With the research, a compilation study was conducted by evaluating the current impact factors of scientific publishing journals in Türkiye through SOBIAD Index sample. It can be thought that the research study will contribute to similar scientific studies that will be carried out after it.

Acknowledgments

We would like to thank SOBIAD company, which opened SOBIAD Index dataset for research, and its esteemed director Prof. Dr. Serdar YAVUZ, TÜBİTAK/ULAKBİM TR Index institution and from İzzet Baysal University Associate Professor Ahmet Tuncay ERDEM who supported the research in the statistical analysis of the study.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Orhan Alav  <https://orcid.org/0000-0003-4577-0984>

REFERENCES

- Akçiğit, U., & Özcan-Tok, E. (2021). Türkiye Bilim Raporu-2020 [Türkiye Science Report-2020]. *Turkish Academy of Sciences*, 1-115. <http://www.tuba.gov.tr/tr/yayinlar/suresiz-yayinlar/raporlar/turkiye-bilim-raporu>
- Al, U. (2020). *Türkiye'nin Bilimsel Yayın Politikası: Atıf Dizinlerine Dayalı Bibliyometrik Bir Yaklaşım [Türkiye's Scientific Publication Policy: A Bibliometric Approach Based on Citation Indexes]* [Unpublished doctoral dissertation]. Hacettepe Üniversitesi. <http://bby.hacettepe.edu.tr/yayinlar/133.pdf>
- Aslan, Mecbure (2021). Trends in Work-Life Balance Research: A Bibliometric Analysis. *International Journal of Applied Engineering & Technology*, 3(2), 29-38. https://www.researchgate.net/publication/357279715_Trends_in_WorkLife_Balance_Research_A_Bibliometric_Analysis
- Clarivate-Journal Citation (2021, May 04). Journal Citation Reports: Reference Guide. https://clarivate.com/webofsciencegroup/wpcontent/uploads/sites/2/2021/06/JCR_2021Reference_Guide.pdf
- DergiPark (2020, July 22). DergiPark Academic Services. *DergiPark*. <https://dergipark.org.tr/pub/page/about>
- DergiPark (2021a, June 22). DergiPark Academic Services. *DergiPark*. <https://dergipark.org.tr/tr/pub/page/about>
- DergiPark (2021b, June 21). DergiPark Academic Institutional. *DergiPark*. <https://dergipark.org.tr/tr/>
- DergiPark (2022, July 02). DergiPark Academic Publishing Data. *DergiPark*. <https://dergipark.org.tr/tr/search?section=articles>
- Durmuşoğlu, A. (2017). Veri Madenciliği Üzerine Bir Araştırma: Türkiye Adresli Yayınlar [A Research on Data Mining: Publications Addressed in Türkiye]. *Elektronik Sosyal Bilimleri Dergisi*, 16(62), 1111-1122.
- Eigenfactor Metrix (2021a, August 13). Eigenfactor and Article Influence. <http://www.eigenfactor.org/index.php>
- Eigenfactor Metrix (2021b, September 07). Metrix Measurement Criteria. <http://www.eigenfactor.org/about.php>
- Elsevier (2020, May 02). Elsevier Research Metrics Guidebook & Scopus: The Primary Data Source for Elsevier's Research Metrics Inside Research Metrics in Scival: Methods and Use. pp.1-68. https://www.elsevier.com/_data/assets/pdf_file/0020/53327/ELSV-13013-Elsevier-Research-Metrics-Book-r12-WEB.pdf
- Elsevier (2021a, 12 May). Elsevier Science Index: Measuring a Journals Impact. *Elsevier*. https://www.elsevier.com/solutions/scopus/howscopusworks/metrics?dgcid=RNAG_Sourced_400000285&gclid=EA1aIQobChMIqY7BgOrY9AIVTrvVCh1-FgyTEAAYASAAEgLwVfD_BwE
- Elsevier (2021b, June 03). Scopus Index: Metric to Show Journal, Article & Author Influence. <https://www.elsevier.com/solutions/scopus/how-scopus-works/metrics>
- Flint, C. (2021, November 24). A Look at Metrics Measuring the Impact of Publications. ICE Publishing is Part of the Institution of Civil Engineers 1 July 2021. https://www.icevirtuallibrary.com/page/ice-news/156-good-journalimpact?gclid=EA1aIQobChMIsGA18%20w9AIVIRoGAB1QXQsZEAAYBCAAEgKMnvD_BwE

- Gauffriau, M. (2021). Counting methods introduced into the bibliometric research literature 1970-2018: A review. *Quantitative Science Studies*, 2(3), 932-975. <https://direct.mit.edu/qss/article/2/3/932/102387/Counting-methods-introduced-into-the-bibliometric>
- Google (2021, April 04). Google Scholar Metrics. *Google*. <https://scholar.google.com/intl/tr/scholar/metrics.html>
- Habibzadeh, F., & Yadollahie, M. (2008). *Journal Weighted Impact Factor: A Proposal*. *Journal of Informetrics*, 2(2), 164-172. <https://www.sciencedirect.com/science/article/pii/S175115770800014X>
- Hicks, D., & Wouters, P. (2015). Bibliometrics: The Leiden Manifesto for Research Measurements. *Nature*, 520, 429-431. <https://www.nature.com/articles/520429a.pdf>
- Karamustafaoglu, Orhan (2007). Citation analysis of papers published by universitybased Turkish physicists in journals listed in SCI. *Ad Astra*, 6(1), 1-8. <https://www.ad-astra.ro/journal/10/karamustafaoglu.pdf>
- Law, R., & Daniel, L. (2019). Journal Impact Factor: A Valid Symbol of Journal quality?. *Tourism Economics the Business and Finance of Tourism and Recreation*, 25(5), 734-742. <https://doi.org/10.1177/1354816619845590>
- SOBIAD Index (2020a, February 02). SOBIAD Index Impact Values. <https://atif.SOBIAD.com/index.jsp?modul=impact-faktoru>
- SOBIAD Index (2020b, 02 February 02). SOBIAD 2019 Journal Metric Statistics Data. <https://atif.SOBIAD.com/index.jsp?modul=impact-faktoru>
- SOBIAD Index (2021c, December 02). SOBIAD Index Co. <https://atif.SOBIAD.com/>
- SOBIAD Index (2021d, July 04). SOBIAD Index Directory Statistics Data. <https://atif.SOBIAD.com/index.jsp?modul=istatistik>
- SOBIAD Index (2021e, July 14). SOBIAD Index Institutional Content Information. <https://atif.SOBIAD.com/index.jsp?modul=kurumsal>
- SOBIAD Index (2020f, April 14). SOBIAD Index 2019 Impact Value Journal Impact Factor Report. https://atif.SOBIAD.com/files/impact_final_2019.pdf
- SOBIAD Index (2021g, August 22). SOBIAD Citation Index User Guide. <https://atif.SOBIAD.com/index.jsp?modul=kullan%C4%B1m-klavuzu>
- SOBIAD Index (2020i, April 10). SOBIAD Index Statistical Data; 2019 Impact Factors. https://atif.SOBIAD.com/files/impact_final_2019.pdf
- Todeschini, R., & Baccini, A. (2016). Bibliographic information published by theDeutsche Nationalbibliothek, and chapter in the book, pp.19-20. <https://onlinelibrary.wiley.com/doi/epdf/10.1002/9783527681969.fmatter>
- Tonta, Y., & Akbulut, M. (2021). Uluslararası Dergilerde Yayımlanan Türkiye Adresli Makalelerin Atıf Etkisini Artıran Faktörler [Türkiye Address Published in International Journals Factors Increasing the Citation Effect of Articles]. *Türk Kütüphaneciliği Dergisi*, 35(3), 388-409. <https://doi.10.24146/tk.933159>
- Tutar, H., & Erdem, A.T. (2020). *Bilimsel araştırma yöntemleri (1.baskı)* [Scientific Research Methods (1st edition)]. Seçkin Yayınevi.
- TUBITAK (2021, April 14). TUBITAK. <https://www.tubitak.gov.tr/>
- TR Index (2021a, August 07). TR Index Office. <https://trdizin.gov.tr/>
- TR Index (2021b, July 22). Türkiye TR Index Ethics Guide. <https://trdizin.gov.tr/rehber/>
- TR Index (2021c, July 07). Türkiye TR Index Journal Search Indexing Standards. <https://trdizin.gov.tr/kriterler/>
- TR Index (2021d, August 02). Scientific Publication of the Provinces of Türkiye Performance Report 2011-2015. <https://cabim.ulakbim.gov.tr/bibliyometrik-analiz/turkiye-bilimsel-yayin-performans-raporlari/>
- WOS (2021, August 13). Web of Science Impact Factor. <https://journals.mejsp.com/blogsingl.e.php?lang=en&id=25&name=Web%20of%20Science%20Impact%20Factor>

To what extent are item discrimination values realistic? A new index for two-dimensional structures

Abdullah Faruk Kilic ^{1,*}, Ibrahim Uysal ²

¹Adıyaman University, Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Adıyaman, Türkiye

²Bolu Abant İzzet Baysal University, Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Bolu, Türkiye

ARTICLE HISTORY

Received: Apr. 05, 2022

Revised: July 18, 2022

Accepted: Sep. 01, 2022

Keywords:

Item discrimination,
Multi-dimensional
structures,
Classical test theory,
Corrected item-total
correlation.

Abstract: Most researchers investigate the corrected item-total correlation of items when analyzing item discrimination in multi-dimensional structures under the Classical Test Theory, which might lead to underestimating item discrimination, thereby removing items from the test. Researchers might investigate the corrected item-total correlation with the factors to which that item belongs; however, getting a general overview of the entire test is impossible. Based on this problem, this study aims to recommend a new index to investigate item discrimination in two-dimensional structures through a Monte Carlo simulation. The new item discrimination index is evaluated by identifying sample size, item discrimination value, inter-factor correlation, and the number of categories. Based upon the results of the study it can be claimed that the proposed item discrimination index proves acceptable performance for two-dimensional structures. Accordingly, using this new item discrimination index could be recommended to researchers when investigating item discrimination in two-dimensional structures.

1. INTRODUCTION

Since the social science field has latent traits that cannot be observed directly, researchers use indicators to identify these traits. When latent traits (concepts) are not clearly expressed hypothetically, researchers often develop a scale to measure them. When scales are developed to measure latent traits like success, attitude, interest, and belief, there are two common measurement theories; namely, the Classical Test Theory (CTT) and Item Response Theory (IRT). Since this research focuses on CTT, this paper only explains this theory and is limited to CTT. The CTT is used in numerous scale development studies due to its typical implementation in the software, easy-to-understand structure, suitability for social sciences, and relatively weak assumptions.

Moreover, CTT results are similar to and have high-level relationship with results obtained from different theories (ex. IRT) in certain situations (DeVellis, 2006; Fan, 1998). However, it is essential to note that there are also disadvantages, such as item and person statistics being

*CONTACT: Abdullah Faruk Kilic ✉ abdullahfarukkilic@gmail.com 📍 Adıyaman University, Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Adıyaman, Türkiye.

dependent on the test and sample (Kohli et al., 2015). Therefore, the sampling procedure that must be representative of the population when developing a scale becomes an important subject. Otherwise, item statistics (discrimination and difficulty) will fail to reflect the reality.

CTT assumes that each score contains the true and error scores related to the examined trait. The normal distribution of the error score is another assumption. Although CTT seems to focus on the items, it focuses on the entire test (DeVellis, 2006). When developing CTT-based scales, it is reasonable to apply item analysis before factor analysis (Kline, 2000) because item analysis can help decide the items to be kept in or removed from the scale (Green & Salkind, 2014). For item analysis, it is necessary to focus on exploratory statistics, item difficulty, and discrimination (Kline, 2005). The validity of test scores depends on the item validity in the test. Especially when the unidimensional structure is considered, a high-level relationship between item analysis and factor analysis is found (Kline, 2000). Therefore, it is reasonable to collect evidence towards the validity and reliability of the scores obtained from the scale after conducting item analysis. Item validity is investigated during item analysis and is frequently determined by item discrimination.

On the other hand, item discrimination is commonly investigated with discrimination index (D) and item-total correlation. The D index compares the lowest and highest performance groups in the test (Kaplan & Saccuzzo, 2018). Accordingly, the difference between the correct numbers of the upper and lower 25% (or 33%) groups is taken and divided by the number of individuals in a group (Brown, 1988; Metsämuuronen, 2020a). Cureton (1957) suggested using 27% for the upper and lower groups. 27% is a critical ratio that separates the tails from the mean in the standard normal distribution of errors. Item discrimination is also the strength of the relationship between an item in the test and other items. Therefore, it also measures the item's relationship with the true score (DeVellis, 2006). In other words, it is the relationship between one item and all items. Therefore, it is called item-total correlation. Item-total correlation is investigated with phi coefficient, tetrachoric, biserial, and point-biserial correlation coefficients for binary (1-0) scored items and Pearson product-moment correlation coefficient for polytomous scored items (ex. open-ended tests) (Kline, 2000). It can be seen that some applications calculate correlation after reducing the investigated item score from the total score. That application was named the corrected item-total correlation (Macdonald & Paunonen, 2002). Values obtained without corrected item-total correlation are biased (Kline, 2000) since correction is essential, especially when 5-6 items are in the test (Kline, 2005). The correlation will be higher than its actual value as item scores will be included in the total score with no correction.

In unidimensional structures, when the item-total correlations are positive and high, these items can distinguish low and high-level individuals from each other in terms of the trait measured by the item, which is the basis of item discrimination. Item-total correlation values show that the item discrimination varies between -1 and 1, like the Pearson product-moment correlation coefficients (Brown, 1988). A negative item discrimination value indicates inverse discrimination between individuals with low and high ability in terms of the measured trait. Negatively discrimination means that while individuals with a high trait have a low score on the item, individuals with a low trait have a high score. The increased discrimination of an item with a positive value indicates that individuals with low and high trait levels are effectively distinguished (Macdonald & Paunonen, 2002). There is a cut-off point for item discrimination. Most researchers state that item-total correlation must be at least .30 (Kline, 2000; Nunnally & Bernstein, 1994).

The related literature review shows relatively more common discrimination coefficients, examples of which include the D index, point-biserial correlation coefficient, biserial correlation coefficient, phi coefficient, tetrachoric correlation coefficient, and rank biserial correlation coefficient. There are less common discrimination indexes such as the B index, the

agreement statistic, Davis discrimination index, Flanagan's correlation coefficient, Flanagan's corrected correlation coefficient, and phi/(phi max) coefficient (Liu, 2008). In other words, researchers related to the discrimination coefficient have always been in a search for what the best discrimination coefficient is since there are currently more than 20 discrimination coefficients available in the literature. Although item discrimination has been investigated for a long time, the research on this subject is still ongoing.

Some studies compare item discrimination indexes or recommend a new index when the current literature is reviewed. For example, Bazaldua et al. (2017) stated that the literature has complicated results regarding item discrimination and compared point-biserial, biserial, and point-biserial with the item-rest score, phi coefficient for binary data which categorize using median value, discrimination index. The estimators showed different performances in the analysis by differentiating test length, item difficulty, item discrimination, and test score distribution. In another study, Liu (2008) compared the point-biserial and biserial correlation coefficients with the D coefficient calculated with different lower and upper group percentages (10%, 27%, 33%, and 50%). Item-factor correlations showed the closest result to the item-total correlation. In recent years, Metsämuuronen (2020a) conducted research in order to generalize the D index, a simple and robust coefficient. D index that gives consistent results even when there are outliers is generalized for items scored in more than two categories while e vector properties are used in generalization. In addition, Metsämuuronen (2020b) recommended Somers' D index as an alternative to item-total correlation and corrected item-total correlation. As a result of the simulation study, the researcher found that Somers' D index estimated values below the real value for items with four and more categories.

Even when multi-dimensional structures are found in CTT-based scale development studies, it is seen that the item-total correlation or corrected item-total correlation is examined when examining the item discrimination (Ak & Alpulu, 2020; Akyıldız, 2020; Çalışkan, 2020; Tarhan & Yıldırım, 2021). However, such analysis might lead to underestimates of item discrimination. Therefore, items that should be included in the scale might be removed from the scale. To avoid item removal, item-factor correlation or corrected item-factor correlation might be investigated (see also Green & Salkind, 2014). However, such an approach requires much effort and fails to provide information about the entire test. Our study built on this problem aims to provide an alternative approach to investigate item discrimination of scales developed or adapted based on the Classical Test Theory (CTT). We proposed a new item discrimination index for two-dimensional structures and tested it using the Monte Carlo simulation under the conditions of sample size, the magnitude of item discrimination, inter-factor correlation, and the number of categories. The newly developed item discrimination index can determine the discrimination of each item at one time by considering the scale's dimensionality. The inter-factor correlation can be considered with this newly proposed index, and a direct relationship can be established between the score for the entire test and items.

Our study contributes to the literature by eliminating the mentioned limitations regarding item discrimination and providing evidence for item discrimination by considering the dimensionality and inter-factor correlation in two-dimensional structures. Therefore, this study is considered necessary and aims to contribute to the literature by a) recommending a new item discrimination index for two-dimensional structures, b) investigating the recommended item discrimination index under numerous simulation conditions, and c) the new recommended discrimination index can be used in scales development studies. The detailed information regarding this index is provided as follows:

1.1. New Index

A vector length in analytic geometry is used by considering the inter-factor correlation to develop a two-dimensional item discrimination index. The item discrimination values

calculated for each dimension of an item create a vector in the space. Let us consider a two-dimensional example: In a two-dimensional structure, an item's correlation with the first and second dimensions is expressed by two values, D_1 and D_2 . These points can be represented as ordered pairs in two-dimensional Euclidean space, which is presented in Figure 1.

Figure 1. D_1 and D_2 points on the plane.

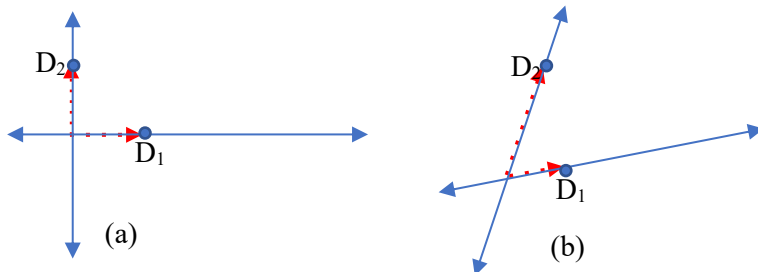


Figure 1-(a) shows that the x and y axes represent vertical two dimensions. D_1 and D_2 points represent the discrimination of an item in each dimension. D_1 and D_2 points can be represented as ordered pairs $(a,0)$ and $(0,b)$. At the same time, these points indicate a vector on a plane. Similarly, D_1 and D_2 points in Figure 1-(b) are points on the affine coordinate system. The affinity of the axes indicates a correlation between the dimensions. The correlation between the dimensions equals the cosine of the angle between these two vectors (Gorsuch, 1974). In this case, the product of these vectors is found to learn about the discrimination on both dimensions. The starting point of this study is this idea. The parallelogram method is applied to find out the product of these points, and the product vector is found as equation 1:

$$\overline{V_b^2} = a^2 + b^2 + 2abc\cos\theta \quad 1$$

(Lange, 2009). Here, a represents the x-axis value, b represents the y-axis value, and θ represents the angle between the x and y axes. Since the axes will have a 90° angle when they are perpendicular, $\cos(90^\circ) = 0$ will give the resultant vector as $\overline{V_B} = \sqrt{a^2 + b^2}$. However, when the axis is affine, the coordinates on these affine systems are first transformed into the rectangular coordinate system. The product vector is calculated as in the perpendicular coordinate system. The transformation matrix in equation 2 is used for this transformation (Deakin, 1998).

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \quad 2$$

Accordingly, X' and Y' values correspond to discrimination in the affine coordinate system, while X and Y values are the correspondence in the rectangular coordinate system. θ is the angle between the two axes. When the equation system in Equation 2 is solved, X and Y values are obtained. Since X' and Y' values correlate with each item's dimension for two-dimensional structures, these are known as numerical values. θ value can be obtained from the correlation between two dimensions. Since the correlation (r_{xy}) between two dimensions is $\cos(\theta)$ (Gorsuch, 1974), which will be $\arccos(r_{xy}) = \theta$, the value obtained here can be used for calculating $\sin(\theta)$. Thus, two unknown values in the equation system will be X and Y . If this equation system is solved:

$$X' = \cos\theta.X + \sin\theta.Y \quad 3$$

$$Y' = -\sin\theta.X + \cos\theta.Y \quad 4$$

will be obtained. Here, if we multiply equation three to $(-\cos\theta)$ and equation four to $(\sin\theta)$, we obtain X and Y variables:

$$-\cos\theta.X' = -\cos^2\theta.X - \cos\theta.\sin\theta.Y \quad 5$$

$$\sin\theta.Y' = -\sin^2\theta.X + \sin\theta.\cos\theta.Y \quad 6$$

equations are obtained. If each side of the Equations 5 and 6 are summed:

$$\sin\theta.Y' - \cos\theta.X' = (\sin^2\theta + \cos^2\theta).X \quad 7$$

$$X = \sin\theta.Y' - \cos\theta.X' \quad 8$$

equations are obtained. Thus, the X variable is found. X variable can be written in Equation 3, and similar operations are followed for the Y variable:

$$Y = -\sin\theta.X' + \cos\theta.Y' \quad 9$$

by writing the X and Y variables obtained from here to Equation 1 *a* and *b* variables, a two-dimensional discrimination index is obtained.

1.2. An Example of a New Index

Let us assume that an item's discrimination index for the first dimension (correlation) is .50, and the discrimination index for the second dimension (correlation) is .20 on a two-dimensional scale, then the inter-factor correlation is .30. Let us calculate the two-dimensional discrimination coefficient of an item obtained from a two-dimensional scale: Here, $X' = .50$ and $Y' = .20$ because the X' and Y' values in the new discrimination index are the correlation of the item for two dimensions. Since the correlation between the two dimensions is given as .30, we have $\cos\theta = .30 \Rightarrow \arccos(0.30) = \theta$. Here, $\theta = 72.54^\circ$ is obtained. When these values are written to Equations 8 and 9:

$$X = \sin(72.54^\circ).0.20 - \cos(72.54^\circ).0.50 \quad 10$$

$$Y = -\sin(72.54^\circ).0.50 + \cos(72.54^\circ).0.20 \quad 11$$

equations are obtained. X and Y values are obtained as 0.0407 and -0.4169, respectively. When X and Y values are written to Equation 1 respectively as *a* and *b* and written to $\cos\theta = .30$:

$$\xi = \sqrt{0.0407^2 + (-0.4169)^2 + 2.(0.0407).(-0.4169).0.30} \quad 12$$

The equation is obtained. ξ results as 0.4065 when equation 12 is completed. Accordingly, for a two-dimensional scale, the correlation of an item with the first and second dimensions is .50 and .20, respectively. The discrimination for both dimensions is obtained as .41 when two dimensions are considered together.

2. METHOD

This study investigated a new item discrimination index for two-dimensional structures in a Monte Carlo simulation. In Monte Carlo simulation studies, the data are generated to fit the desired distribution properties (Bandalos & Leite, 2013) and analyzed in line with the purpose of the study.

2.1. Simulation Conditions

In this study, the sample size (200, 500, and 1000), the magnitude of item discrimination (.30, .50, and .70), inter-factor correlation (.00, .30, and .50), and the number of categories (2, 3, 5 and 7) were the simulation conditions and the fixed simulation condition was two-dimensional structures (see the further details in the data analysis section).

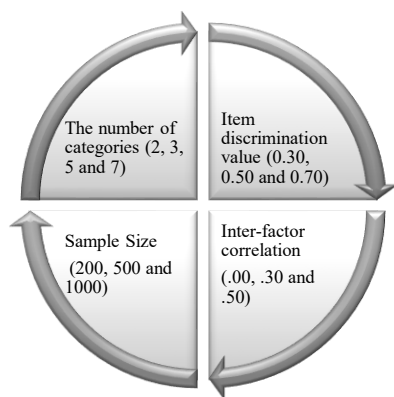
200, 500, and 1000 conditions were determined for the sample size. Kılıç and Koyuncu (2017) reviewed the scale adaptation studies in Turkey and found that more than half of the studies had samples between 100 and 349, and the average was 244. On the other hand, another study investigating the scale development studies in Turkey (Koyuncu & Kılıç, 2019) reported that more than half of the studies investigated included 300 or more individuals. Goretzko et al. (2021) systematically reviewed scale development studies and stated that more than half of the studies had a 400 or higher sample size. For this reason, considering the item discrimination index mainly reported for scale development and adaptation studies, the sample size was selected as 200 and 500. The 1000 sample size condition was included in this study to investigate the effects of increased sample size on the results.

.30, .50, and .70 conditions were determined for the magnitude of item discrimination. Since item discrimination between the .30-.39 range suggest that the item can be directly included in the scale/test (Crocker & Algina, 2008), the .30 condition was added to the study. On the other hand, since the item discrimination was desired to be .40 and above, the .50 and .70 conditions were added to the research as the conditions where the item discrimination was medium and high, respectively.

.00, .30 and .50 conditions were investigated for inter-factor correlation. .00 inter-factor correlation suggests no relationship between the dimensions, i.e., the dimensions are perpendicular. The .00 inter-factor correlation condition was added since item-total correlation was investigated while the item discrimination was calculated. Thus, it was aimed to examine the results that would emerge when the total score is taken in a situation where the total score should not be taken. On the other hand, the inter-factor correlation is generally reported and investigated as .30 in empirical (Li, 2016) and simulation studies (Cho et al., 2009; Curran et al., 1996; Flora & Curran, 2004; Foldnes & Grønneberg, 2017). Therefore, this simulation condition was added to the study. .70 inter-factor correlation condition was added due to high correlation between the dimensions in order to investigate the item-total score correlation results when getting a total score would cause no problems.

The number of categories of variables was manipulated as 2, 3, 5, and 7 in this study. The scale items are often Likert-type, and Likert-type items are generally scored as five-point scores (Lozano et al., 2008). Therefore, five was added as the category number to the study. On the other hand, three category conditions were added to the study since 3-point scales were used for children. Two conditions were added since there might be achievement tests with multiple options, yes/no, or a control list. Lastly, seven category number condition was included in the study to investigate the effects of increased category number on the discrimination index. [Figure 2](#) briefly shows the simulation conditions.

Figure 2. Simulation conditions.

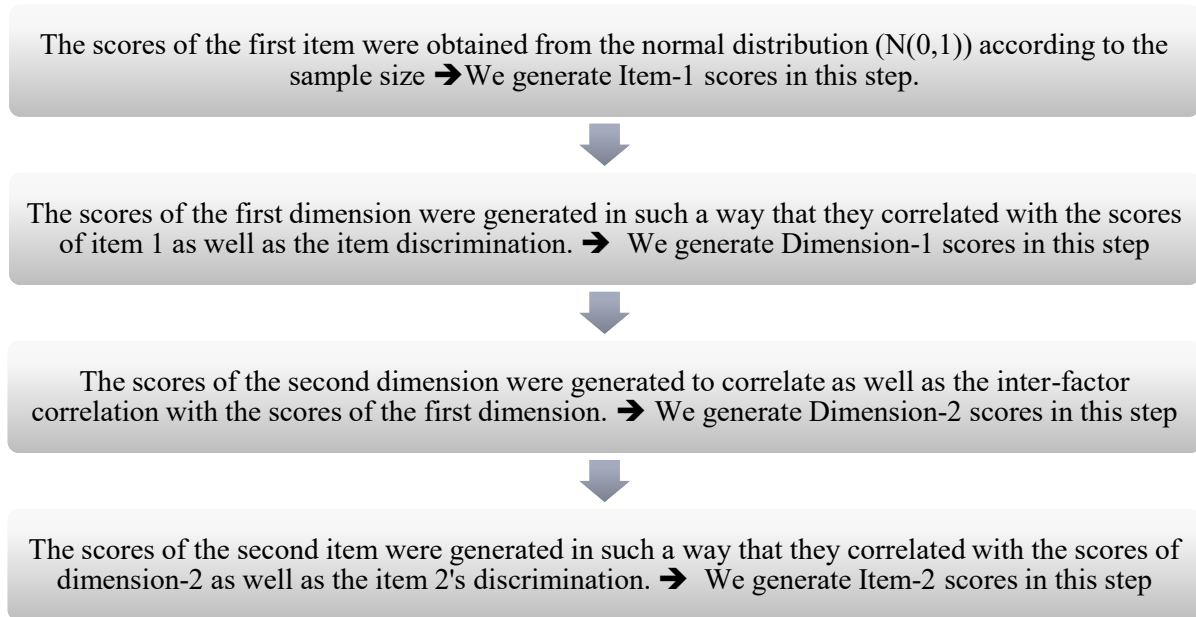


This simulation study was carried out with fully cross design. As seen in [Figure 2](#), a fully crossed design was applied, and the simulation was run for $4 \times 3 \times 3 \times 3 = 108$ conditions. 1000 replication was applied for each condition.

2.2. Data Analysis

We used R software (R Core Team, 2021) for data generation and came up with four variables: the first was item 1 scores, the second one was total score for dimension 1, the third one was total score for dimension 2, and last one was item 2 scores. The data generation process was given in [Figure 3](#). Also, the data generation R codes were added (see [Appendix 1](#)).

Figure 3. Data generation process.



After data generation was performed, we added the scores of dimension-1 and dimension-2 to obtain the total scale score. We calculated item-total correlation using items (item-1 and item-2) scores and total scale scores. Thus, the correlations of the items with the scores obtained from the whole test were examined.

The R software's stats (R Core Team, 2021) package was used to calculate the proposed two-dimensional item discrimination index. The item-total and item-factor correlation were examined and the proposed two-dimensional discrimination index worked under simulation conditions was determined. Therefore, the graphics show the item-factor correlation, item-total correlation, and two-dimensional item discrimination index results. These results were used for a descriptive inference. Additionally, a one-way analysis of variance (ANOVA) was applied to investigate which conditions would have a more effect on item discrimination.

3. FINDINGS

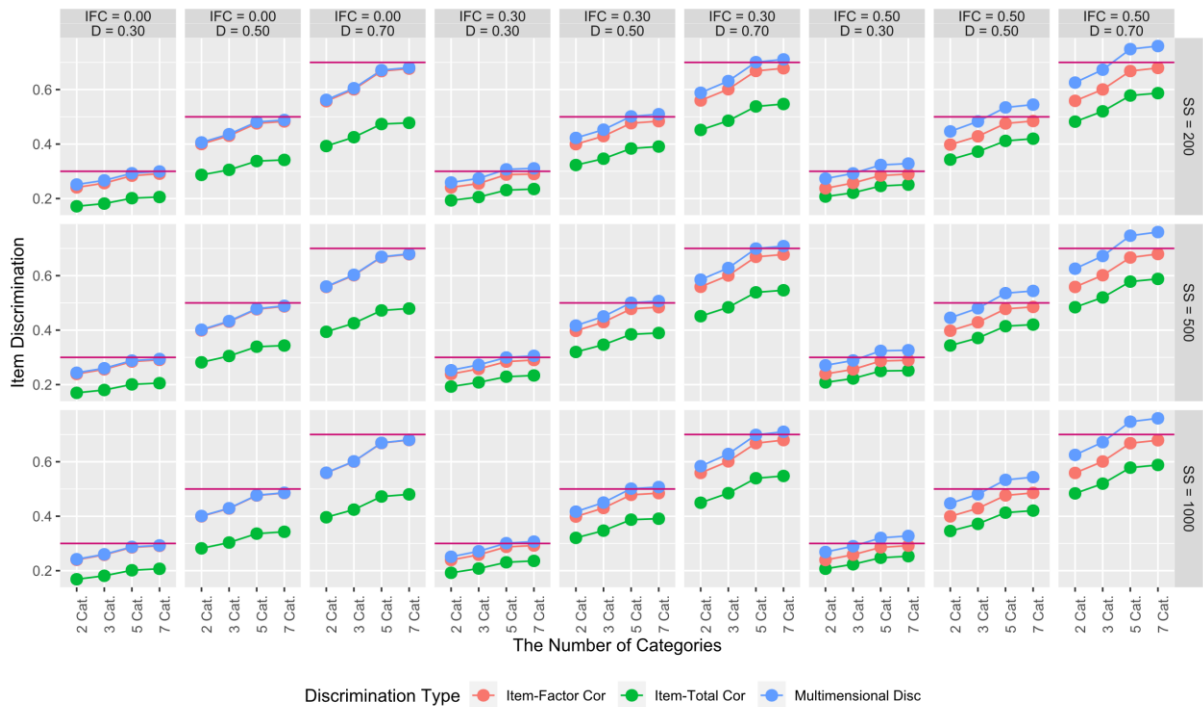
The average values obtained from item discrimination methods as a result of this study are given in [Figure 4](#). Additionally, these values are given in the [Appendix-2](#) for researchers who want to take a detailed look at these results. [Figure 4](#) shows that the correlation between the dimensions is .00, i.e., when two dimensions are orthogonal, and the recommended two-dimensional discrimination index and the item-factor correlation revealed similar results. The calculated values are more accurate since the data became closer to continuous as the category number increased.

When the inter-factor correlation was .00, it could be stated that the item-total correlation was underestimated for all magnitude of item discrimination conditions. One reason is to examine

item-total correlation by taking the total score from two vertical dimensions. The simulation conditions with a .00 inter-factor correlation between the dimensions showed a better performance for the newly recommended method.

When the inter-factor correlation was .30, it is possible to say that the recommended item discrimination index had a higher value than that of the item-factor correlation. Since the correlation coefficients made a more reasonable estimation with the increased category number, the item discrimination indexes increased in 7-category items. However, the graphic shows that the corrected item-total score correlations fail to give results close to the actual values in any conditions.

Figure 4. Discrimination indexes obtained from simulation conditions.



When the conditions with a .50 inter-factor correlation and discrimination were investigated, it was observed that the two-dimensional discrimination index was overestimated. Although increasing the inter-factor correlation and magnitude of the item discrimination to .70 deviates the results of the recommended two-dimensional item discrimination index from the actual value, the value should be .70 and estimated as .76 at most. Accordingly, overestimation could be stated as approximately 9%.

The one-way analysis of variance conducted to investigate which simulation conditions affected the values obtained from the item discrimination methods revealed that the sample size had no significant effect on the item discrimination [$F_{(2,312)}=.04, p=.97$]. There is a significant difference between category number [$F_{(3,312)}=189.21, p=.00$], inter-factor correlation [$F_{(2,312)}=70.33, p=.00$], magnitude of item discrimination [$F_{(2,312)}=4906.54, p=.00$], and item discrimination methods [$F_{(2,312)}=668.99, p=.00$]. When the effect size was investigated, the eta-square value was found to be .97 for the magnitude of item discrimination conditions, .81 for item discrimination method, .65 for category number, and .31 for inter-factor correlation. Accordingly, the most impactful factor on item discrimination estimations was the magnitude of item discrimination. According to Green and Salkind (2014), the eta-square value of .14 shows a high impact size. Based on this, it could be stated that the eta-square values obtained for the magnitude of item discrimination, item discrimination method, category number, and inter-factor correlation had a significantly high impact.

4. DISCUSSION and CONCLUSION

A new item discrimination index was obtained for two-dimensional structures in the current study, which was carried out based on the discrepancies in examining item-total score correlations for item discrimination in multi-dimensional constructs based on a test score or factor score. After numerous investigations on category number, sample size, the magnitude of item discrimination, and inter-factor correlation based on Monte Carlo simulation, the newly obtained item discrimination index can be used for two-dimensional structures. This study shows a significant difference between item-total correlation, item-factor correlation, and recommended item discrimination for two-dimensional structures. This finding matches the results of Bazaldúa et al. (2017) as they failed to find similarities between item discrimination methods when multiple item discrimination methods were compared. Such results support the hypothesis stated in the problem situation of the current research.

Item-factor correlations showed similar results with the newly recommended index, especially when the correlation between the factors was extremely low (correlation was taken as .00 to exemplify this condition in the current study). It can be seen that the recommended item discrimination index for two-dimensional structures showed adequate performance when the fact that item-factor correlations should be investigated for two-dimensional structures. Moreover, the recommended two-dimensional discrimination index could be used when the correlation between dimensions was extremely low. Also, the results in item-factor correlations provide ideas about the factor, not the entire test. Considering that item-total correlations underestimate the discrimination, it is beneficial to use the newly item discrimination index for two-dimensional structures that can be calculated at once.

When the inter-factor correlation increased to .30, although the item-total correlation was closer to the actual value of the item discrimination, the value was deficient. This situation may cause researchers to be mistaken when making decisions about items. Item-factor correlations of two-dimensional structures and newly item discrimination index revealed similar values. Although the item-factor correlations were highly close, these revealed slightly lower results than the actual values. When the factor correlation was .00 or .30, the item-total and item-factor score correlations showed differences. However, contrary to this finding, Liu (2008) stated that the item-total and item-factor correlations had similar results. It is believed that the difference between our specific study and Liu's (2008) study was due to mixed-format test usage.

As the inter-factor correlation increased to .50 and the discrimination value to .70, the highest value was obtained for two-dimensional structures in the new item discrimination index. When inter-factor correlation was .50 and item discrimination was .70, the behaviors of the item discrimination methods differentiated more. When the inter-factor correlation was .50, the item-total correlations were underestimated; the item-factor correlations were estimated close to the actual value, and the new two-dimensional item discrimination index was overestimated. The overestimation percentage for the new two-dimensional discrimination index was 9 at most.

The sample size was not found as a significant independent variable to impact the estimation of item discrimination. One of the reasons might be that the smallest sample size was 200. In addition, the magnitude of item discrimination and inter-factor correlations were found as significant independent variables. Therefore, inter-factor correlation and magnitude of item discrimination should be considered by researchers when item discrimination is investigated. It is important to note that the item discrimination index for two-dimensions might be slightly overestimated when the inter-factor correlation is high (approximately .70).

There is another important finding in this study. The methods can identify the item discrimination more accurately as the category number increases. However, Metsämuuronen (2020b) recommended that Somers' D coefficient estimates two-category data better. The difference between the new item discrimination index in the current study and

Metsämuuronen's (2020b) study is due to the different mathematical basis. The recommended index in this study is in line with corrected item-total and item-factor correlations. In addition, Metsämuuronen (2020a) generalized the D index for items scored in more than two categories by using vectors in the study to generalize the D index. Similarly, calculations of the discrimination index for two-dimensional structures in the current study were based on vectors. The item discriminations were underestimated when the category number was low. However, the literature does not show different cut-off points for item-total correlations according to category number. Although there is no rule of thumb, the cut-off point for the new item discrimination index can be determined as .30 when data are in 5 and 7 categories. Considering that the data have 2 or 3 categories, it is rational to accept the new item discrimination index up to .20. Accordingly, different cut-off points can be determined for different categories and discrimination indexes by conducting simulation studies in future studies.

This study has certain limitations. Since the recommended discrimination index is newly developed, we investigate it for only two-dimensional structures. Future studies can focus on three or more dimensional structures. Moreover, the item discrimination index for two-dimensional structures might be revised based on the studies with 3, 4, 5, or higher dimensions and added to open-source software (Python, R, etc.). This study has not covered items with cross-loading. In future studies, the performance of the developed item discrimination index can be examined in cases where items have cross-loading.

The item discrimination index for two-dimensional structures revealed as a result of this study can be recommended only for two-dimensional structures. We named the recommended item discrimination index as ξ coefficient. Therefore, researchers using the recommended index for two-dimensional structure could show the index as a ξ coefficient.

Acknowledgments

This study was presented as an oral presentation at the International Congresses on Education (ERPA) in 2021.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Abdullah Faruk Kilic: Investigation, methodology, developing process, visualization, software, formal analysis, and writing-original draft. **Ibrahim Uysal:** Developing process, methodology, resources, validation, and writing-original draft.

Orcid

Abdullah Faruk Kilic  <https://orcid.org/0000-0003-3129-1763>

Ibrahim Uysal  <https://orcid.org/0000-0002-6767-0362>

REFERENCES

- Ak, M.O., & Alpullu, A. (2020). Alpak akış ölçeği geliştirme ve Doğu Batı üniversitelerinin karşılaştırılması [Alpak flow scale development and comparison of east west universities]. *E-Journal of New World Sciences Academy*, 15(1), 1-16. <https://doi.org/10.12739/NWSA.2019.14.4.2B0122>
- Akyıldız, S. (2020). Eğitim programı okuryazarlığı kavramının kavramsal yönden analizi: Bir ölçek geliştirme çalışması [A conceptual analysis of curriculum literacy concept: A study of scale development]. *Electronic Journal of Social Sciences*, 19(73), 315–332. <https://doi.org/10.17755/esosder.554205>

- Bandalos, D.L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.). Information Age.
- Bazaldúa, D.A.L., Lee, Y.-S., Keller, B., & Fellers, L. (2017). Assessing the performance of classical test theory item discrimination estimators in Monte Carlo simulations. *Asia Pacific Education Review, 18*, 585–598. <https://doi.org/10.1007/s12564-017-9507-4>
- Brown, J.D. (1988). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5*(1), 19–31. <https://doi.org/10.1177/026553228800500102>
- Cho, S.-J., Li, F., & Bandalos, D.L. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement, 69*(5), 748–759. <https://doi.org/10.1177/0013164409332229>
- Crocker, L., & Algina, J. (2008). *Introduction of classical and modern test theory*. Cengage Learning.
- Cureton, E.E. (1957). The upper and lower twenty-seven per cent rule. *Psychometrika, 22*, 293–296. <https://doi.org/10.1007/BF02289130>
- Curran, P.J., West, S.G., & Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16–29. <https://doi.org/10.1037/1082-989X.1.1.16>
- Çalışkan, A. (2020). Kriz yönetimi: Bir ölçek geliştirme çalışması [Crisis management: A scale development study]. *Journal of Turkish Social Sciences Research, 5*(2), 106–120.
- Deakin, R. (1998). 3-D coordinate transformations. *Surveying and Land Information Systems, 58*(4), 223–234.
- DeVellis, R.F. (2006). Classical test theory. *Medical Care, 44*(11), 50–59. <https://doi.org/10.1097/01.mlr.0000245426.10853.30>
- Fan, X. (1998). Item Response Theory and Classical Test Theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357–381. <https://doi.org/10.1177/0013164498058003001>
- Flora, D.B., & Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491. <https://doi.org/10.1037/1082-989X.9.4.466>
- Foldnes, N., & Grønneberg, S. (2017). The asymptotic covariance matrix and its use in simulation studies. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(6), 881–896. <https://doi.org/10.1080/10705511.2017.1341320>
- Goretzko, D., Pham, T.T.H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology, 40*(7), 3510–3521. <https://doi.org/10.1007/s12144-019-00300-2>
- Gorsuch, R.L. (1974). *Factor analysis*. W. B. Saunders.
- Green, S.B., & Salkind, N.J. (2014). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (7th ed.). Pearson Education.
- Kaplan, R.M., & Saccuzzo, D.P. (2018). *Psychological testing: Principles, applications, and issues*. Cengage Learning.
- Kılıç, A.F., & Koyuncu, İ. (2017). Ölçek uyarlama çalışmalarının yapı geçerliği açısından incelenmesi [Examination of scale adaptation studies in terms of construct validity]. In Ö. Demirel & S. Dinçer (Ed.), *Küreselleşen dünyada eğitim* [Education in the globalized world] (pp. 1202–1205). Pegem.
- Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). Routledge.
- Kline, T.J.B. (2005). *Psychological testing: A practical approach to design and evaluation* (3rd ed.). Sage.

- Kohli, N., Koran, J., & Henn, L. (2015). Relationships among classical test theory and item response theory frameworks via factor analytic models. *Educational and Psychological Measurement*, 75(3), 389–405. <https://doi.org/10.1177/0013164414559071>
- Koyuncu, İ., & Kılıç, A.F. (2019). The use of exploratory and confirmatory factor analyses: A document analysis. *Education and Science*, 44(198), 361-388. <https://doi.org/10.15390/EB.2019.7665>
- Lange, M. (2009). A tale of two vectors. *Dialectica*, 63(4), 397-431. <https://doi.org/10.1111/j.1746-8361.2009.01207.x>
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936–949. <https://doi.org/10.3758/s13428-015-0619-7>
- Liu, F. (2008). Comparison of several popular discrimination indices based on different criteria and their application in item analysis [Master of Arts, University of Georgia]. http://getd.libs.uga.edu/pdfs/liu_fu_200808_ma.pdf
- Lozano, L.M., García-Cueto, E., & Muñoz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, 4(2), 73–79. <https://doi.org/10.1027/1614-2241.4.2.73>
- Macdonald, P., & Paunonen, S.V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, 62(6), 921-943. <https://doi.org/10.1177/0013164402238082>
- Metsämuuronen, J. (2020a). Generalized discrimination index. *International Journal of Educational Methodology*, 6(2), 237-257. <https://doi.org/10.12973/ijem.6.2.237>
- Metsämuuronen, J. (2020b). Somers' D as an alternative for the item–test and item–rest correlation coefficients in the educational measurement settings. *International Journal of Educational Methodology*, 6(1), 207-221. <https://doi.org/10.12973/ijem.6.1.207>
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). McGraw Hill.
- R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. <https://www.r-project.org/>
- Tarhan, M., & Yıldırım, A. (2021). Bir ölçek geliştirme çalışması: Hemşirelikte geçiş şoku ölçeği [A scale development study: Nursing Transition Shock Scale]. *University of Health Sciences Journal of Nursing*, 3(1), 7-14. <https://doi.org/10.48071/sbuhemsirelik.818123>

APPENDIX

Appendix 1. R Codes for data generation.

```
generate_data <- function(seed, discrimination, interfactor_cor, sample_size) {
  #Set the seed and generate the parameters
  set.seed(seed)
  i_1 <- rnorm(sample_size, 0, 1)
  t_1 <- rnorm(sample_size, discrimination*i_1, sqrt(1-discrimination^2))
  t_2 <- rnorm(sample_size, interfactor_cor*t_1, sqrt(1-interfactor_cor^2))
  i_2 <- rnorm(sample_size, discrimination*t_2, sqrt(1-discrimination^2))
  tidy::tibble(i_1, i_2, t_1, t_2, scale_score = t_1 + t_2)
}
```

Appendix 2. Discrimination indexes values obtained from simulation conditions.

The Number of Categories Type of Item Discrimination		Sample Size																										
		200									500									1000								
		Inter-factor Correlation																										
		.00			.30			.50			.00			.30			.50			.00			.30			.50		
		Item Discrimination																										
		.30	.50	.70	.30	.50	.70	.30	.50	.70	.30	.50	.70	.30	.50	.70	.30	.50	.70	.30	.50	.70	.30	.50	.70	.30	.50	.70
2	IFC	.24	.40	.56	.24	.40	.56	.24	.40	.56	.24	.40	.56	.24	.40	.56	.24	.40	.56	.24	.40	.56	.24	.40	.56	.24	.40	.56
	ITC	.17	.29	.39	.19	.32	.45	.21	.34	.48	.17	.28	.39	.19	.32	.45	.21	.34	.48	.17	.28	.40	.19	.32	.45	.21	.35	.48
	MD	.25	.41	.56	.26	.42	.59	.27	.45	.63	.24	.40	.56	.25	.42	.59	.27	.45	.63	.24	.40	.56	.25	.42	.58	.27	.45	.63
3	IFC	.26	.43	.60	.26	.43	.60	.26	.43	.60	.26	.43	.60	.26	.43	.60	.26	.43	.60	.26	.43	.60	.26	.43	.60	.26	.43	.60
	ITC	.18	.31	.43	.21	.35	.49	.22	.37	.52	.18	.31	.43	.21	.35	.48	.22	.37	.52	.18	.30	.42	.21	.35	.48	.22	.37	.52
	MD	.27	.44	.61	.27	.45	.63	.29	.48	.67	.26	.43	.60	.27	.45	.63	.29	.48	.67	.26	.43	.60	.27	.45	.63	.29	.48	.67
5	IFC	.28	.48	.67	.29	.48	.67	.28	.48	.67	.28	.48	.67	.28	.48	.67	.29	.48	.67	.29	.48	.67	.29	.48	.67	.29	.48	.67
	ITC	.20	.34	.47	.23	.38	.54	.25	.41	.58	.20	.34	.47	.23	.38	.54	.25	.41	.58	.20	.34	.47	.23	.39	.54	.25	.41	.58
	MD	.29	.48	.67	.31	.50	.70	.32	.53	.75	.29	.48	.67	.30	.50	.70	.32	.54	.75	.29	.48	.67	.30	.50	.70	.32	.53	.75
7	IFC	.29	.48	.68	.29	.48	.68	.29	.48	.68	.29	.49	.68	.29	.48	.68	.29	.49	.68	.29	.49	.68	.29	.48	.68	.29	.49	.68
	ITC	.21	.34	.48	.24	.39	.55	.25	.42	.59	.21	.34	.48	.23	.39	.55	.25	.42	.59	.21	.34	.48	.24	.39	.55	.25	.42	.59
	MD	.30	.49	.68	.31	.51	.71	.33	.54	.76	.29	.49	.68	.31	.51	.71	.33	.54	.76	.29	.49	.68	.31	.51	.71	.33	.54	.76

IFC: Item-Factor Correlation, ITC: Item-Total Correlation, MD: Multidimensional Discrimination

Pamukkale critical thinking skill scale: a validity and reliability study

Erdinc Duru¹, Sevgi Ozgungor¹, Ozen Yildirim^{1,*}, Asuman Duatepe-Paksu², Sibel Duru¹

¹Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye

²Pamukkale University, Faculty of Education, Department of Mathematics and Science Education, Denizli, Türkiye

ARTICLE HISTORY

Received: Apr. 04, 2022

Accepted: Sep. 01, 2022

Keywords:

Critical thinking,
Test development,
University students,
Validity and reliability.

Abstract: The aim of this study is to develop a valid and reliable measurement tool that measures critical thinking skills of university students. Pamukkale Critical Thinking Skills Scale was developed as two separate forms; multiple choice and open-ended. The validity and reliability studies of the multiple-choice form were constructed on two different theoretical frameworks as classical test theory and item-response theory. According to classical test theory, exploratory and confirmatory factor analyses were performed, to item-response theory, the Generalized Partial Credit Model (GPCM) for one-dimensional and multi-category scales was tested for the construct validity of the multiple-choice form of the scale. Analysis results supported the unidimensional structure of the scale. The reliability analyzes showed that the internal consistency coefficient of the scale and the item-total correlation values were high enough. The test-retest analysis results supported that the scale shows stability over time regarding the field it measures. The results of the item-response theory-based analysis also showed that the scale met the item-model fit assumptions. In the evaluation of the open-ended form of the scale, a rubric was used. Several studies were conducted on the validity and reliability of the open-ended form, and the results of the analysis provided psychometric support for the validity and reliability. As a result, Pamukkale Critical Thinking Skills Scale, which was developed in two forms, is a valid and reliable measurement tool to measure critical thinking skills of university students. The findings were discussed in the light of the literature and some suggestions were given.

1. INTRODUCTION

All living things on earth have genus specific biological and cognitive resources that enable them to adapt to the world (Tolman, 1932, p. 374). The main skill that distinguishes humans in terms of these resources is thinking. Thinking as a core concept for all cognitive actions of human beings includes many important sub-processes. In this sense, the literature distinguishes among different high-level thinking processes such as reflective thinking, creative thinking, and critical thinking. Reflective thinking, like critical thinking conceptualized by Dewey (1933) and

*CONTACT: Ozen Yildirim ✉ ozen19@gmail.com 📧 Pamukkale University, Faculty of Education, Department of Measurement and Assessment, Denizli, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

often used interchangeably, can be defined as the process of creating a new understanding that makes the meaning and importance of the phenomenon apparent through processing the phenomenon and related intellectual experience analytically (Boyd & Fales, 1983). In other words, reflective thinking includes the process of making judgments for controlling and improving the learning process by actively thinking about what is known, what is lacking and how the discrepancy can be eliminated (Dewey, 1933). In the most general sense, creative thinking is defined as the ability to create new products (Parkhurst, 1999) by offering solutions or perspectives that have not been offered yet. On the other hand, critical thinking, referred by Paul (2005) as thinking about thinking, is defined as the evaluation and meaning making process (Mazer et al., 2007) of identifying the main themes and assumptions behind the claims presented in light of reasons and evidence independently of the effects of current prejudice and past experiences (Paul & Elder, 2001), discovering relationships, drawing conclusions based on existing evidence and considering whether these conclusions are valid based on the evidence (Pascarella & Terezini, 1991). Although each one of the aforementioned thinking processes is important in carrying out the daily life's actions and tasks which are getting more complex, diverse and requiring multi-dimensional perspective day by day, the critical thinking is no longer just an ability with extra advantage, rather, as the base of all other thinking processes (Paul & Nosich, 1991; Ruggiero, 1990), it has become an indispensable perquisite for the adaptation to today's world where information is temporary, intense and often misleading. As a matter of fact, in today's world, defined as the information age by many researchers in recent years, there has been numerous calls for critical thinking to be an indispensable part of the educational process (Facione, 2015; Lipman, 1988; Siegel, 1988; Uzuntiryaki & Capa-Aydin, 2013) as a necessary skill needed in every aspect of life including workplace performance and leadership skills (Flores et al., 2012), crisis prevention management, which has become more important under the threat of global warming (Comfort, 2007) and ensuring the continuation and preservation of democracy through knowledge management as a citizen under the information bombardment (Rezaee et al., 2012).

In spite of the existing need and it has a deep-rooted history dating back to Socrates (Bailey & Mentz, 2015), a common conceptualization that can provide scientific understanding and consistency in the literature has only emerged as late as 1990's as a result of the Delphi project (Mpfu & Maphalala, 2017). Awareness of the importance of critical thinking increased due to the regained popularity of the 1980's educational approach emphasizing the importance of inquiry-based high-level thinking skills where students are the main actors in the education process (Facione, 1990a). As a result of this awareness, The Delphi project was initiated to create a consensus-based conceptualization that could be used in critical thinking teaching by means of a holistic definition (Facione, 1990a). Within the scope of the project, 46 experts who are known for their contributions to the field from philosophy, education, social and natural sciences formed a committee and worked for two years to determine what critical thinking is and its most important components, skills, and related behaviors. After intense exchanges among prominent figures in the field such as Dave Ellis, Richard Paul, and Peter A. Facione, the committee determined that critical thinking has two inseparable components: critical thinking skills and critical thinking dispositions. Critical thinking skills further consist of interpretation, analysis, evaluation, inference, explanation, and self-regulation sub-skills. Interpretation skills, which are defined as understanding the content and importance of the text by critically considering the material at hand, independently of their own subjective thoughts, include cognitive skills such as recognizing the problem, defining it objectively, defining the content in one's own words, and defining the author's point of view. Analyzing is to identify the inferential relationships between elements by identifying the ideas and arguments presented in the content. In this framework, analysis includes actions such as finding the source of the claims in the content, identifying the similarities and differences between different options. Evaluation

includes skills such as deciding based on the credibility of a content to determine the reliability of the judgment, belief, decision, or ideas presented in the content and the reasons presented for these ideas, whether the evidence presented sufficiently supports the conclusion reached, or whether the reason, judgment or ideas put forward are within the framework of logic, existing situation, and evidence. Inference involves reasoning and drawing conclusions by questioning presented arguments and assumptions in the light of available evidence. In this framework, creating a consistent content-based synthesis, predicting the next step, and identifying possible outcomes are among the examples of making inferences. Explanation, which is another sub-dimension of critical thinking skills, is the ability to present the content as a coherent whole by synthesizing the ideas reached in the critical thinking process and is exemplified by actions such as presenting criteria that reflect the logic behind the decisions reached, turning the content into graphics. Self-regulation skill, which consists of self-testing and correction sub-skills, includes behaviors such as examining content objectively, reviewing previous decisions and ideas, referencing objective sources to be sure, and rearranging when erroneous inferences are noticed. In other words, self-regulation is the process of regulating one's own critical thinking process by critically addressing it. In addition to these sub-skills, Paul and Elder (2002) distinguish between weak critical thinking, which includes an objective analysis of the individual's content and is characterized as external to the individual, and strong critical thinking skills, which also includes the individual's monitoring of their own cognitive processes (p.38). Paul and Elder state that individuals with strong critical thinking listen to others even when they have completely different ideas from their own, try to understand by valuing their perspectives, and are able to change their own perspectives based on others' rationality. In other words, strong critical thinking also includes creating an objective reality by listening to others and evaluating events beyond their own personal needs within the framework of all other perspectives on the situation. This type of evaluation enables individuals to fully understand the others by putting themselves in the shoes of others and thus to develop a holistic understanding including the thinking of others and the underlying logic of this thought. For this reason, it can be argued that an important last sub-dimension of critical thinking is perspective taking, although it is not included in the Delphi project. In this context, perspective taking is the ability to approach the content from the perspective of others to express ideas based on the synthesis of different perspectives (Carpendale & Lewis, 2006) and to develop an original perspective.

One of the common deductions of the experts involved in the Delphi project is that these skills can be taught by training. As a matter of fact, the literature supports this inference. Bensley et al., (2010) reported that a significant increase was observed in the critical thinking scores of the psychology program students who took the research methods course when they received training on the critical thinking process in the first three weeks of the course, but the critical thinking scores of the students who did not receive this type of training did not change. Similarly, Cisneros (2009) stated that the critical thinking scores of the pharmacy students who did not receive any explicit critical thinking training in the study did not improve throughout the year, even though they had above the average scores compared to what is reported in the literature. On the other hand, there are many studies showing the positive effects of critical thinking activities when critical thinking is clearly instructed or when these activities are presented as a natural part of the course (see Abrami et al., 2008; Huber & Kuncel, 2016; Marin & Halpern, 2011; Mpofu & Maphalala, 2017; Msila, 2014; Sahool & Mohammed, 2018; Puig et al., 2019, for more detail).

The general conclusion drawn when the findings of these studies are taken together is that critical thinking can be supported by experience, strategic information and practice (Snyder & Snyder, 2008) gained through especially a teaching process that includes questions encouraging the analysis and evaluations of the claims behind the idea and arguments within the scope of

healthy skepticism (Browne & Freeman, 2000). At the same time, it was pointed out that the university life provides valuable experiences in the development of critical thinking (Huber & Kuncel, 2016; Pascarella & Terenzini, 2005), which naturally brings together many students with different cultural and life experiences together, and where discussion and analysis is supported more than previous educational experiences.

In many of the higher education institutions in Western societies, different methods are applied to support critical thinking skills as a result of this awareness (eg, Dumitru, et al., 2018). In Turkey, interest in critical thinking has increased recently, and although most of the studies have been carried out in the field of educational sciences (Batur & Özcan, 2020), the number of studies on critical thinking skills in different fields, especially health, business and economics, continues to increase. Most of these studies aim to determine the current situation (Batur & Özcan, 2020) and evaluate the competence levels of individuals' critical thinking skills. These studies indicate that university students' critical thinking skills are at medium or low level (e.g., Doğanay et al., 2007; Özmen, 2008). Another noteworthy point is that existing studies in the literature mostly use the concepts of skill and dispositions interchangeably and ignore the conceptual nuances between the two.

The critical thinking disposition, which constitutes the curiosity and motivation necessary for an individual to think critically, expresses the tendency or willingness of the individual to critical thinking skills such as questioning, thinking of alternatives and searching for evidence (Facione, 1990a). Facione identified seven critical thinking dispositions: analyticity, truth-seeking, self-confidence, maturity, open-mindedness, systematicity, and inquisitiveness. Although critical thinking dispositions are useful in predicting critical thinking skills as an integral part of critical thinking skills (Facione, 1997), unlike skills measured based on performance, it expresses a tendency to critical thinking and is measured through self-reports based on subjective evaluations. However, as mentioned above, critical thinking skills include performance based on deep processing of content through cognitive actions such as interpretation, analysis, evaluation, inference and explanation. Therefore, these skills could be measured only through testing the participant's ability to apply these skills instead of subjective evaluations of a person's motivation to critically think.

Despite this distinction, the studies conducted in Turkey mostly use skills and tendencies synonymously, and dispositions are often tested in the evaluation of programs that claim to support critical thinking skills (e.g., Atay, Ekim, Gökkaya, & Sağım, 2009; Güçlü & Evcili, 2021; Naçalı, et al., 2016; Özmen, 2008). In his comprehensive study that analyzed the historical development of critical thinking measurement tools in Turkey, Doğan (2013) stated that the psychometric properties of the scales for measuring skills have more psychometric issues compared to those measure dispositions. He also stressed the inadequacy of measurement tools based on adaptation studies as well as the need for the national psychometrically strong scale development studies.

Despite these apparent differences, an important reason why these two terms are used interchangeably is the limitations regarding the availability of a valid and reliable scale adapted to Turkish culture to measure critical thinking skills in the adult population. A series of tests have been developed in the literature to measure critical thinking. The most widely used of these tests in the literature are the Watson-Glaser Critical Reasoning Scale (WGCTA- Watson-Glaser Critical Thinking Appraisal, Watson & Glaser, 1980), The California Critical Thinking Skills Test (Facione & Facione, 1992), Cornell Critical Thinking Test Level X and Level Z (Cornell Critical Thinking Test Level X- Level Z) (Ennis & Millman, 1985) and New Jersey Test of Reasoning Skills (Shipman, 1983). Despite the intense work in the literature and the development of many measurement tools, the debate about the validity and reliability levels of these tests continues, and the findings are that the psychometric levels of these tests are not

ideal, or the findings are inconsistent (Abrami et al., 2008). In Turkey, validity and reliability studies were conducted on only a few of these scales -Watson-Glaser Critical Reasoning Scale and the California Critical Thinking Skills Test and the Cornell Critical Thinking test, which was developed to measure the critical thinking skills of preschool children- and many studies reported psychometric properties that were far from ideal.

Ayberk and Çelik (2007) collected data from pre-service teachers and reported reliability coefficients values ranging from .10 to .35 for the subscales of Watson-Glaser Critical Reasoning Scale, where the reliability coefficient for the whole scale was only .38. They pointed out that these numbers were similar to the values of .29 and .53 obtained by Evcen and Çıkrıkçı-Demirtaşlı (2002). On the other hand, the only subscale of The California Critical Thinking Skills Test commonly used in Turkey is the one measuring dispositions. The subscale of critical thinking skills has been shown to have reasonable psychometric values in studies conducted abroad (Facione, 1990b; Facione, 1990c), and although these findings were supported across different cultures, there are also call for caution regarding the use of the scale. For example, Jacobs (1995) reported that although the reliability coefficients of the A and B forms of the scale were .56 and .59, the reliability coefficients of the subscales were as low as .14. Moreover, although the scale has been translated into Turkish, it continues to be a measurement tool with very low accessibility since it is subject to a practically an unreachable fee in Turkey's conditions and is not equally open to all researchers.

Today, although critical thinking skills are needed in all areas of life and have become a prerequisite for the healthy functioning of society, there is currently no accessible scale to measure students' critical thinking skills in university that prepare individuals for working and living conditions and hence expected to support critical thinking skills. However, the lack of an accessible scale that can measure the critical thinking skills of students in university environments where critical thinking opportunities and development potential are abundant, makes it difficult to monitor whether the required improvements are achieved as a result of the current educational experiences offered in higher education institutions. At best, the lack of a valid scale limits the research scope to making predictions about critical thinking skills through dispositions.

In the light of the literature above, the need for an economical, accessible, valid and reliable measurement tool developed in Turkish culture is evident. In this context, the main purpose of this study is to develop a valid and reliable measurement tool for measuring critical thinking skills of university students in the context of Turkish culture.

2. METHOD

2.1. Participants

In the research, the aim was to develop both multiple choice and open-ended forms of the critical thinking skill scale, and for this purpose, data were collected from students studying in the field of teaching in different age groups and different departments. The data were collected according to convenient sampling method from prospective teachers studying in the 1st, 2nd, 3rd and 4th grades of Pamukkale University Faculty of Education between the 2019-2021 academic years.

During the construction of the open-ended form, data were collected from the participants in order to develop the rubric and to determine the response distributions, and then 15 participants were asked to answer the scale again for the reliability analysis of the test.

The data for the multiple-choice form were collected from two different groups. First, data were collected from 355 participants and analyzes based on Exploratory Factor Analysis (EFA) and Item Response Theory (IRT) were conducted. 29% (103 people) of the participants are male

and 70% (251) are female. The average age of the participants is 20.75. The [Table 1](#) gives the distribution of participants by grade level.

Table 1. *The distribution of participants by grade level for EFA.*

Grade Level	Frequencies (<i>f</i>)	Percentage (%)
First	168	47.323
Second	80	22.535
Third	34	9.577
Fourth	73	20.281
total	355	100.00

The majority of the sample (47.32%) consists of 1st year prospective teachers. While the distributions of the 2nd (22.53%) and 4th grades (20.28%) are close to each other, it is seen that there are at least (9.57) 3rd year prospective teachers in the sample. The distribution of the participants according to the departments is given [Table 2](#).

Table 2. *The distribution of the participants by the departments.*

	Frequencies (<i>f</i>)	Percentage (%)
Mathematics and Science	82	23.119
Turkish and Social Studies	47	13.260
Foreign languages	94	26.478
Special education	44	12.651
Guidance and Psychological Counseling	88	24.507
Total	355	100.000

The distribution of the participants participating in the research in the fields of mathematics and science (23.12%), foreign languages (26.48%) and guidance and psychological counseling (24.51%) is close to each other. In addition, the rate of these fields is higher than the fields of Turkish and social studies (13.26%) and special education (12.65%).

The scale was applied to 156 participants for Confirmatory Factor Analysis (CFA), which is used to determine the construct validity of the multiple-choice test. 26.00% of the participants are male (40 people), 74.00% are female (116). The predominance of female students in education faculties is a reflection of the sampling in the research. The average age of the participants was calculated as 21.91. [Table 3](#) gives the distribution of the participants according to their grade levels.

Table 3. *Distribution of participants by grade level for the CFA.*

Grade level	Frequencies (<i>f</i>)	Percentage (%)
Second	56	35.897
Third	70	44.871
Fourth	30	19.232
Total	156	100.000

36% of the participants are second graders, 45% are third graders and 19% are fourth graders. The number of fourth graders in the sample is less than the second and third grades. The distribution of the participants according to their departments is given in [Table 4](#).

Table 4. *The distribution of the participants by the departments.*

	Frequencies (<i>f</i>)	Percentage (%)
Mathematics and Science	49	31.410
Foreign languages	45	28.846
Guidance and Psychological Counseling	62	39.743
Total	156	100.000

Data was collected from three different departments. The number of participants participating in mathematics and science (31%) and foreign languages (29%) is close to each other, while the number of participants participating in Guidance and Psychological Counseling (40%) is higher.

2.2. Data Collection

In the research process, the theoretical framework was decided by analyzing the literature and existing scales to determine the type of measurement tool used to measure critical thinking skills (see Doğan, 2013, for more detail). The existing scales developed abroad (Watson-Glaser Critical Reasoning Scale, California Critical Thinking Skills Test, Cornell Critical Thinking Test Level X and Level Z, New Jersey Thinking Skills Test), as well as the Critical Thinking Skills Test developed in Turkey (Eğmir & Ocak, 2016) were mostly observed to be in multiple-choice test format and in the form of independent questions. It was decided to form open-ended questions based on the text in order to evaluate the respondent's behavior at different cognitive levels, given an existing situation in the presentation of the relevant structure.

Selecting the text is a critical process in terms of guiding the further steps of the research. At the first step of the writing process of the essay, the topic was determined. The text was selected based on its relatedness to real life so that it could capture the respondents' attention, its depth and its suitability for preparing questions to tap different cognitive levels. In addition, the prior knowledge of the respondents and the difficulty of the text were taken into account, as it may affect the reader's understanding (Mullis et al., 2009). Among the different topics suggested by the researchers, *vaccines and today's reflections* were chosen as the subject. In the text, speculations based on the relationship between vaccine and autism and possible side effects of the vaccine are mentioned. It is an informative compilation text created by bringing together the information from different sources. Two Turkish language experts examined the text in terms of the criteria and grammar mentioned above, and the text was finalized by making the relevant corrections. An example of a short paragraph from the text is given below.

“While developing technology provides many conveniences in our lives, it has also brought some discussions. One of these debates is whether the vaccines made to protect our children from diseases by strengthening the immune system are associated with autism or not. In the last 20 years, cases of autism in developed countries have increased dramatically. While the probability of autism in a child born in the United States in 1992 was one in 150, this number increased to one in 68 in 2004.”

Upon construction of the text, open ended questions were written in light of the cognitive processes of critical thinking proposed in the literature (eg, Ennis, 1991; Facione, 1990a; Irani et al., 2007; Lippman, 1988; Norris & Ennis, 1990; Watson & Glaser, 1980) and therefore were decided to develop around seven cognitive processes. The cognitive processes that are considered in the preparation of the questions and their definitions are as follows.

Interpretation: Understanding and expressing the meaning or significance of a wide variety of experiences, situations, data, events, judgments, conventions, beliefs, rules, steps or criteria. Sub-skills are classification, inferring and clarifying meanings.

Analyzing: Identifying inferential relationships between phrases, questions, concepts, explanations, or different forms of expression intended to express belief, judgment, experience, reason, knowledge, or opinions. Sub-skills are examining ideas, identifying arguments, and analyzing arguments.

Evaluation: Determining the reliability of explanations or definitions or statements made about perceptions, experiences, situations, decisions, beliefs or opinions, as well as; evaluating the logical strength of inferential relationships between statements, definitions, questions, or other representations. Sub-skills are evaluation of claims and evaluation of arguments.

Inference: Identifying the elements necessary to reach a logical conclusion, forming assumptions and hypotheses correctly, considering relevant information, and revealing results obtained from statements, principles, evidence, ideas, beliefs, opinions, concepts, questions and other forms of representation. Sub-skills are questioning evidence and reasoning and drawing conclusions about alternatives.

Explanation: Presenting one's reasoning results in a convincing and coherent way means being able to look at the big picture. Sub-skills are determining conclusions, justifying the steps, and presenting the arguments.

Self-regulation: Applying the cognitive activities, the elements used in these activities, and especially the skills of analysis and evaluation from the perspectives of questioning, validation, validation or correction to one's own inferential decisions. Sub-skills are self-testing and self-correction.

Perspective taking: Bringing different perspectives together and establishing cognitive empathy. In this sense, it can be said that perspective taking is a form of cognitive empathy.

Detailed information about the content of cognitive processes is included in the handbook of the scale. In this structure, the assumption that cognitive processes progress from an easy structure to a more complex structure has been accepted.

A total of 10 questions were composed/written based on two interpretations, two analyses, one evaluation, one inference, one explanation, two self-regulations and one perspective taking, based on the criteria specified above. In order to see the clarity of the questions, a pilot application was made with a sample of 10 participants, and the participants were asked questions that they did not understand or had difficulties. When the data were examined, it was observed that the desired answers could not be obtained, especially in the perspective taking question. Later, this question was reconsidered and revised by the researchers. The questions were rearranged by taking the opinions of a total of five experts in the field of critical thinking and measurement and evaluation before the pilot implementation.

The open-ended form consists of 10 items. The scale was applied to 136 participants within the scope of the research in order to develop the Rubric used in the evaluation of the scale. Then, the answers given to each question were brought together separately and analyzed and grouped from the most correct answer to the wrong answer. The answers given were grouped between one point and five points.

Based on the data received from experts and participants during the process, it was decided to develop a multiple-choice form in order to reduce the scoring bias of the scale, to facilitate the scoring and to enable it to be answered in a shorter time, in other words, to increase its usefulness (Cohen et al., 1992; Ebel, 1972). As with open-ended questions, multiple-choice questions have options ranging from one to five points. The highest score that can be obtained from the test is 50 and the lowest score is 5. If the respondent receives zero (blank), one or two points from one of these two questions, he is deemed to have received zero points from the remaining items. Answers in the remainder of the test are not scored. In this case, the student gets zero points in total. Even if the student has not answered any question correctly, he/she can get zero points. While creating the options of the multiple-choice form, attention was paid to the followings:

- ✓ Harmony with the root in terms of grammar and meaning,
- ✓ Similar lengths of the options,
- ✓ Compatibility of the closeness of the distractors to the correct answer and the planned difficulty level of the items,
- ✓ The use of participants common mistakes in distractors (Bilican, 2021).

In addition, while preparing the test, the followings were considered in the test order;

- ✓ not to not placed the correct answers of the items in a certain pattern,
- ✓ to leave a certain gap between the items, the item root and the options,
- ✓ the suitability of the number of selected options to the level of the respondent,
- ✓ the first items are suitable for the lower level of the cognitive level and the last items are suitable for the last level of the cognitive level,
- ✓ to put a directive informing the students at the beginning of the test (Haladyna, 1997).

After the questions and options were written, five different experts working in the field of critical thinking (two critical thinking, one language, two assessment and evaluation) were asked to examine the multiple-choice test by considering the table of specification of the scale. The final version of the scale was determined according to the feedback received. In order to determine the psychometric properties of the scale, an application was made on two different study groups of 355 and 156 participants. One of the points to be considered in the scoring of the multiple-choice test and the open-ended test is that the first two interpretation skill questions in the test are criterion items. If the respondent gets zero (blank), one or two points from one of these two questions, she/he is deemed to have received zero points from the remaining items. In other words, in order to get points from the whole scale, the respondent must not score less than 2 in the first two questions. Under the leadership of Bloom, one of the most significant names in the education literature, many contemporary education researchers have conceptualized thinking skills in a spectrum ranging from basic processes such as knowledge, understanding and comprehension to thinking processes such as higher-level analysis and synthesis based on these basic processes (e.g. Anderson & Krathwohl, 2001; Crockett, 2019; Dwyer et al., 2014). The common argument of these conceptualizations is that the inferences reached by the reader who cannot grasp what the content means correctly will be wrong, and therefore, low-level comprehension skills form the basis of critical thinking skills (Dwyer et al., 2014). As a matter of fact, Williams et al. (2003) showed that a program for the development of critical thinking skills did not cause an increase in the critical thinking scores of students with low academic skills despite having the same feedback and practical experiences as other students, in other words, those who already have problems in understanding the text demonstrated the need for additional support to develop critical thinking. In this framework, the scores obtained from the other items measuring high-level skills such as analysis and evaluation, which should be formed within the scope of this basic understanding, were not calculated by the students who did not correctly answer the first two questions about the comprehension level of the text, and the scores of these students regarding their critical thinking skills were recorded as low. In such a case, it is assumed that the participants do not have the ability to answer other questions correctly and guess the answers by chance.

The rubric development process for the open-ended form of the scale was reconsidered after the development of the multiple-choice test. It was decided to give score points to the whole answer given by the student to each question and a holistic rubric was prepared. For this purpose, the answers received from 136 participants were examined and scores were graded for each level from the highest to the lowest. It was deemed appropriate to make the scoring between one point and five points. Participants who did not answer the question or answered meaninglessly were given zero points. In addition, what is expected from the respondent for each success level is written with clear descriptive statements. Using participants' responses based on these definitions, possible examples are given. Open-ended questions and the developed holistic rubric were finalized by taking the opinions of three assessment and evaluation experts. In order to determine the scoring reliability of the rubric, the open-ended test was applied to a similar sample of 15 participants. Responses were scored by five

researchers and three independent experts. In order to determine the consistency between the scores, the intraclass correlation coefficient between the five researchers and between a randomly selected expert from the research group and three independent experts was examined. A sample item and rubric are given below.

LEVEL: INTERPRETATION (Classifying, inferring and clarifying meanings)

QUESTION 1. What do you think is the best title for this text?

Score	Evaluation Criteria	Sample answers
5	Reflects the main theme (content/scope/focus of the text) of the text in the title, Explanation: The title fully reflects the relationship between vaccines and autism, which constitutes the content of the text.	<ul style="list-style-type: none"> • Relationship/link between vaccines and autism • Vaccination and autism • Discussions on the Relationship Between Vaccination and Autism
4	Although includes the main argument(s)/discussions of the text in the title, narrows the scope partially. Explanation: While examining the main focus of the text (the relationship between vaccines and autism), the title narrows it down to imply a causal relationship.	<ul style="list-style-type: none"> • Effects of vaccine on autism • Is Vaccine a Cause of Autism? • Do vaccines really cause autism?
3	Mentions only one of the main points that constitute the text content in the title (ya da mentions only one of texts content's main points in the title) Explanation: Although the main focus of the text is the relationship between shot and autism, limits the content by mentioning either only shot or autism in the title. Or even though mentions both, narrows the scope of at least one to the degree it does not reflect the text anymore.	<ul style="list-style-type: none"> • Is the vaccine our friend or not? • Autism and Its Causes • Vaccine and its importance • The relationship between triple vaccine combination and autism
2	Although the title emphasizes the focus/most important elements/main elements of the text, it narrows the scope causing significant misunderstanding. Explanation: Although it mentions vaccine and/or autism in the title, it uses expressions that cause misunderstanding in a way that cannot be excluded from the scope of the text. Sets an irrelevant title that does not reflect the scope of the text.	<ul style="list-style-type: none"> • Does the developing technology trigger autism? • Vaccine-Autism Theory or Technology and Neuropsychiatry • Autism and infectious diseases • Increase in Diseases and Vaccination • Are vaccines killing our children?
1	Explanation: It does not include any statement that will reflect the relationship between vaccine and autism, which is the main element of the text.	<ul style="list-style-type: none"> • Incorrect treatment and possible consequences • Can technology make us worse while improving it? • Severe consequences of unfair ignorance • Science and diseases

The open-ended form and the multiple-choice form were applied separately to different groups in the classroom environment under the supervision of the researchers. While it took 20-30 minutes to answer the open-ended form, it took 10-15 minutes to answer the multiple-choice form.

2.3. Data Analysis

2.3.1. Validity and reliability analysis of critical thinking multiple choice form according to Classical Test Theory

Exploratory Factor Analysis (EFA) was performed to test the construct validity of the scale and to identify the items that best revealed the construct. Principal Axis Factoring Method, one of the factor determination methods, was preferred in EFA. Before the analyses, the assumptions of the factor analysis were tested. Univariate and multivariate outliers and missing values were examined in the data collected from a total of 355 participants, and finally, analyzes were carried out with a sample of 336 participants. Since the number of missing values was low (less than 5%) (Bennett, 2001; Shaffer, 1999) no data imputation method was used and they were excluded from the sample. The correlations between the ten items in the scale are given in Table 5.

Table 5. Inter-item correlation coefficients for EFA.

Item no	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
I1	1.000**									
I2	.304**	1.000**								
I3	.496**	.281**	1.000**							
I4	.504**	.211**	.464**	1.000**						
I5	.544**	.290**	.447**	.544**	1.000**					
I6	.569**	.281**	.519**	.591**	.592**	1.000**				
I7	.627**	.307**	.577**	.578**	.630**	.705**	1.000**			
I8	.572**	.314**	.524**	.534**	.577**	.682**	.768**	1.000**		
I9	.559**	.298**	.472**	.553**	.542**	.583**	.702**	.739**	1.000**	
I10	.540**	.322**	.487**	.515**	.585**	.596**	.659**	.636**	.585**	1.000**

** $p < 0.001$

The correlation coefficients between the items vary between 0.211 and 0.768. Although the correlation of I2 with other items is observed to be somewhat low (0.211 to 0.322), there are significant correlations between the variables according to the result of the Barlett test, which tests the significance of the correlation matrix and the suitability of the data for analysis, the data set is suitable for analysis ($p < 0.01$). Finally, the KMO (Kaiser Mayer Olkin) value, which gives information about the suitability of the sample size for each variable and the whole model, was calculated as 0.947. It means the number of samples (336) used in the analysis is sufficient for the analysis.

While deciding the number of factors in EFA, it was tested with a parallel analysis in addition to the analysis results. The number of factors obtained from the factor analysis and the number of factors suggested by the additional analysis were compared in the scatter plot. The proof of reliability of the scale was calculated with the Cronbach's Alpha reliability coefficient, which gives information about internal consistency, and it was examined item discrimination, based on item-total test correlation, and the difference between the lower and upper 27% groups.

Test-retest reliability was also tested in order to obtain additional information about the reliability (in terms of stability) of the scale. For this purpose, the multiple-choice form of the critical thinking scale was applied twice to a similar sample group of 35 participants, one month apart, and the correlation between the first and second applications of the students was examined. In addition, the difference between the pretest-posttest scores of the critical thinking variable were examined using the paired sample t-test, and it was determined whether the variable changed over time. SPSS 26 was used for EFA and reliability analysis and Jamovi 2.3 program was used for parallel analysis.

To test the construct validity of the scale, confirmatory factor analysis (CFA) was performed with the data set collected from a different sample (156 participants) at the last stage. Before

the analysis, univariate and multivariate outliers were tested and two data were excluded from the analysis. In addition, the multivariate normality assumption was tested using Mardia's skewness and kurtosis coefficients. As a result of the analysis, it was seen that the data set did not meet the multivariate normality assumption ($\chi^2=509$ $p<0.001$). For this reason, Robust Maximum Likelihood (MLR) was used as the estimation method in CFA. Before the analysis, the adequacy of the correlation coefficients between the variables was examined. Table 6 shows the correlation coefficients between the items used for CFA.

Table 6. Inter-item correlation coefficients for Table CFA.

Item No	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
I1	1.000									
I2	.333**	1.000								
I3	.578**	.217**	1.000							
I4	.578**	.128	.430**	1.000						
I5	.597**	.166*	.392**	.570**	1.000					
I6	.632**	.197*	.558**	.615**	.628**	1.000				
I7	.671**	.187*	.581**	.591**	.574**	.745**	1.000			
I8	.645**	.184*	.492**	.530**	.546**	.699**	.776**	1.000		
I9	.626**	.172*	.444**	.576**	.488**	.599**	.728**	.782**	1.000	
I10	.588**	.203**	.461**	.546**	.614**	.679**	.675**	.621**	.618**	1.000

** $p<0.001$, * $p<0.05$

Except for I2 in the scale, correlations between items vary between 0.392 and 0.782. Although correlations between I2 and other items were significant at the 0.05 level, only one value ($rm2-m4=0.128$, $p>0.01$) was not significant. When the distribution of participants' answers to the I2 item was examined, it was determined that 73% (116) of 156 participants had the most correct answer, and the distribution was less than the other options. This may indicate a problem regarding the distinctiveness of the item. It was observed that the item measures the "interpretational behavior", which is an important step of critical thinking, and no problems were encountered in its writing or in the process of understanding the options. Due to the significant correlations between I2 and other variables, it was decided by the researchers that the item should remain on the scale. The decision of whether the item remains on the scale was decided according to the results of the CFA. Jamovi 2.3 program was used for CFA analysis.

2.3.2. Validity and reliability of critical thinking multiple choice form according to Item Response Theory

Measurement tools can be developed based on different theories, the validity and reliability proofs of the multiple-choice form of critical thinking based on Classical Test Theory (CTT) are given above. Traditionally, CTT is used in development tools. However, there are some limitations brought by the CTT, for example, the psychometric properties of a tool developed according to the CTT are affected by the characteristics of the individuals who answered the test. In another theory, Item Response Theory (IRT), item parameters can be evaluated independently of group characteristics and group characteristics can be evaluated independently of item sample (Hambleton & Swaminathan, 1985). For this reason, validity and reliability analyzes of the Critical Thinking Scale based on IRT were also tested. Due to the structure of the scale, parameter estimations were made using the Generalized Partial Credit Model (GPCM) for one-dimensional and multi-category scales. GPCM is a generalization of the 2-parameter logistic model (2PLM) used for items scored in two categories. For item discrimination a parameter and the difficulty b parameter is used which is one less than the number of categories. In addition, since GPCM is basically a logistic model, a value of 1.702 was used as the D scaling coefficient to approximate this model to the more mathematically complex *ogive* models. Analyzes were conducted on 336 participants. During the analysis,

catIRT tools (Aybek, 2021) and *mirt* (Chalmers, 2012) packages in R (R core team, 2022) were used in the creation of graphics. Before proceeding to the IRT analysis, the assumptions of unidimensionality, local independence and item model fit were tested. For the unidimensionality assumption, factor analytical methods were evaluated and the results of the EFA were examined, and for local independence, Yen's Q3 local independence statistic (Yen, 1993) was calculated. The critical cut-off point was accepted as 0.30 (Røe, Damsgård, Fors, & Anke, 2014). For item-model fit, RMSEA values were analyzed in the S_{χ^2} statistic.

2.3.3. Validity and reliability analysis of critical thinking open-ended form

In order to ensure the reliability of the measurement tool, text and text-based questions were applied to 15 participants who were randomly selected and had sample characteristics, and then five experts who conducted the research scored the answers of the participants to each item based on rubric. Each item in the scale is scored multiple times. In determining the reliability of scores obtained from multiple-scored measurement tools, the inter-rater reliability coefficient can be determined by the intraclass correlation coefficient (ICC), which gives consistency between raters. As the evaluation of the ICC approaches 1.00, which can be interpreted as the evaluation of the correlation coefficient, the consistency between the raters increases, while the consistency decreases as it approaches 0.00. The suggestion of Portney and Watskins (2000) was taken into account in the evaluation of the coefficient obtained. Accordingly, when the sample size is less than 30 and the number of raters is less than 3, below 0.50 indicates weak reliability, 0.5-0.75 shows moderate reliability, 0.75-0.90 implies good reliability, and above 0.90 indicates excellent reliability.

Considering that the raters in the research group were together during both the development of the scale questions and the development of the rubrics, the consistency between the scores of three independent experts in the field of critical thinking and a randomly selected expert from the research group was evaluated by looking at the intra class correlation. The intraclass correlation coefficient was examined for both the item and the total scores obtained from the scale. 15 participants' responses were re-scored for one month by an expert selected from the research group in order to determine whether there was a difference between the scoring of the rater at two different times (intra-rater reliability). The correlations between the total scores given by the rater to each participant based on the first and second measurement results were examined. SPSS 26 program was used in the analysis.

In the research, it was tried to measure the same structure according to different measurement methods with the multiple choice and open-ended tests. The correlation between the scores obtained from these two scales in the study can also be considered as evidence for validity. Both tests were administered to 11 participants at different time intervals and the correlation between the scores was checked. Due to the small number of individuals, non-parametric Spearman Brown Rank Differences correlation analysis was performed.

3. FINDINGS

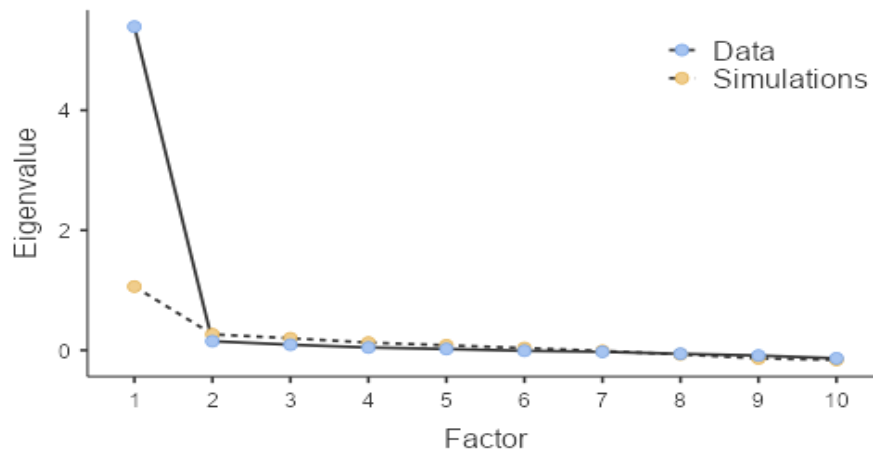
3.1. Validity and reliability findings of critical thinking multiple choice form according to Classical Test Theory

In the exploratory factor analysis, the contributions of ten items in the scale were examined and it was determined that except I2, they varied between 0.398 and 0.721. The contribution of I2 was calculated as 0.150, and the process was continued without removing the item from the analysis due to the reasons stated in the data analysis section. The explained total variance was examined to find the number of factors. Table 7 shows the explained total variance and eigenvalues.

Table 7. Explained Eigenvalues and total variance.

Factor	Total	Total %
I1	5.806	58.055
I2	.861	66.666
I3	.590	72.563
I4	.549	78.053
I5	.484	82.890
I6	.465	87.544
I7	.397	91.513
I8	.387	95.388
I9	.245	97.838
I10	.216	100.000

When [Table 7](#) was examined, it was determined that there was only one factor (5.806) with an eigenvalue greater than 1 and this factor explained 58% of the total variance. It can be said that ten items were gathered under a single factor and explained more than half of the variance. In addition, when the parallel analysis scatter plot results based on the observed and expected values are examined, it is confirmed that ten items are grouped under a single factor.

Figure 1. Parallel analysis scatter plot.

Since the items were collected under a single factor, factor rotation was not performed. Finally, factor loading values were examined. [Table 8](#) gives the factor loading values.

Table 8. The factor loading values.

Item No	Factor Loading
I1	.721
I2	.380
I3	.645
I4	.690
I5	.729
I6	.797
I7	.878
I8	.841
I9	.784
I10	.759

It is seen that the factor loads of the items are quite high (0.645-0.878). It was determined that only the factor load of I2 was 0.38, but this value was higher than the critical cut-off point of 0.30 (Tabachnick & Fidell, 2013). In addition, the RMSEA value of the model fit indices was

calculated as 0.047, which indicates a good fit (Browne & Cudek, 1993). Ten items show a single-factor structure.

After deciding on the number of factors, the reliability of the scale and the discrimination of the items were examined. The internal consistency coefficient of the scale was calculated with high reliability of 0.92. Although the number of items is small, it can be said that the scale is quite reliable in terms of internal consistency. The item-total test correlations are given in Table 9.

Table 9. Item-total correlations and reliability values.

Item No	r_{jx}	Cronbach's Alpha value when item is removed
I1	0.693**	0.910**
I2	0.367**	0.923**
I3	0.618**	0.913**
I4	0.664**	0.910**
I5	0.699**	0.908**
I6	0.762**	0.904**
I7	0.838**	0.900**
I8	0.804**	0.901**
I9	0.748**	0.905**
I10	0.728**	0.907**

** $p < 0.001$

When the item-total test correlations were examined, the lowest 0.37 and the highest 0.838 correlation values were calculated. The item-test correlation value is above 0.30, it indicates that there is a sufficient relationship between the item and the construct to be measured (Tabachnick & Fidell, 2013). Although the item was removed from the model, it was observed that there was no significant change in the Cronbach Alpha value, and the Cronbach Alpha value calculated with ten items was quite high.

In order to support the validity and reliability of the scale, item discrimination was also calculated according to the lower and upper 27% groups. Table 10 gives item discrimination according to 27% lower-upper groups.

Table 10. Independent samples *t*-test between lower-upper 27% groups.

Item No	<i>t</i> -value
I1	11.858**
I2	8.822**
I3	11.859**
I4	10.871**
I5	17.189**
I6	13.098**
I7	14.779**
I8	16.589**
I9	17.127**
I10	17.008**

** $p < 0.001$

The critical thinking scores of the participants in the lower and upper groups differ significantly for each item ($p < 0.001$). The scores of participants with high critical thinking skills can be distinguished from the scores of participants with low critical thinking skills with the scale.

To calculate the reliability of the scale as stability, the test-retest reliability was checked and a correlation of 0.52 was calculated between the first and second application. There is a moderately significant positive correlation between the two measurements ($p < 0.001$). The scores of the participants did not change between the first and second applications. In addition,

for the data obtained from these applications, the difference between the pretest-posttest scores of the critical thinking variable was not significant ($p > .001$). No change was observed in the participants' critical thinking skills during the process. Paired sample t-test results are given in Table 11.

Table 11. Paired sample t-test result.

Application	Mean	N	SD	Mean difference	SD	t	df	p
Pretest	38.085	35	5.537					
Posttest	37.314	35	4.581	0.771	5.041	0.905	34	0.372

CFA was performed to confirm the structure of the scale, which was found as a single factor. The overall goodness of fit values obtained when all items were added to the model and no modifications were made: $\chi^2/df=2.43$ (Good), SRMR= 0.039 (Good), RMSEA= 0.095 (Poor), CFI=0.095 (Acceptable), TLI=0.94 (Low). According to Browne and Cudek (1993), a RMSEA value greater than 0.08 in the model indicates poor model-data fit. In addition, CFI and TLI values higher than 0.95 indicate good fit, while values between 0.90 and 0.95 indicate acceptable fit (Bentler, 1990). When the parameter estimations were examined, the standardized regression coefficients ranged between 0.628 and 0.884, while the standardized beta coefficient of I2 was significant at the 0.05 level (Beta= 0.248, $p < 0.05$). While the variance rates explained by the items ranged between 0.39 and 0.78, I2 had the lowest explained variance ($R^2 = 0.06$). The model should be revised according to the obtained values. According to these results, I2 was removed from the model and the analysis was repeated, the overall goodness of fit values obtained: $\chi^2/df = 2.72$ (Good), SRMR= 0.037 (Good), RMSEA= 0.104 (Very Poor), CFI=0.095 (Acceptable), TLI Calculated as =0.94 (Low). It was observed that there was no change in the model fit values when I2 was added or removed from the model, and even when it was removed, the RMSEA value increased, and the model weakened.

Considering the cognitive process measured by I2, the researchers decided that I2 should remain in the scale, considering that I1 and I2 should be prerequisite items in scoring the scale, and that the prerequisite item should be measured with more than one item rather than a single item. In addition, instead of taking the average of all items in scale scoring, the validity of the presented answer was tested first. That is, although critical thinking is defined as a whole of multidimensional cognitive activities such as interpretation, understanding, and analysis (eg, Paul, 1990), the emergence of higher-level critical skills such as analysis and evaluation, and basic cognitive activities such as understanding and interpretation would not be possible without it. Therefore, in the present study, it is necessary to observe the cases where the questions I1 and I2, which measure the basic skills of understanding and interpretation, are answered incorrectly.

When the modifications are examined to increase the model fit, the error variances of I8 and I9, which measure self-regulation skills, are connected, the goodness of fit values increase ($\chi^2/df=62.8/34=1.85$ (excellent), SRMR= 0.035 (Good), RMSEA= 0.073 (Acceptable) CFI=0.097 (Good), TLI=0.96 (Good)) and model data fit was observed. Since the distribution of the answers to I8 and I9 is similar and the items measure similar cognitive levels (self-regulation), this arrangement between errors was found appropriate by the researchers. The parameter values obtained from the model are given in Table 12.

When the standardized beta coefficients giving the relationships between the items and the factor were examined, it was observed that the lowest correlation was I2 (0.253) and the highest correlation was I7 (0.880). However, most of the items have a regression coefficient above 0.60.

When looking at the variance explaining the factor, while the contribution of I7 is the highest (0.774), the contribution of I2 is the least (0.064).

Table 12. CFA model parameter estimation values.

Item No	B	SH	β	z	R ²
I1	1.000	0.000	0.787		0.618
I2	0.224	0.094	0.253	2.37*	0.064
I3	1.184	0.099	0.638	11.89**	0.406
I4	1.090	0.088	0.704	12.30**	0.495
I5	1.386	0.112	0.707	12.30**	0.500
I6	1.540	0.114	0.846	13.45**	0.715
I7	1.585	0.115	0.880	13.70**	0.774
I8	1.603	0.117	0.823	13.68**	0.677
I9	1.570	0.118	0.773	13.23**	0.596
I10	1.598	0.122	0.778	13.04**	0.606

** $p < 0.001$, * $p < 0.05$

3.2. Validity and Reliability Analysis Findings of Critical Thinking Multiple Choice Test Based on Item Response Theory

When the EFA results, which were conducted to determine the unidimensionality of the critical thinking scale, it was determined that there was only one factor with an eigenvalue greater than 1 (5,806) and this factor explained 58% of the total variance. Accordingly, it was found that ten items were gathered under a single factor and explained more than half of the variance.

Violation of local independence may affect individual parameter estimates, reliability and validity estimates of the scale (Marais, 2009; Yen 1993). For this reason, the second assumption of the IRT, local independence, was tested and it was seen that all items were below the critical cut-off point (0.30) according to the Yen's Q3 local independence test and did not violate local independence. As the last assumption, item model fit was examined, and item calibrations were made according to GPCM. The S_{χ^2} statistic for item concordance is given in Table 13.

Table 13. Item fit indices.

	S_{χ^2}	df	RMSEA
I1	28.620	27	0.013
I2	97.837	28	0.086
I3	78.936	55	0.036
I4	54.391	45	0.025
I5	75.403	47	0.042
I6	56.867	43	0.031
I7	45.416	35	0.030
I8	37.517	34	0.018
I9	62.882	41	0.040
I10	83.608	47	0.048

It was determined that the RMSEA values of nine items in the scale ranged between 0.013 and 0.048, and these items fit well with the model. The RMSEA value of I2 was calculated as 0.086. This item has low agreement with the model. A similar situation was observed in both the item discrimination and the contribution of the item to the model in the EFA and CFA analyzes, but it was decided to keep the item based on expert opinion.

After deciding on the model item fit, item parameters and standard errors of these parameters were calculated. The values of the parameters are given in [Table 14](#).

Table 14. Parameter values and standard errors of items according to GKPM.

Item no	<i>a</i>	<i>b</i> ₁ (0-1)	<i>b</i> ₂ (1-2)	<i>b</i> ₃ (2-3)	<i>b</i> ₄ (3-4)	<i>b</i> ₅ (4-5)
I1	0.827 (0.114)	NA	-3.570 (0.505)	-1.235 (0.317)	-0.259 (0.279)	-1.549 (0.339)
I2	0.347 (0.064)	NA	NA	-5.473 (1.165)	2.457 (0.897)	-6.702 (1.406)
I3	0.848 (0.124)	-0.652 (0.267)	-0.064 (0.254)	0.013 (0.240)	0.310 (0.219)	0.809 (0.228)
I4	1.010 (0.131)	0.067 (0.343)	-1.815 (0.367)	0.588 (0.163)	0.960 (0.202)	1.096 (0.248)
I5	1.153 (0.177)	-0.612 (0.247)	-0.260 (0.231)	-0.315(0.202)	0.818 (0.190)	-0.129 (0.200)
I6	0.996 (0.147)	0.080 (0.392)	-0.490 (0.385)	-1.298 (0.360)	0.200 (0.174)	0.309 (0.166)
I7	1.399 (0.221)	0.380 (0.492)	-1.365 (0.470)	-0.856 (0.263)	-0.121 (0.141)	0.657 (0.123)
I8	1.603 (0.276)	-0.439 (0.268)	-0.530 (0.251)	0.053 (0.186)	-0.509 (0.194)	0.318 (0.108)
I9	1.868 (0.374)	-0.865 (0.178)	0.420 (0.194)	0.019 (0.195)	-0.427 (0.198)	0.112 (0.120)
I10	1.002 (0.169)	-0.531 (0.284)	0.261 (0.307)	-0.325 (0.312)	-0.239 (0.249)	-0.558 (0.240)

According to [Table 14](#), it is observed that the discrimination parameters of the items (*a*) are close to 1.00. According to Baker (2001), 0.01-0.34 is considered very low, 0.35-0.64 low, 0.65-1.34 moderate, 1.35-1.69 high, and 1.70 and above very high. Item discrimination gives information about the power of the item to distinguish students according to their abilities. The higher the discrimination, the better the item can distinguish individuals according to the relevant structure. Accordingly, six items (I1, I3, I4, I5, I6, I10) have medium discrimination, 2 items (I7 and I8) have high discrimination, and one item (I9) has very high discrimination. The discrimination of I2 is low. The other predicted parameter is the “*b_i*” (option response function) parameter, which gives information about the item difficulty or the item response frequency. In GPCM, the number of option response functions is one less than the number of possible options. Since the scale was scored between 0 and 5, five alternative response functions were calculated. However, when the response pattern of I1 and I2 was examined, *b*₁ and *b*₂ of these items could not be calculated since there were no students who got zero points in I1 and no students who got zero and one points in I2. Option response parameters ranged from -6,720 to 2,457. When the *b* parameters are examined, it is seen that the *b* parameters of the items other than the 1st and 2nd items used as prerequisites include individuals with both low and high critical thinking levels. Item characteristic curves and item test information functions of ten items in the scale are given in [Figure 2](#) and [Figure 3](#).

When the item probability functions and item information functions are examined together, it is seen that I2 does not provide information for all ability levels. The item probability function of this item focused especially on two score categories. These categories are 2 (P2) and 5 (P5). Therefore, individuals below -2 ability level are more likely to get 3 points from this item, while individuals above -2 ability level are more likely to get 5 points from this item. Other score categories for this item could not be differentiated for different ability levels. On the other hand, I9 provides very high information especially for individuals between -2 and +2 skill levels.

Figure 2. Item probability functions.

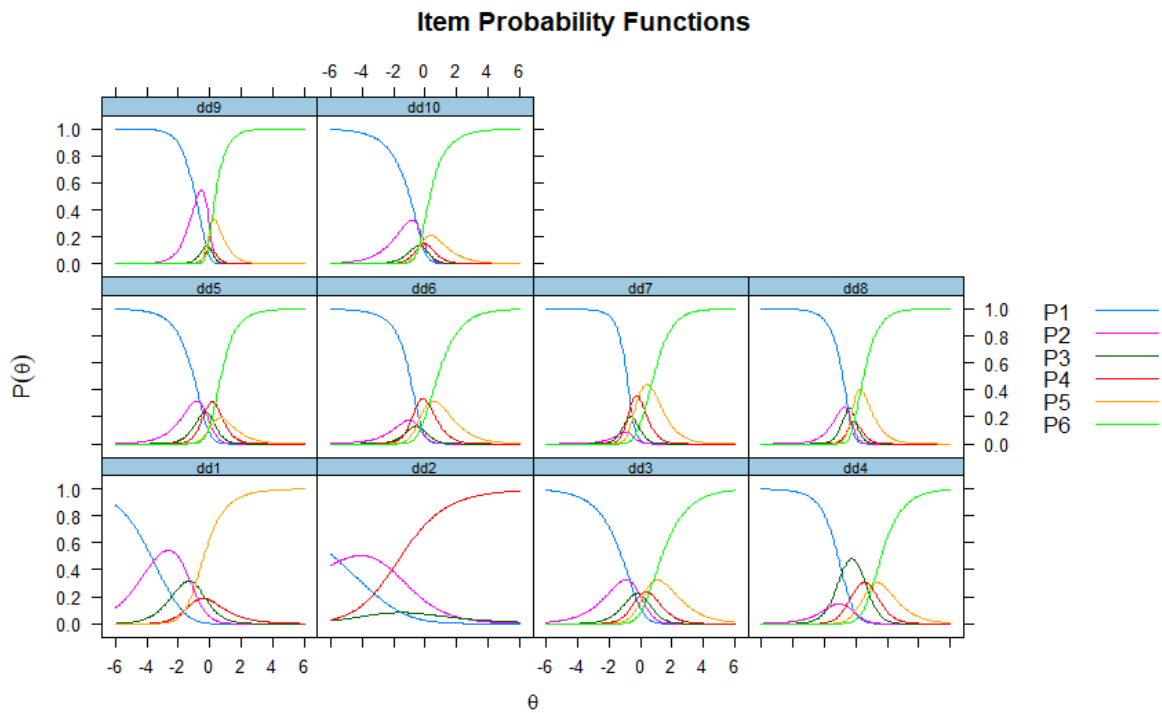
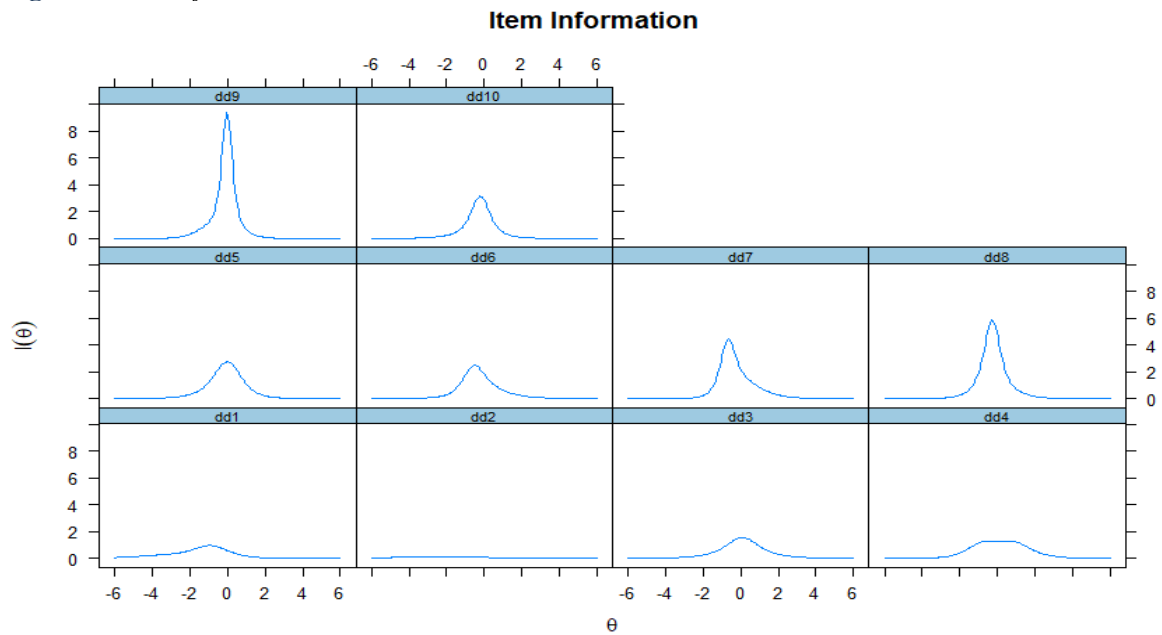
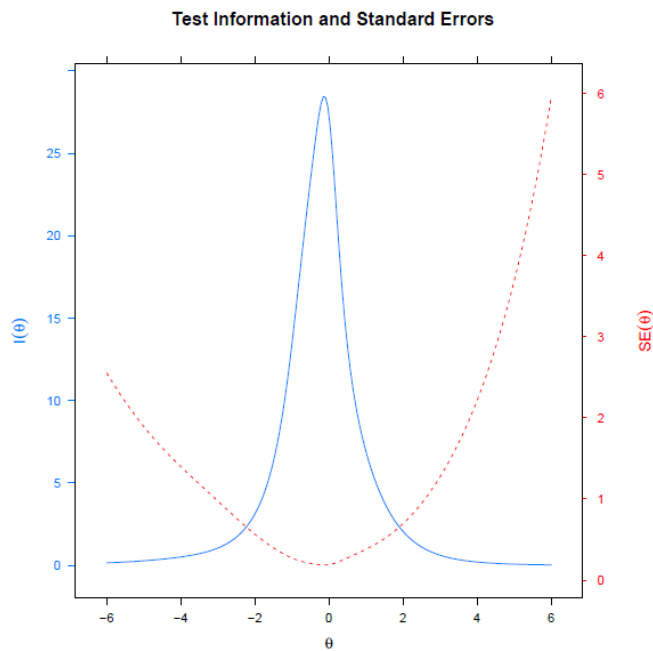


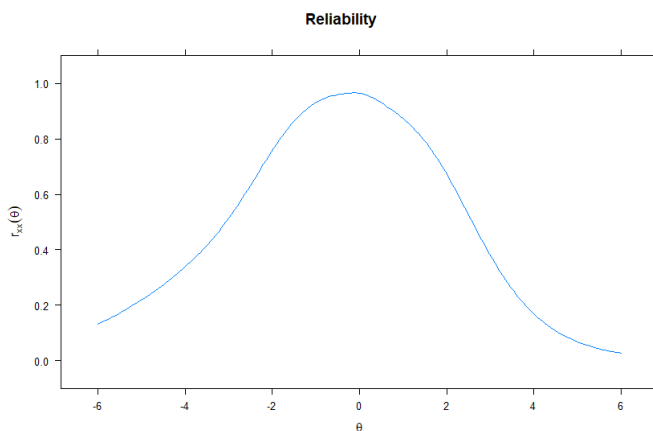
Figure 3. Item information.



The test information function was evaluated (Figure 4), it was seen that the scale provided more information for individuals whose critical thinking levels were between -2 and +2, in other words, it distinguished these individuals better.

Figure 4. Test information functions.

The empirical reliability coefficient of the scale based on IRT was calculated as .91. In addition, when the reliability function obtained for all ability levels is examined, the scale measures with a reliability above .80 especially for individuals between -2 and +2 ability levels. [Figure 5](#) shows the reliability of the test.

Figure 5. Reliability of the test.

3.3. Critical Thinking Open-Ended Form Holistic Rubric Validity and Reliability Findings

To ensure the validity steps of the holistic rubric, the steps described in detail in the data collection section were followed. Extra care has been taken to ensure inter-research coherence in arranging the definitions based on each question and possible participant responses to these definitions. In addition, the developed scoring tool was finalized by taking the opinions of experts in two critical thinking, one Turkish language and three measurement and evaluation fields and making necessary adjustments.

To determine the reliability of the scoring key, the consistency between the raters was examined. The correlation coefficients between the scores given by the five experts in the study to the answers of 15 participants are given in [Table 15](#).

Table 15. *Intraclass correlation coefficient (five experts).*

Item no	Intraclass correlation coefficient
I1	0.992**
I2	0.956**
I3	0.926**
I4	0.993**
I5	0.994**
I6	0.974**
I7	0.993**
I8	0.968**
I9	0.991**
I10	0.985**
Total score	0.966**

** $p < 0.001$

Table 15 displays that the intra-class correlation coefficients range between the raters according to the items vary between 0.956 and 0.992. It can be said that the consistency between raters is quite high. When the total scores given by the raters to 10 items were compared, it was observed that the correlation coefficient between the raters was again very high (0.966).

When the correlations between independent raters were examined in order to support the consistency between raters and to avoid bias, Table 16 correlation coefficients were obtained.

Table 16. *Intraclass correlation coefficient (four independent experts).*

Item no	Intraclass correlation coefficient
I1	0.983**
I2	0.841**
I3	0.628**
I4	0.962**
I5	0.746**
I6	0.956**
I7	0.832**
I8	0.956**
I9	0.876**
I10	0.877**
Total score	0.925**

** $p < 0.001$

The consistency between the scores given by the four experts to each item varies between 0.628 and 0.962. While the consistency between the raters was quite high for nine items, it was observed that the consistency between raters was moderate (0.628) in I3. In addition, there is a high (0.925) consistency among the raters according to the total scores. The evidence obtained reveals that reliable scoring can be done with the holistic rubric developed in scoring responses.

The correlation coefficient obtained as a result of the same rater scoring 15 participants at different times was calculated as 0.765 ($p < 0.001$). There is a positive high-significant relationship between the rater's first and second evaluations made with a one-month interval. This situation reveals that the scale also provides reliability against time.

In the evaluation of the same structure according to different measurement types, the correlation between the scores of 11 participants from multiple choice and open-ended tests varied between 0.461 and 0.658 for 5 raters. Accordingly, there was a positive moderate significant correlation

between the scores obtained from the multiple-choice test and the open-ended test ($p < 0.001$). It is expected that the correlation coefficient between the same constructs will be high, but it should be considered that the answers to the multiple-choice test are more structured, and the objectivity is strong in the evaluation, whereas the bias stemming from the expert opinion in the evaluation of the answers to the open-ended test should be considered. Therefore, this may be the reason for the possible decrease in the correlation coefficient.

4. DISCUSSION, CONCLUSION and SUGGESTIONS

The main purpose of this study is to develop a measurement tool with high validity and reliability that measures critical thinking skills of university students. In this context, a series of studies were conducted on different participant groups. The results of the analysis support that the psychometric properties of the developed scale are acceptable and can be used to evaluate critical thinking skills of university students.

In accordance with the scale development procedures, first of all, the literature was searched and the existing scales in the literature were examined (Ennis & Millman, 1985; Facione & Facione, 1992; Shipman, 1983; Watson & Glaser, 1980). The existing scales are mostly in multiple choice test format and in the form of independent questions/ the literature mainly includes scales in multiple choice test format and in the form of independent questions. In addition, these scales do not have strong psychometric properties or that the findings are not supported by different research results (Abrami et al., 2008). A similar situation seems to be valid for a few scales adapted for use in our country. In these studies, psychometric properties that are far from expected regarding the adapted scales were reported (Ayberk & Çelik, 2007). Therefore, in order to capture the critical thinking potential; It is thought that it is necessary to use performance-based evaluation rather than self-reports, critical thinking consists of related abstract cognitive structures, and it is more appropriate to conduct a holistic evaluation by evaluating the cognitive level reactions of the participants while presenting a case to identify these structures. For this purpose, it was decided to develop two separate forms of the Pamukkale Critical Thinking Skills Scale, which is structured based on the selected text; multiple choice format and open-ended format. The validity and reliability studies of the multiple-choice critical thinking scale are based on two different theoretical frameworks: classical test theory and item-response theory. In the evaluation of the open-ended form of the scale, the developed "rubric" was used.

According to classical test theory, a series of analyzes were conducted to test the construct validity of the multiple-choice form of the critical thinking scale. Whether the partial correlations between the items and the correlation matrix were suitable for factor analysis were examined with the Kaiser-Meyer-Olkin (KMO) coefficient and the Barlett test (Fayers & Machin, 1995). The analyzes showed that the KMO value was high and the Barlett test result was significant. In order to determine the construct validity of the scale, factors with an eigenvalue above 1.00 according to Kaiser normalization were taken as the criteria. The findings showed that the items were collected on a single factor of 5,806 eigenvalues, which constituted 58% of the total variance. Considering that the variance explained in social sciences should be at least 40% and above (Stevens, 1992), the results of the analysis seem significant. When the items that make up the scale were analyzed in terms of factor loads, it was observed that the factor loads of the items ranged from .38 to .84. The fact that the factor loads are above .30 is considered important in terms of showing the high representativeness power of the items in the scale. Similarly, the break point of the graph supports that the breakout occurred after the first factor.

After the exploratory factor analysis, Confirmatory Factor Analysis (CFA) was performed on a different study group to confirm the single-factor structure of the scale. When all items were added to the model and no modifications were made, some of the general goodness of fit values

were higher than expected ($\chi^2/df=2.43$, SRMR= .039, RMSEA= .95, CFI=.95, TLI=.94). For example, according to Browne and Cudek (1993), a RMSEA value greater than .08 in the model indicates poor model-data fit. When the parameter estimations were examined, the standardized regression coefficients ranged from .62 to .88, while the beta coefficient of I2 was low (.25) but significant at the .05 level. All items except for I2 have a beta coefficient over .60. According to the obtained values, the model should be revised, and some modifications should be made. According to these results, I2 was removed from the model and the analysis was repeated, but it was observed that there was no change in the model fit values when I2 was added or removed from the model and even RMSEA value increased, and the model weakened when I2 was removed. Considering the cognitive feature measured by I2, it was decided to keep I2 in the scale. In addition, some modifications were made to increase model compatibility. It was observed that when the error variances of I8 and I9 were connected, the goodness of fit values increased ($\chi^2/df=62.8/34=1.85$, SRMR= 0.035, RMSEA= 0.073, CFI=0.097, TLI=0.96) and the model data fit increased. It can be said that this arrangement between error variances is appropriate since the distribution of the answers given by the students to I8 and I9 is similar, and the items measure similar cognitive characteristics (self-regulation). In summary, the results of EFA and CFA analysis support the one-dimensional structure of the scale.

Related to the reliability studies of the scale, the internal consistency was calculated with the Cronbach's Alpha reliability coefficient, the item-total test correlations were examined and the level of discrimination between the upper and lower groups of the items was examined, and the test-retest method was used to test the measurement stability. In the data analyzes, the internal consistency coefficient of the scale was calculated as .92. This result shows that the similarity of the items and the consistency of the responses to the items are high.

When the item-total correlations and correlation matrix of the scale were examined, it was observed that the correlation values ranged between .37 and .84. If the item-test correlation value is above .30, it indicates that there is a sufficient relationship between the item and the construct to be measured (Tabachnick & Fidell, 2013). According to these results, it can be said that the items of the scale are positively and significantly related to each other and the whole scale.

In order to support the validity and reliability of the scale, item discriminations were also calculated according to the lower and upper 27 % groups, and it was observed that the critical thinking scores of the participants in the lower and upper groups differed significantly for each item. According to these results, it can be said that the scale can significantly distinguish the scores of participants with high critical thinking skills from the scores of participants with low critical thinking skills.

The test-retest method was used to test the measurement stability. For this purpose, the scale was administered to the participants with an interval of three weeks, and the Pearson Correlation Coefficient between the two applications was found to be significant at the level of .52. The results of the analysis also showed that there was no significant change in the scores of the participants between the first and second applications. These results indicate that the scale shows stability over time regarding the behavioral domain it measures. In other words, no significant change was observed in participants' critical thinking skills over time.

Measurement tools can be developed based on different theories, the validity and reliability evidence of the multiple-choice form of critical thinking based on Classical Test Theory (CTT) are given above. In addition, validity and reliability analyzes of the Critical Thinking Scale based on IRT were also conducted. Before proceeding to the IRT analysis, the assumptions of unidimensionality, local independence and item model fit were tested. Considering the one-dimensional assumption of the theory, EFA and CFA results support that Pamukkale Critical Thinking Skills Scale is one-dimensional. In addition, local independence, the second

assumption of the ITC, was tested and it was seen that according to the Yen's Q3 local independence test, the critical cut-off point for all items was below .30 and did not violate local independence. As the final assumption, item model fit was examined, and item calibrations were examined. According to the results of the analysis, it was observed that the RMSEA values of most of the items in the scale ranged between .013 and .048. Only the RMSEA of I2 was slightly higher than expected (.086).

After deciding on the model item fit, the item parameters and the standard errors of these parameters were calculated, and it was observed that the discrimination parameters (a) of the items were close to 1.00. Accordingly, six items (I1, I3, I4, I5, I6, I10) have medium discrimination, 2 items (I7 and I8) have high discrimination, and one item (I9) has very high discrimination. The discrimination of I2 is low. Since the scale was scored between 0 and 5, five alternative response functions were calculated. When the b parameters were examined, it was seen that the b parameters of the items other than the 1st and 2nd items used as prerequisites in scoring included both individuals with low and high critical thinking levels.

When the alternative response functions and item information functions are examined together, it can be said that the scale provides more information for individuals whose critical thinking levels are between -2 and +2, in other words, it distinguishes these individuals. The reliability coefficient of the scale based on IRT was calculated as .91. In addition, when the reliability function obtained for all skill levels is examined, the scale measures with a reliability above .80 especially for individuals between -2 and +2 skill levels. As a result, IRT-based analysis results of the scale; unidimensionality, local independence, and item-model fit assumptions.

In addition, a number of studies were conducted on the validity and reliability of the open-ended form of the Pamukkale Critical Thinking Skills Scale, and the detailed steps given in the data collection section were followed in order to develop the rubric. In order to determine the reliability of the scoring key, the consistency between the raters was examined. According to the analysis results, the inter-class correlation coefficients between raters ranged from .95 to .99. When the total scores given by the raters to the 10 items were compared, it was observed that the correlation coefficient between the raters was again quite high (.97). These results can be considered as an indication that the rubric is well structured and therefore the consistency between raters is high. Correlations between independent raters were also examined to control for the possibility of inter-rater bias. For this, the evaluations of four experts were used. The consistency between the scores given by the four experts to each item varies between .62 and .96. In addition, it was observed that there was a very high (.92) consistency between the raters according to the total scores. The evidence obtained reveals that reliable scoring can be done with the holistic rubric developed in scoring participant responses. In addition, the correlation coefficient obtained as a result of the same rater scoring 15 participants at different times was calculated as .76. There is a highly significant positive correlation between the rater's first and second evaluations made one-month apart. This situation can be evaluated as an indication that the scale has reliability over time.

Finally, it was examined whether two separate scale forms developed to measure critical thinking skills could make similar evaluations. The results of the analysis showed that the correlation between multiple choice and open-ended tests scores of 11 participants was .46 and .65. According to these results, it can be said that there is a moderate positive correlation between the scores obtained from the multiple-choice test and the open-ended test. On the other hand, considering that the answers to the multiple-choice test are structured, the objectivity is strong in the evaluation, and that there may be some limitations in the evaluation of the answers to the open-ended test due to expert opinion, this result seems significant.

In summary, the main purpose of this study was to develop a valid and reliable measurement tool that measures critical thinking skills in university students. The results of the analysis

provided psychometric support that the measurement tool developed in two forms is valid and reliable and can be used to measure critical thinking skills of university students. Considering the limited number of measurement tools that measure critical thinking skills based on performance, it can be said that the study contributes to the literature (Abrami et al., 2008; Facione & Facione, 1992; Ennis & Millman, 1985; Shipman, 1983; Watson & Glaser, 1980). In addition, the study contributes to the literature in terms of conceptual perspective as well as scale forms. A new conceptual dimension called "Taking Perspective" was added to the existing critical thinking dimensions and this was supported by the findings of the research. As a meaningful component of critical thinking, perspective taking requires the individual to be able to both connect with the person, text, situation, or theme and stay objective by keeping a distance from them. In addition, different from the Delphi project, the operational measurement of the "Self-Regulation" skill and its inclusion as a basic component in the content of the developed scale can be considered as another contribution to the literature. Therefore, the conceptual framework of the study can form the basis for the structuring of educational programs in the processes of developing and teaching critical thinking, which is conceptualized as an important 21st century skill (Duru et al., 2020; Trilling & Fadel, 2009; Van Laar et al., 2019; Voogt & Roblin, 2012).

Similarly, the development of the Pamukkale Critical Thinking Skills Scale in two separate forms, open-ended and multiple-choice, is another important contribution to the literature. While the open-ended form allows to evaluate critical thinking skill in a holistic and performance-based manner, the use of multiple-choice form, free from chance factor, seems to be advantageous in terms of practical, economic, accessible and time, besides holistic evaluation. Therefore, it can be expected that the scale will help field experts and educators both in understanding the level of critical thinking skills of university students and in evaluating the contribution of curriculum and practices to the development of critical thinking skills. In addition, critical thinking is one of the higher-order thinking skills, and individuals with this potential can be considered qualified human resources. Therefore, the conceptual framework related to scale can contribute to policymakers in determining qualified human resources and creating, developing, and planning education policies for this resource. Finally, it can be said that the two most important features that distinguish the Pamukkale Critical Thinking Scale (PCTS) from similar scales in the literature are that it measures critical thinking skills on a performance-based way with a text and can evaluate the individual as a whole in terms of critical thinking skills.

In the light of the above explanations, the findings of this study should be evaluated within the framework of some limitations. First, in this study, the psychometric properties of the Pamukkale Critical Thinking Skills Scale were tested on the students of the Faculty of Education. Therefore, examining the psychometric properties of the scales on different study groups and in different universities may contribute to the validity and reliability of the scale and the generalizability of the findings. Secondly, the fact that the text created within the scope of the study is related to the field of social sciences may have increased the bias in the measurement. For this reason, repeating the measurement on a different text related to the quantitative field in which tables and graphics are used can serve the purpose of testing the conceptual framework used in developing the scale. Third, in this study, predictive and discriminant validity studies of Pamukkale Critical Thinking Skills Scale were not conducted. New studies to be carried out in this context may contribute to the strengthening of the psychometric properties of the scale. Fourth, the structure of the scale was not tested in groups with different characteristics in this study. Future studies, with new research on the measurement invariance of the scale; It can serve the purpose of testing the structure of the scale in groups with different characteristics such as gender, socio-economic level, verbal-numerical domain.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Pamukkale University/Social and Humanities Ethics Committee, 93803232-622.02.

Authorship Contribution Statement

Erdinc Duru: Conceptualization, research design, supervision, fundings, interpretation, literature review, writing, critical review. **Sevgi Ozungor:** Conceptualization, research design, supervision, fundings, data collection and processing, interpretation, literature review, writing, critical review **Ozen Yildirim:** Research design, supervision, fundings, data collection and processing, data analysis and interpretation, literature review, writing, critical review **Asuman Duatepe-Paksu:** Research design, supervision, fundings, data collection and processing, critical review **Sibel Duru:** Research design, supervision, fundings, data collection and processing, critical review.

Orcid

Erdinc Duru  <https://orcid.org/0000-0001-7027-4937>

Sevgi Ozungor  <https://orcid.org/0000-0003-4954-1572>

Ozen Yildirim  <https://orcid.org/0000-0003-2098-285X>

Asuman Duatepe-Paksu  <https://orcid.org/0000-0003-2504-6294>

Sibel Duru  <https://orcid.org/0000-0002-8152-8610>

REFERENCES

- Abrami, P.C., Bernard, R.M., Borokhovski, E., Wade, A., Surkes, M. ., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, 78(4), 1102-1134.
- Anderson, L.W., & Krathwohl, D.R. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives: Complete Edition*. Longman.
- Atay, S., Ekim, E., Gökçaya, S., & Sağım, E. (2009). Sağlık Yüksekokulu öğrencilerinin eleştirel düşünme düzeyleri. [Critical thinking tendencies of Health School students] *Sağlık Bilimleri Fakültesi Hemşirelik Dergisi*, 39-46.
- Ayberk, B., & Çelik, M. (2007). Watson-Glaser Eleştirel, Akıl Yürütme Gücü Ölçeği'nin (W-GEAYGÖ) üniversite ikinci, üçüncü ve dördüncü sınıf İngilizce bölümü öğretmen adayları üzerindeki güvenlik çalışması. [Reliability study related to the power of Watson-Glaser critical thinking appraisal scale on university second, third and fourth-grade English department teacher candidates] *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 16 (1), 101-112
- Aybek, E.C. (2021). CatIRT tools: A “Shiny” application for item response theory calibration and computerized adaptive testing simulation. *Journal of Applied Testing Technology*, 22(1), 23-27.
- Baker, F.B. (2001). The basics of item response theory. College Park, ERIC, Clearinghouse on Assessment and Evaluation.
- Batur, Z., & Özcan, H.Z. (2020). Eleştirel düşünme üzerine yazılan lisansüstü tezlerinin bibliyometrik analizi. [Bibliometric analysis of graduate theses written on critical thinking] *Uluslararası Türkçe Edebiyat Kültür Eğitim Dergisi*, 9(2), 834-854.
- Bailey, R., & Mentz, E. (2015). IT teachers' experience of teaching-learning strategies to promote critical thinking. *Issues in Informing Science and Information Technology*, 12(1), 141-152.

- Bensley, D.A., Crowe, D.S., Bernhardt, P., Buckner, C., & Allman, A.L. (2010). Teaching and assessing critical thinking skills for argument analysis in psychology. *Teaching of Psychology*, 37(2), 91–96. <https://doi.org/10.1080/00986281003626656>
- Bennett, D.A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25(5), 464–469.
- Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238–246. <https://doi.org/10.1037/0033-2909.107.2.238>
- Bilican, S.D. (2021). Başarı testlerinin geliştirilmesi ve madde yazımı [Development of achievement tests and item writing] Yıldırım, Ö. ve Kartal, S.K. (Ed.), *Eğitimde Ölçme ve Değerlendirme [Measurement and Evaluation in Education]* (125-163) 1. Baskı, Lisans Yayıncılık.
- Browne, N., & Freeman, K. (2000). Distinguishing features of critical thinking classrooms. *Teaching in Higher Education*. 5(3), 301-309. <https://doi.org/10.1080/713699143>
- Boyd, E.M., & Fales, A.W. (1983). Reflective learning: Key to learning from experience. *Journal of Humanistic Psychology*, 23(2), 99-117. <https://doi.org/10.1177/0022167883232011>
- Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. K. A. Bollen and J.S. Long (Ed.), *Testing structural equation models* (pp. 136-162). Sage.
- Carpendale, J.I., & Lewis, C. (2006). *How children develop social understanding*, Blackwell.
- 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>
- Cisneros R .M. (2009). Assessment of critical thinking in pharmacy students. *American Journal of Pharmaceutical Education*, 73(4), 66. <https://doi.org/10.5688/aj730466>
- Cohen, R.J., Swerdlik, M.E., Smith, D.K., & Cohen, R.J. (1992). *Psychological testing and assessment: An introduction to tests and measurement*. Mayfield Pub. Co.
- Comfort, L.K. (2007). Crisis management in hindsight: cognition, communication, coordination, and control. *Public Administration Review*, 67(1), 189–197.
- Crockett, L. (2019). *Future-focused Learning: 10 essential shifts of everyday practice*. Solution Tree Press.
- Dewey, J. (1933). *How we think*. DC Herman.
- Doğan, N. (2013). Eleştirel düşünmenin ölçülmesi [Measuring the Critical Thinking]. *Cito Eğitim: Kuram ve Uygulama*, 22(1), 29-42.
- Doğanay, A., Akbulut-Taş, M., & Erden, Ş. (2007). Üniversite öğrencilerinin bir güncel tartışmalı konu bağlamında eleştirel düşünme becerilerinin değerlendirilmesi. [Assessing university students' critical thinking skills in the context of a current controversial issues]. *Kuram ve Uygulamada Eğitim Yönetimi*, 52(1), 511-546.
- Dumitru, D., Bîgu, D., Elen, J., Jiang, L., Railiene, A., Penkauskiene, D., Papathanasiou, I.V., Tsaras, K., Fradelos, E.C., Ahern, A.K., McNally, C., O'Sullivan, J., Verburch, A.P., Jarošová, E., Lorencová, H., Poce, A., Agrusti, F., Re, M.R., Puig, B., Blanco, P., Mosquera, I., Crujeiras-Pérez, B., Dominguez, C., Cruz, G., Silva, H., & Morais, M.D., Nascimento, M.M., & Payan-Carreira, R. (2018). *A European review on Critical Thinking educational practices in Higher Education Institutions*. <http://hdl.handle.net/10197/9865>
- Duru, E., Duatepe-Paksu, A., Balkıs, M., Duru, S., & Bakay, E. (2020). Examination of 21st century competencies and skills of graduates from the perspective of sector representatives and academicians. *Journal of Qualitative Research in Education*, 8(4), 1059-1079. <https://doi.org/10.14689/issn.2148-2624.8c.4s.1m>
- Dwyer, C.P., Hogan, M.J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity*, 12(1), 43-52. <https://doi.org/10.1016/j.ts.c.2013.12.004>

- Ebel, R.L. (1972). *Essentials of educational measurement*. Prentice-Hall.
- Eğmir, E., & Ocak, G. (2016). Eleştirel düşünme becerisini ölçmeye yönelik bir başarı testi geliştirme. [Developing an achievement test towards evaluating critical thinking skill]. *Turkish Studies*, 11(19), 337-360.
- Ennis, R. (1991). Critical thinking: A streamlined conception. *Teaching Philosophy*, 14(1), 15-24. <http://dx.doi.org/10.5840/teachphil19911412>
- Ennis, R. H., & Millman, J. (1985). *Cornell Critical Thinking Test (Level X)*. Critical Thinking Press & Software.
- Facione, P.A. (1990a). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. executive summary: "The Delphi Report"*. The California Academic Press.
- Facione, P.A. (1990b). *The California Critical Thinking Skills Test-College Level Technical Report 1: Experimental Validation and Content Validity*. The California Academic Press.
- Facione, P.A. (1990c). *The California Critical Thinking Skills Test-College Level Technical Report 2: Factors Predictive of Critical Thinking Skills*. The California Academic Press.
- Facione, N.C. (1997) *Critical thinking assessment in nursing education programs An aggregate data analysis*. The California Academic Press.
- Facione, P.A. (2015). Critical thinking: What it is and why it counts. http://www.student.uwa.edu.au/_data/assets/pdf_file/0003/1922502/Critical-Thinking-What-it-is-and-why-it-counts.pdf
- Facione, P.A., & Facione N.C. (1992). *The california critical thinking dispositions inventory*. The California Academic Press.
- Fayers, P., & Machin, D. (1995). Factor analysis for assessing validity. *Quality of Life Research*, 4(5), 424.
- Flores, K., Matkin, G.S., Burbach, M.E., Quinn, C.E., & Harding, H. (2012). Deficient critical thinking skills among college graduates: implications for leadership. *Educational Philosophy and Theory*, 44(2), 212-230. <https://doi.org/10.1111/j.1469-5812.2010.00672.x>
- Güçlü, G., & Evcili, F. (2021). Sağlık hizmetleri meslek yüksekokulu öğrencilerinin eleştirel düşünme yetileri ve boyun eğici davranış eğilimlerinin incelenmesi. [Investigation of critical thinking qualifications and submissive behavior tendency of health services vocational school students]. *Turkish Journal of Science and Health*. 2(1), 31-39.
- Haladyna, T.M. (1997). *Writing test item to evaluate higher order thinking*. Allyn & Bacon.
- Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory*. Kluwer-Nijhoff Publishing.
- Huber, C.R., & Kuncel, N.R. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research*, 86(2), 431-468. <https://doi.org/10.3102/0034654315605917>
- Jacobs, S.S. (1995). Technical characteristics and some correlates of the California Critical Thinking Skills Test, Forms A and B. *Research in Higher Education*, 36(1), 89-108. <https://doi.org/10.1007/BF0220776>
- Irani, T., Rudd, R., Gallo, M., Ricke, J., Friedel, C., & Rhoades, E. (2007) *Critical thinking instrumentation manual*. University of Florida.
- Lipman, M. (1988) Critical thinking—What can it be? *Educational Leadership*, 46(1), 38–43.
- Marais, I. (2009). Response dependence and the measurement of change. *Journal of Applied Measurement*, 10(1), 17-29.
- Marin, L., & Halpern, D. (2011). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity*, 6(1), 1–13.

- Mazer, J.P., Hunt, S.K., & Kuznekoff, J.H. (2007). Revising general education: Assessing a critical thinking instructional model in the basic communication course. *The Journal of General Education* 56(3), 173-199. <https://doi.org/10.1353/jge.0.0000>
- Mpofu, N., & Maphalala, M.C. (2017). Fostering critical thinking in initial teacher education curriculums: A comprehensive literature review. *Gender and Behaviour*, 15(2), 9342–9351.
- Msila, V. (2014). Critical Thinking in open and distance learning programmes: Lessons from the University of South Africa's NPDE Programme. *Journal of Social Sciences*, 38(1), 33–42
- Nalçacı, A., Meral, E., & Şahin, İ.F. (2016). Sosyal bilgiler öğretmen adaylarının eleştirel düşünme ile medya okuryazarlıkları arasındaki ilişki [Correlation between critical thinking and media literacy of social sciences pre-service teachers]. *Doğu Coğrafya Dergisi*, 21(36), 1-12. <https://doi.org/10.17295/dcd.99051>
- Norris, S.P., & Ennis, R.H. (1990). *The practitioners' guide to teaching thinking series. Evaluating Critical Thinking*. Hawker Bronlow Education.
- Özmen, K.S. (2008). İngilizce öğretmeni eğitiminde eleştirel düşünce: Bir vaka çalışması. [Critical Thinking in English teacher education: A case study]. *Ekev Akademi Dergisi*, 12(36), 253-266.
- Orhan, A., & Çeviker-Ay, Ş. (2022). Developing the critical thinking skill test for high school students: A validity and reliability study. *International Journal of Psychology and Educational Studies*, 9(1), 132-144. <https://dx.doi.org/10.52380/ijpes.2022.9.1.561>
- Parkhurst, H.B. (1999). Confusion, lack of consensus, and the definition of creativity as a construct. *Journal of Creative Behavior*, 33(1), 1–21.
- Paul, R. (2005) The state of critical thinking today, *New Directions for Community Colleges*, 130(1), 27–38.
- Paul, R., & Elder, L. (2001) Critical thinking: Inert information, activated ignorance, and activated knowledge, *Journal of Developmental Education*, 25(2), 36–37.
- Paul, R.W., & Elder, L. (2002). *Critical thinking: Tools for taking charge of your professional and personal life*. Pearson Education Inc.
- Paul, R., & Nosich, G. (1991). *A proposal for the national assessment of higher-order thinking*. Paper commissioned by the U.S. Department of Education Office of Educational Research and Improving National Center for Education Statistics.
- Pascarella, E.T., & Terenzini, P.T. (1991). *How college affects students: Findings and insights from twenty years of research*. Jossey-Bass.
- Pascarella, E.T., & Terenzini, P.T. (2005). *How college affects students: A third decade of research*. Jossey-Bass.
- Portney, L.G., & Watkins, M.P. (2000) *Foundations of clinical research: Applications to practice*. 2nd Edition, Prentice Hall.
- Puig, B., Blanco-Anaya, P., Bargiela, I.M., & Crujeiras-Pérez, B. (2019). A systematic review on critical thinking intervention studies in higher education across professional fields. *Studies in Higher Education*, 44(5), 860-869.
- R Core Team (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.
- Rezaee, M., Farahian, M., & Ahmadi, A. (2012). Critical thinking in higher education: Unfulfilled expectations. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 3(2), 64-73.
- Ruggiero V.R. (1990) *Beyond feelings. A guide to critical thinking*, (3rd ed.) Mayfield Publishing.
- Røe, C., Damsgård, E., Fors, T., & Anke, A. (2014). Psychometric properties of the pain stages of change questionnaire as evaluated by Rasch analysis in patients with chronic

- musculoskeletal pain. *BMC Musculoskelet Disord*, 15(1), 95. <https://dx.doi.org/10.1186/1471-2474-15-95>, PubMed 24646065
- Sahool, S., & Mohammed, C.A. (2018). Fostering critical thinking and collaborative learning skills among medical students through a research protocol writing activity in the curriculum. *Korean J Med Educ*, 30(2), 109-118. <https://doi.org/10.3946/kjme.2018.86>
- Schafer, J.L. (1999). Multiple imputation: a primer. *Stat Methods in Med.*, 8(1), 3–15. <https://doi.org/10.1191/096228099671525676>
- Shipman, V. (1983). *New Jersey test of reasoning skills*. IAPC, Test Division, Montclair State College.
- Siegel, H. (1988). *Educating for reason: Rationality, critical thinking, and education*. Routledge.
- Snyder, L.G., & Snyder, M.J. (2008). Teaching critical thinking and problem solving skills. *The Journal of Research in Business Education*, 50 (2), 90–99.
- Stevens, J. (1992). *Applied multivariate statistics for the social sciences*. Second Edition, Lawrence Erlbaum Associates.
- Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics* (6th ed.), Allyn and Bacon.
- Tolman, E.C. (1932). *Purposive behavior in animals and men*. Century/Random House UK.
- Trilling, B., & Fadel, C. (2009). *21st-century skills: Learning for life in our times*. John Wiley & Sons.
- Uzuntiryaki-Kondakçı, E., & Çapa-Aydın, Y. (2013). Predicting critical thinking skills of university students through metacognitive self-regulation skills and chemistry self-efficacy. *Educational Sciences: Theory & Practice*, 13(1), 666-670.
- Van Laar, E., van Deursen, A.J.A. M., van Dijk, J.A.G.M., & de Haan, J. (2019). Determinants of 21st-century digital skills: A large-scale survey among working professionals. *Computers in Human Behavior*, 100, 93–104. <https://doi.org/10.1016/j.chb.2019.06.017>
- Voogt, J., & Roblin, N.P. (2012). A comparative analysis of international frameworks for 21st-century competencies: implications for national curriculum policies. *Journal of Curriculum Studies*, 44(3), 299–321.
- Watson, G., & Glaser, E.M. (1980). *Watson-Glaser critical thinking appraisal*. Psychological Corporation
- Williams, R.L., Oliver, R., Allin, J.L., Winn, B., & Booher, C.S. (2003). Psychological critical thinking as a course predictor and outcome variable. *Teaching of Psychology*, 30(3), 220–223. https://doi.org/10.1207/S15328023TOP3003_04
- Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(1), 187-213.

APPENDIX

Sample Multiple Choice Questions and Some Options

1. Aşağıdakilerden hangisi metnin amacını **en iyi** açıklar?

Metnin amacı,

- Aşıların otizme neden olup olmadığını göstermektir.
- Aşı ile otizm arasındaki ilişkiye dair bazı tartışmaları sunmaktır.
- Otizmin nedenlerini yapılan araştırmaları karşılaştırarak açıklamaktır.

2. Hangi seçenekte metinden çıkarılabilecek gerekçelerin tamamı birlikte verilmiştir?

- Son 20 yılda gelişen teknolojiyle birlikte otizm vakalarının artması- Aşı tartışmaları sonucunda ailelerin çocuklarına aşı yaptırmaması- Aşıların içindeki cıvanın otizme neden olması- Fazla miktarda balık tüketimi olması
- Otizimli 12 çocukla yapılan araştırma sonuçları- Gelişen teknolojinin insan sağlığını tehdit etmesi- Cıva içeren balıkların tüketiminin nörolojik hastalıklara neden olması- Otizm ile çocuklardaki alüminyum oranı arasındaki ilişki
- Otizimli çocukların çoğunluğunun aşılı olması- Otizm ve aşı arasındaki ilişkilerin araştırma sonuçlarına dayanması- Son yıllarda otizmin artması- Sayısal verilerle otizm ile aşı arasındaki ilişkinin desteklenmesi

3. Aşağıdakilerden hangisinde bebeklerine aşı yaptırmama konusunda kararsız kalan ebeveynlere, metinden çıkarılacak gerekçelere dayalı **en uygun** öneri verilmiştir?

- Araştırma sonuçlarından çıkarılacağı gibi aşı yaptırmamalarını önerirdim. Çünkü aşı olmayan birçok insan günümüzde sağlıklı bir şekilde hayatlarına devam edebilmektedir.
- Farklı kaynaklardan araştırmalarını ve uzmanlara sormalarını önerirdim. Çünkü aşı yaptırırlarsa otizm olma, yaptırmazlarsa bulaşıcı hastalıklara yakalanma olasılığı söz konusudur.
- Farklı kaynaklardan araştırıp, uzmanlara danışmalarını, sonucunda aşı yaptırmalarını önerirdim. Çünkü aşı yapılmadığı takdirde bulaşıcı hastalıklarda artış gözlenmiştir.

4. Aşı yaptırmayı savunan bir çocuk doktorunun bu metni okuduktan sonraki düşüncelerini aşağıdakilerden hangisi **en iyi** yansıtır?

- Aşılar gereklidir. Ancak aşıların olası yan etkileri ve ebeveynlerin kaygıları dikkate alındığında başka araştırmaların da incelenmesi önemlidir.
- Aşı önemlidir, aşı olmayan çocukların bulaşıcı hastalıklara karşı bağışıklıkları düşük olduğundan, bebeklere küçük yaştan itibaren aşı yapılmalıdır.
- Çocukların daha sağlıklı büyüebilmesi için bazı aşılar zamanında yapılmalıdır ve en kısa sürede tekli aşı sistemine geçilmelidir.

A novel approach for calculating the item discrimination for Likert type of scales

Umit Celen^{1,*}, Eren Can Aybek²

¹Amasya University, Faculty of Education, Department of Educational Sciences, Amasya, Türkiye

²Pamukkale University, Faculty of Education, Department of Educational Sciences, Denizli, Türkiye

ARTICLE HISTORY

Received: May. 21, 2022

Accepted: Sep. 04, 2022

Keywords:

Item discrimination index,
Likert-type scale,
Exploratory factor
analysis,
Slope coefficient,
Monte-Carlo simulation.

Abstract: Item analysis is performed by developers as an integral part of the scale development process. Thus, items are excluded from the scale depending on the item analysis prior to the factor analysis. Existing item discrimination indices are calculated based on correlation, yet items with different response patterns are likely to have a similar item discrimination index. This study proposed a new item discrimination index that can be used in Likert type of scales and examined its effect on factor analysis results. For this purpose, simulative datasets were generated, and items were excluded from the analysis according to the .20, .30 and .35 item discrimination index criteria, and exploratory factor analysis was performed for a single factor. Accordingly, it was found that more variance could be explained by a single factor with fewer items compared to other discrimination indices when the .20 criterion of the slope coefficient was used as suggested in this study. Similar findings were obtained using the .35 criterion with other discrimination indices. In this context, it is recommended to use the slope coefficient as an additional discrimination index calculation method in the scale development process.

1. INTRODUCTION

Although validity and reliability are the features related to the scores obtained with the measurement tool, and not the measurement tool itself, the qualities of the measurement tool affect the validity and reliability of the scores obtained with that instrument. Qualities such as having items that include the characteristic to be measured and being prepared in accordance with the guidelines of item writing can be examined with the help of expert opinion. On the other hand, statistical methods are used to measure the difficulty levels of the items or whether they can distinguish among the individuals who more or less have the characteristic to be measured. According to the classical test theory, the correct answer rate of an item by the group constitutes the item difficulty while a wide variety of statistics have been developed to detect item discrimination. Long and Sandiford reported that 23 different methods were defined to calculate the item discrimination index even in 1935 (as cited in Oosterhof, 1976). Kelley (1939) presented that among these methods, the most appropriate findings could be obtained in

*CONTACT: Umit Celen ✉ umit.celen@amasya.edu.tr 📍 Amasya University, Faculty of Education, Department of Educational Sciences, Amasya, Türkiye

e-ISSN: 2148-7456 /© IJATE 2022

the method based on the comparison of the lower and upper groups when the group sizes were 27% and Johnson (1951), on the other hand, made corrections in this formula and suggested the formula for the lower - upper 27% groups method used today. In addition, methods based on the correlation between the item and the total score are also frequently used to determine item discrimination. However, this way of calculating the correlation also gave rise to different methods.

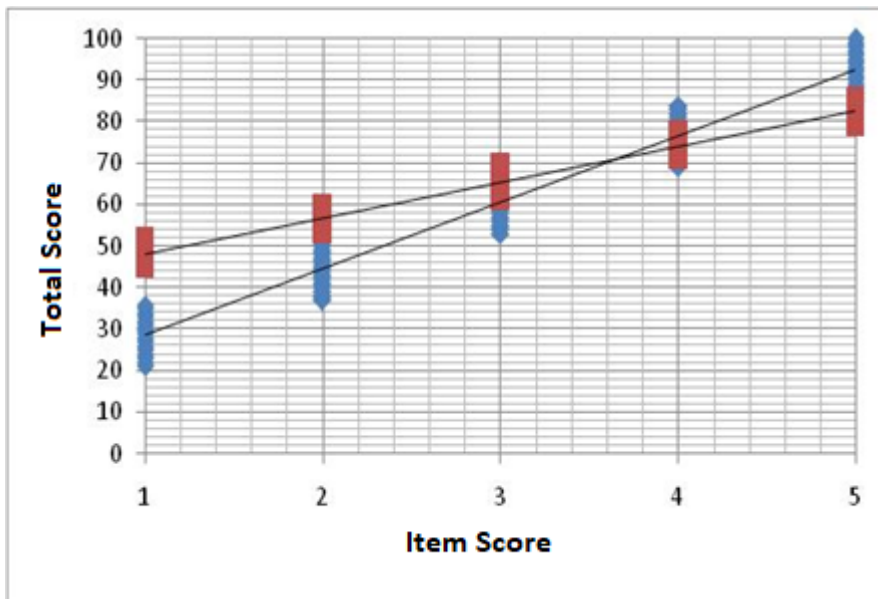
For dichotomous items, the item score has two categories, while the total score is a continuous variable. For this reason, biserial or point biserial correlation coefficients are used to calculate the correlation between a two-category discontinuous variable and a continuous variable (Popham, 2014). The main difference between these two correlation coefficients is that the point biserial correlation coefficient assume that the variable (item score) is true categorical in nature, whereas the biserial correlation coefficient assumes that the categorical variable actually has a continuous nature but has been artificially made discontinuous (Crocker & Algina, 2008). On the other hand, Guilford (1965) suggested using the point biserial correlation coefficient as it provides more information about the contribution of each item to the predictive validity of the test. Henrysson (1971) stated that the biserial correlation coefficient can be used if the total score is normally distributed (as cited in Oosterhof, 1976).

Although these methods suggested for item discrimination index were first generated for items scored 1-0, they are also used for items scored in polytomous categories (e.g. Likert type). However, in this case, the item-total correlation is calculated with the Pearson product moments correlation coefficient instead of the biserial or point biserial correlation coefficient and since there are no correct and incorrect answers, the upper-lower groups method is calculated by taking into account the difference between the item score averages of the upper and lower groups. However, since the total score obtained from the test also includes the item score whose discrimination is to be calculated, the correlation coefficient calculated between the item and the total score gives an overestimate of discrimination. For this reason, the correlation coefficient between the item and the total score obtained from the other items in the test (item-rest correlation) is also used in the calculation of item discrimination. On the other hand, polyserial correlation also could be used instead of Pearson correlation, if one assumes item score as ordinal and total score as continuous (Moses, 2017).

Either by using a correlation-based or group comparison-based item discrimination coefficient, the main purpose of an item discrimination index is to show whether individuals more or less exhibiting the measured trait also respond to the item in a similar way. However, it is stated (Livingston & Dorans, 2004) that a graphical method should also be followed in the examination of item discrimination. In this context, it is seen that when the items with the same item discrimination index are examined graphically, they distinguish individuals differently. [Figure 1](#) provides a sample graphic. As [Figure 1](#) shows, the two items indicated by blue and red dots distinguish individuals differently. However, the correlation coefficient between the item and the total score for both items is obtained as .98. This case reveals the necessity of considering different methods in conjunction when item discrimination index is calculated.

On the other hand, the test development process includes performing exploratory factor analysis to obtain proof of construct validity after the implementation of the draft items and investigating the common structure under which the items are joined (Tabachnick & Fidell, 2013). Before exploratory factor analysis, item discrimination is performed to exclude the items with low discrimination from the analysis at the very beginning.

Figure 1. Response plot for two different items.



1.1. Current Study

This study proposed a new item discrimination index to be used alongside the existing item discrimination indices. Equation 1 provides this coefficient, called the slope coefficient.

$$\text{Slope Coefficient} = \frac{\bar{X}_n - \bar{X}_m}{n - m} \quad (1)$$

In this equation, \bar{X}_n is the total mean score of the participants choosing the highest category; \bar{X}_m is the total mean score of the participants choosing the lowest category; k is the number of items in the scale; n represents the point value of the highest category and m represents the point value of the lowest category. When calculating the average score of individuals, some researchers reduce the individual's score to the response category range by dividing the total scores by the number of items, instead of the average of the total score obtained from the scale. In this situation, for example, the individual's score from the scale is obtained in the range of 1-5, for a 5-point Likert scale. In this case, there is no need to use the k value in the equation.

The present study aimed to examine the effect of *slope coefficient* (sc) on the total variance explained by the exploratory factor analysis and in this context, to compare the performance of sc with other item discrimination indices.

2. METHOD

2.1. Data

The research data were generated in R (R Core Team, 2022) using the *genPolyMatrix* function of the *catR v3.16* (Magis & Raiche, 2012) package. The *catR* package generates data based on Item Response Theory (IRT). Although this research was carried out according to the Classical Test Theory (CTT), this package was preferred because of its convenience in producing the item pool and response pattern.

Both the item pool and the sample size were controlled during data generation. Accordingly, the item pool size was 10, 30, 50 and 100, respectively and sample size was assigned as 50, 100, 250, 500 and 1000, respectively. Hence, a total of 20 different response patterns were generated simulatively, in four different item pools and five different samples. In addition, the number of replications was determined to be 100 and each simulation was repeated 100 times. The response category was chosen as 5. Since the item parameters were generated according to

IRT with *catR*, the item discrimination indices of the items would be quite high. To prevent this, the item discrimination a parameters of 30% of the generated items were corrected. For this, parameter a was randomly assigned from a normal distribution with a mean of 0.3 and a standard deviation of 0.05. Then, for each item pool, the response pattern with the specified sample size was generated using the *genPattern* function. As a result, a total of 2000 data files with a total of 20 conditions and 100 replications with different item pools and sample sizes were examined in the study.

2.2. Data Analysis

Simulative data were examined in the study and five different coefficients were calculated for item discrimination: Polyserial correlation coefficient, item-total correlation coefficient, item-rest correlation coefficients upper-lower 27% groups method and slope coefficient. The polyserial correlation coefficients were calculated by using the *polyserial* function in the *psych* v2.2.5 (Revelle, 2022) package and the item discrimination indices found via item-total, item-rest, and upper-lower 27% groups method were calculated by using the *ItemAnalysis* function in the *ShinyItemAnalysis* v1.4.1 (Martinkova & Drabinova, 2018) package. The slope coefficient, constituting the essence of the research, was calculated by transferring Equation-1 to R.

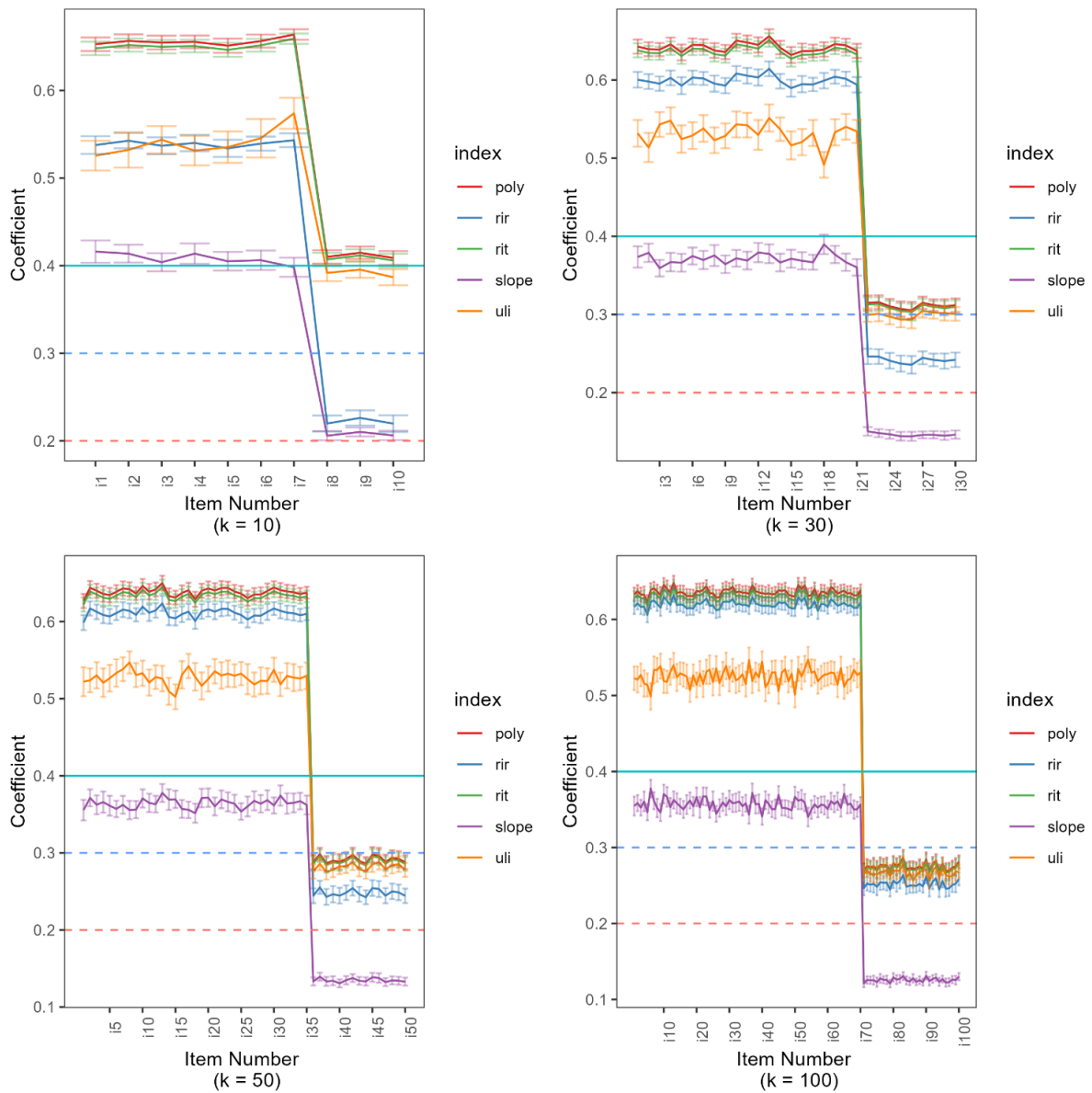
The relevant coefficients were separately calculated for each replication and then the item discrimination index averages and 95% confidence interval values for each item were visualized with the help of the *dplyr* v1.0.9 (Wickham et.al., 2022) and *ggplot2* v3.3.6 (Wickham, 2016) packages. Then, .20, .30 and .35 values were accepted as criteria, respectively, and the items below the criteria were excluded from the data set and factor analysis was performed with the remaining items. The number of items included in the factor analysis and the variance rates explained by these items in a single factor were reported by calculating the mean and 95% confidence intervals. The R script used in data generation and analysis can be accessed via <https://www.github.com/anonym> [The full URL will be provided if the manuscript is approved. URL is hidden for the purpose of anonymity].

3. RESULTS

The item discrimination indices were the first findings obtained as a result of data analysis. For ease of interpretation, the calculated item discrimination indices for each item pool and sample size were plotted, but in order not to disrupt the flow of the text, only the plot for the sample size of 1000 was provided in Figure 2. Appendix lists the plots obtained for all sample sizes. The mean value of 100 replications for each item was taken as the basis for obtaining the graphs, but the 95% confidence intervals were also presented in the graph.

Accordingly, for each item pool size, item discrimination values were found to decrease in the last 30% of the item pool. While creating the simulation conditions, the last 30% of each item pool was manipulated and deliberately reduced. Therefore, this was an expected result. On the other hand, the slope coefficient for each item pool size generated significantly lower values than the other item discrimination indices, except when the item pool size was 10. This can be seen from the fact that the confidence intervals did not intersect. In addition, when the number of items exceeds 50, the item discrimination indices obtained by the item-total, item-rest, and upper-lower methods were quite similar to each other; polyserial correlation coefficient was significantly lower and the slope coefficient provided the lowest value among the five item discrimination indices.

Figure 2. Item discrimination indices for $n = 1000$ in different item pool sizes.



poly: polyserial correlation; *rir*: item-rest correlation; *rit*: item-total correlation; *uli*: upper-lower 27% groups; *slope*: slope coefficient.

However, item discrimination indices provided similar results for items specifically when the number of items was 100, in the last 30% of the item pool and expected to be non-discriminatory, excluding the slope coefficient. Another finding that emerged after examining the graphics in Figure-2 and Annex-1 showed that the slope coefficient generally generated lower values; on the other hand, the remaining four item discrimination indices and items expected to have poor discrimination values were above .20. Then, using .20, .30 and .35 criteria, respectively, items with an item discrimination index below this criterion were excluded from the test, and exploratory factor analysis was performed with the remaining items. Tables 1, 2, and 3 provide the mean number of remaining items, the mean of variance explained by a single factor, and the 95% confidence intervals for both values.

Table 1. The number of items included in the EFA and the mean ratio of variance explained by a single factor for item discrimination index criterion .20

n	Method	k = 10		k = 30		k = 50		k = 100	
		k_r	Var	k_r	Var	k_r	Var	k_r	Var
		[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]
50	SC	8.72 [8.53-8.92]	.36 [.35-.37]	22.84 [22.45-23.23]	.34 [.33-.34]	36.89 [36.32-37.46]	.40 [.40-.41]	72.56 [71.65-73.47]	.40 [.40-.41]
	Poly	9.90 [9.84-9.96]	.32 [.31-.33]	28.32 [28.08-28.56]	.33 [.32-.34]	46.77 [46.42-46.12]	.34 [.33-.35]	91.47 [90.94-92.00]	.34 [.34-.35]
	ULI	9.68 [9.57-9.79]	.33 [.32-.34]	27.67 [27.04-27.94]	.34 [.33-.34]	45.75 [45.40-46.10]	.34 [.34-.35]	89.73 [89.12-90.34]	.35 [.34-.35]
	RIT	9.90 [9.84-9.96]	.32 [.31-.33]	28.26 [28.02-28.50]	.33 [.32-.34]	46.61 [46.26-46.96]	.34 [.33-.35]	91.28 [90.75-91.82]	.34 [.34-.35]
	RIR	8.80 [8.61-8.99]	.36 [.35-.37]	26.71 [26.40-27.02]	.35 [.34-.36]	45.12 [44.69-45.55]	.35 [.34-.36]	89.72 [89.13-90.31]	.35 [.34-.35]
	SC	8.78 [8.60-8.96]	.35 [.34-.36]	22.35 [22.09-22.62]	.39 [.39-.40]	36.81 [36.46-37.17]	.40 [.39-.40]	71.80 [71.25-72.35]	.40 [.40-.41]
100	Poly	9.97 [9.94-10.0]	.31 [.30-.32]	28.99 [28.84-29.14]	.32 [.31-.32]	47.42 [47.09-47.76]	.33 [.32-.33]	93.55 [93.21-94.11]	.33 [.32-.33]
	ULI	9.95 [9.91-9.99]	.31 [.30-.32]	28.39 [28.19-28.59]	.32 [.32-.33]	46.80 [46.48-47.13]	.33 [.32-.33]	92.26 [91.76-92.76]	.33 [.33-.34]
	RIT	9.97 [9.94-10.0]	.31 [.30-.32]	28.96 [28.80-29.12]	.32 [.31-.32]	47.31 [46.98-47.64]	.33 [.32-.33]	93.35 [92.89-93.81]	.33 [.33-.33]
	RIR	8.86 [8.69-9.03]	.31 [.30-.32]	27.18 [26.90-27.46]	.34 [.33-.34]	45.59 [45.20-45.98]	.34 [.33-.34]	91.28 [90.74-91.83]	.34 [.33-.34]
	SC	8.84 [8.66-9.02]	.34 [.33-.34]	21.72 [21.54-21.90]	.40 [.39-.40]	35.56 [35.39-35.73]	.40 [.39-.40]	70.79 [70.56-71.03]	.40 [.40-.40]
	Poly	10.0 [10.0-10.0]	.30 [.29-.31]	29.52 [29.31-29.65]	.31 [.30-.31]	48.55 [48.31-48.79]	.31 [.31-.32]	95.44 [95.05-95.83]	.32 [.31-.32]
250	ULI	9.99 [9.97-10.0]	.30 [.29-.31]	29.15 [28.97-29.33]	.31 [.31-.32]	47.84 [47.57-48.11]	.32 [.31-.32]	94.09 [93.60-94.58]	.32 [.32-.32]
	RIT	10.0 [10.0-10.0]	.30 [.29-.31]	29.50 [29.37-29.63]	.31 [.30-.31]	48.43 [48.19-48.67]	.31 [.31-.32]	95.23 [94.82-95.65]	.32 [.31-.32]
	RIR	9.00 [8.84-9.16]	.33 [.33-.34]	27.56 [27.28-27.84]	.33 [.32-.33]	46.23 [45.90-46.56]	.33 [.32-.33]	92.77 [92.21-93.33]	.32 [.32-.33]
	SC	8.90 [8.73-9.07]	.33 [.33-.34]	21.44 [21.31-21.58]	.40 [.39-.40]	35.30 [35.19-35.41]	.40 [.40-.40]	70.25 [70.12-70.39]	.40 [.40-.40]
	Poly	10.0 [10.0-10.0]	.30 [.30-.30]	29.85 [29.77-29.93]	.30 [.30-.31]	49.23 [49.06-49.40]	.31 [.30-.31]	97.00 [96.65-97.35]	.31 [.31-31]
	ULI	9.99 [9.97-10.0]	.30 [.30-.30]	29.65 [29.53-29.77]	.31 [.30-.31]	48.76 [48.57-48.95]	.31 [.31-.31]	96.05 [95.64-96.46]	.31 [.31-31]
500	RIT	10.0 [10.0-10.0]	.30 [.30-.30]	29.83 [29.75-29.91]	.30 [.30-.31]	49.18 [49.00-49.36]	.31 [.30-.31]	96.86 [96.51-97.21]	.31 [.31-.31]
	RIR	9.11 [8.96-9.26]	.33 [.32-.33]	28.03 [27.81-28.26]	.32 [.32-.32]	46.88 [46.55-47.21]	.32 [.32-.32]	94.54 [94.08-95.00]	.32 [.31-.32]
	SC	8.76 [8.59-8.93]	.34 [.33-.34]	21.30 [21.18-21.42]	.40 [.39-.40]	35.29 [35.18-35.40]	.40 [.39-.40]	70.18 [70.10-70.26]	.40 [.40-.40]
	Poly	10.0 [10.0-10.0]	.30 [.29-.30]	29.88 [29.81-29.95]	.30 [.30-.30]	49.56 [49.43-49.69]	.30 [.30-.31]	98.01 [97.73-98.29]	.31 [.30-.31]
	ULI	10.0 [10.0-10.0]	.30 [.29-.30]	29.77 [29.68-29.86]	.30 [.30-.31]	49.31 [49.17-49.47]	.30 [.30-.31]	97.46 [97.14-97.79]	.31 [.30-.31]
	RIT	10.0 [10.0-10.0]	.30 [.29-.30]	29.88 [29.81-29.95]	.30 [.30-.30]	49.53 [49.40-49.66]	.30 [.30-.31]	97.91 [97.62-98.20]	.31 [.30-.31]
1000	RIR	9.06 [8.90-9.22]	.33 [.33-.34]	28.30 [28.04-28.56]	.32 [.31-.32]	47.41 [47.13-47.69]	.31 [.31-.32]	95.50 [95.03-95.97]	.31 [.31-.31]

SC: Slope coefficient; Poly: Polyserial correlation; ULI: Upper-Lower index; RIT: Item-total correlation; RIR: Item-rest correlation.

When the criterion for the item discrimination index was set to .20, it was observed that a similar number of items remained in the test with the help of the slope coefficient for 10 items and the item-rest correlation and a similar variance was explained by a single factor. Other item discrimination indices left more items in the test but explained a lower rate of variance. However, in cases where the number of items in the pool exceeded 30 and the sample size was 100 or more, the slope coefficient excluded more items than the others, and accordingly, a higher rate of variance could be explained by a single factor. For example, when the sample

size was 1000 and the item pool size was 100; an average of 70.18 [70.10-70.39] items were included in the factor analysis with the slope coefficient, while an average of 95.50 [95.03-95.97] items were included in the factor analysis according to the item-rest correlation.

Table 2. The number of items included in the EFA and the mean ratio of variance explained by a single factor for item discrimination index criterion .30

n	Method	k = 10		k = 30		k = 50		k = 100	
		k _r [95% CI]	Var [95% CI]	k _r [95% CI]	Var [95% CI]	k _r [95% CI]	Var [95% CI]	k _r [95% CI]	Var [95% CI]
50	SC	6.73 [6.47-6.99]	.42 [.41-.43]	16.63 [15.91-17.35]	.44 [.43-.44]	27.19 [26.11-28.26]	.45 [.44-.45]	52.66 [50.50-54.83]	.45 [.44-.45]
	Poly	9.47 [9.34-9.60]	.34 [.33-.34]	26.11 [25.82-26.40]	.35 [.35-.36]	42.85 [42.40-43.30]	.36 [.36-.37]	83.57 [82.89-84.25]	.37 [.36-.38]
	ULI	8.88 [8.68-9.08]	.35 [.34-.36]	24.46 [24.10-24.82]	.37 [.36-.37]	40.42 [39.89-40.99]	.38 [.37-.38]	79.70 [78.89-80.51]	.38 [.37-.38]
	RIT	9.43 [9.30-9.56]	.34 [.33-.35]	25.97 [25.67-26.27]	.36 [.35-.36]	42.60 [42.16-43.04]	.37 [.36-.37]	83.03 [82.37-83.69]	.37 [.37-.38]
	RIR	7.78 [7.60-7.96]	.39 [.38-.40]	24.33 [24.02-24.64]	.37 [.37-.38]	40.52 [40.05-40.99]	.38 [.37-.39]	81.19 [82.37-83.69]	.38 [.37-.38]
	100	SC	6.77 [6.61-6.93]	.41 [.40-.42]	17.99 [17.45-18.53]	.42 [.42-.43]	28.23 [27.26-29.20]	.43 [.43-.43]	55.32 [53.48-57.16]
Poly	9.81 [9.72-9.90]	.31 [.31-.32]	26.21 [25.93-26.49]	.35 [.34-.35]	42.64 [42.25-43.03]	.36 [.35-.36]	82.55 [81.99-83.11]	.37 [.36-.37]	
ULI	9.33 [9.17-9.49]	.33 [.32-.34]	25.31 [25.04-25.58]	.36 [.35-.36]	41.42 [41.00-41.84]	.36 [.36-.37]	80.25 [79.54-80.96]	.37 [.36-.37]	
RIT	9.75 [9.65-9.85]	.32 [.31-.32]	26.09 [25.80-26.38]	.35 [.34-.36]	42.44 [42.05-42.83]	.36 [.35-.36]	82.10 [81.55-82.65]	.37 [.36-.37]	
RIR	7.62 [7.47-7.77]	.39 [.38-.40]	23.66 [23.39-23.93]	.38 [.37-.38]	40.23 [39.80-40.66]	.37 [.37-.38]	80.13 [79.59-80.67]	.37 [.37-.38]	
250	SC	6.83 [6.72-6.94]	.40 [.39-.41]	18.47 [18.07-18.87]	.41 [.41-.42]	29.23 [28.52-29.94]	.42 [.41-.42]	56.35 [55.02-57.68]	.42 [.42-.42]
	Poly	9.93 [9.88-9.98]	.30 [.30-.31]	26.17 [25.85-26.49]	.34 [.34-.35]	42.07 [41.69-42.45]	.35 [.35-.36]	80.98 [80.34-81.62]	.36 [.36-.37]
	ULI	9.72 [9.62-9.82]	.31 [.30-.31]	25.23 [24.92-25.54]	.35 [.35-.36]	40.84 [40.39-41.29]	.36 [.35-.36]	79.06 [78.45-79.67]	.37 [.36-.37]
	RIT	9.92 [9.87-9.92]	.30 [.30-.31]	26.00 [25.69-26.31]	.34 [.34-.35]	41.78 [41.38-42.18]	.35 [.35-.36]	80.55 [79.92-81.18]	.36 [.36-.37]
	RIR	7.43 [7.30-7.56]	.38 [.38-.39]	22.86 [22.60-23.12]	.38 [.38-.39]	38.76 [38.42-39.10]	.38 [.37-.38]	77.51 [76.96-78.06]	.37 [.37-.38]
	500	SC	6.92 [6.85-6.99]	.40 [.40-.41]	18.97 [18.64-19.30]	.41 [.41-.41]	29.64 [29.10-30.18]	.41 [.41-.42]	57.64 [56.61-58.67]
Poly	9.97 [9.94-10.0]	.30 [.30-.30]	26.46 [26.18-26.74]	.34 [.33-.34]	41.75 [41.40-42.10]	.35 [.35-.35]	80.39 [79.80-80.98]	.36 [.36-.36]	
ULI	9.85 [9.77-9.93]	.30 [.30-.31]	25.64 [25.34-25.94]	.34 [.34-.35]	40.43 [40.03-40.83]	.36 [.36-.36]	78.15 [77.57-78.73]	.37 [.37-.37]	
RIT	9.95 [9.91-9.99]	.30 [.30-.31]	26.34 [26.05-26.63]	.34 [.33-.34]	41.44 [41.07-41.81]	.35 [.35-.36]	80.03 [79.46-80.67]	.36 [.36-.37]	
RIR	7.20 [7.12-7.28]	.30 [.30-.30]	22.57 [22.30-22.84]	.38 [.38-.39]	37.73 [37.43-38.03]	.38 [.38-.38]	76.34 [75.79-76.89]	.38 [.37-.38]	
1000	SC	6.96 [6.91-7.01]	.40 [.39-.40]	18.90 [18.60-19.20]	.41 [.41-.41]	29.70 [29.13-30.27]	.41 [.41-.41]	57.85 [57.01-58.69]	.41 [.41-.41]
	Poly	9.99 [9.97-10.0]	.30 [.29-.30]	26.38 [26.09-26.67]	.34 [.33-.34]	41.29 [40.90-41.68]	.35 [.35-.36]	79.20 [78.64-79.76]	.36 [.36-.37]
	ULI	9.86 [9.79-9.93]	.30 [.30-.30]	25.24 [24.91-25.57]	.35 [.34-.35]	40.21 [39.84-40.58]	.36 [.36-.36]	77.02 [76.49-77.55]	.37 [.37-.37]
	RIT	9.99 [9.97-10.0]	.30 [.29-.30]	26.16 [25.86-26.46]	.34 [.33-.34]	41.08 [40.70-41.46]	.35 [.35-.36]	78.62 [78.09-79.15]	.37 [.36-.37]
	RIR	7.11 [7.05-7.17]	.39 [.39-.40]	22.02 [21.83-22.21]	.39 [.38-.39]	37.39 [37.10-37.68]	.38 [.38-.38]	75.14 [74.68-75.60]	.38 [.38-.38]

SC: Slope coefficient; Poly: Polyserial correlation; ULI: Upper-Lower index; RIT: Item-total correlation; RIR: Item-rest correlation.

The rate of variance explained by a single factor was 40% on average in the slope coefficient while it was 31% in item-rest correlation. In other words, when the item discrimination index measure was taken as .20 before factor analysis, the slope coefficient evaluated 30% of the item pool as non-discriminatory items, and the explained variance rate was approximately 9%

higher. Considering the fact that 30% of the item pool is generally consciously assigned as items with low discrimination during item production, it was seen that the findings obtained by using the slope coefficient and the .20 criterion were very close to the real situation.

Table 3. The number of items included in the EFA and the mean ratio of variance explained by a single factor for item discrimination index criterion .35

n	Method	k = 10		k = 30		k = 50		k = 100	
		k _r [95% CI]	Var [95% CI]	k _r [95% CI]	Var [95% CI]	k _r [95% CI]	Var [95% CI]	k _r [95% CI]	Var [95% CI]
50	SC	5.35 [5.03-5.67]	.45 [.44-46]	12.16 [11.34-12.98]	.46 [.45-.47]	19.79 [18.43-21.15]	.47 [.46-.47]	36.70 [34.27-39.13]	.47 [.46-.47]
	Poly	9.10 [8.95-9.25]	.35 [.34-.36]	24.88 [24.58-25.18]	.37 [.36-.38]	40.29 [39.81-40.77]	.38 [.38-.39]	79.18 [78.55-79.81]	.38 [.38-.39]
	ULI	8.19 [7.96-8.42]	.36 [.35-.38]	22.58 [22.16-23.00]	.38 [.37-.39]	37.06 [36.40-37.72]	.39 [.39-.40]	73.44 [72.39-74.49]	.39 [.39-.40]
	RIT	9.05 [8.89-9.21]	.35 [.34-.36]	24.73 [24.43-25.03]	.37 [.36-.38]	39.94 [39.47-40.41]	.38 [.38-.39]	78.70 [78.08-79.32]	.39 [.38-.39]
	RIR	7.28 [7.08-7.48]	.41 [.40-.42]	23.01 [22.70-23.32]	.39 [.38-.40]	38.26 [37.80-38.72]	.40 [.39-.40]	77.00 [76.40-77.60]	.39 [.39-.40]
	SC	5.48 [5.20-5.76]	.43 [.42-.44]	12.73 [12.02-13.44]	.45 [.44-.45]	19.96 [18.84-21.08]	.45 [.44-.45]	37.50 [35.40-39.60]	.45 [.45-.46]
100	Poly	9.39 [9.27-9.51]	.33 [.32-.34]	24.41 [24.13-24.69]	.37 [.36-.38]	39.84 [39.45-40.23]	.38 [.37-.38]	77.75 [77.21-78.29]	.38 [.38-.39]
	ULI	8.72 [8.54-8.90]	.34 [.33-.35]	23.09 [22.83-23.35]	.38 [.37-.38]	37.66 [37.17-38.15]	.38 [.38-.39]	72.83 [71.96-73.70]	.39 [.38-.40]
	RIT	9.37 [9.25-9.49]	.33 [.32-.34]	24.22 [23.94-24.50]	.37 [.36-.38]	39.60 [39.22-39.98]	.38 [.37-.38]	77.32 [76.81-77.83]	.38 [.38-.39]
	RIR	7.25 [7.14-7.36]	.40 [.39-.41]	22.35 [22.11-22.59]	.39 [.39-.40]	37.73 [37.40-38.15]	.39 [.38-.40]	75.32 [74.82-75.82]	.39 [.39-.40]
	SC	5.76 [5.51-6.01]	.41 [.41-.42]	12.75 [12.08-13.42]	.43 [.43-.44]	19.63 [18.65-20.61]	.43 [.43-.44]	36.30 [34.69-37.91]	.44 [.43-.44]
	Poly	9.59 [9.48-9.70]	.31 [.31-.32]	23.48 [23.20-23.76]	.37 [.37-.38]	38.21 [37.90-38.52]	.38 [.37-.38]	74.41 [73.95-74.87]	.39 [.38-.39]
250	ULI	9.10 [8.95-9.25]	.32 [.32-.33]	22.66 [22.37-22.95]	.38 [.37-.38]	36.66 [36.31-37.01]	.38 [.38-.39]	71.90 [71.39-72.41]	.39 [.39-.39]
	RIT	9.52 [9.40-9.64]	.31 [.31-.32]	23.28 [23.01-23.55]	.38 [.37-.38]	38.05 [37.74-38.36]	.38 [.38-.38]	74.14 [73.71-74.57]	.39 [.38-.39]
	RIR	7.11 [7.04-7.18]	.39 [.39-.40]	21.60 [21.44-21.76]	.40 [.39-.40]	36.17 [35.97-36.37]	.39 [.39-.39]	72.65 [72.31-72.99]	.39 [.39-.40]
	SC	5.88 [5.67-6.09]	.41 [.41-.42]	12.86 [12.29-13.43]	.43 [.42-.43]	19.64 [18.78-20.50]	.43 [.43-.43]	35.77 [34.35-37.19]	.43 [.43-.43]
	Poly	9.79 [9.71-9.87]	.31 [.30-.31]	23.47 [23.19-23.75]	.37 [.37-.37]	37.22 [36.93-37.51]	.38 [.38-.39]	72.73 [72.37-73.09]	.39 [.39-.39]
	ULI	9.21 [9.06-9.36]	.32 [.31-.32]	22.31 [22.04-22.58]	.38 [.38-.38]	36.11 [35.77-36.45]	.39 [.38-.39]	70.47 [70.02-70.92]	.39 [.39-.40]
500	RIT	9.76 [9.67-9.85]	.31 [.30-.31]	23.27 [22.98-23.56]	.37 [.37-.38]	36.98 [36.71-37.25]	.39 [.38-.39]	72.53 [72.19-72.87]	.39 [.39-.39]
	RIR	7.02 [6.99-7.05]	.40 [.39-.40]	21.25 [21.16-21.34]	.40 [.39-.40]	35.58 [35.43-35.73]	.40 [.39-.40]	71.16 [70.93-71.39]	.40 [.39-.40]
	SC	5.96 [5.77-6.15]	.41 [.40-.41]	12.98 [12.44-13.52]	.42 [.42-.43]	19.31 [18.59-20.03]	.43 [.42-.43]	35.82 [34.78-36.86]	.43 [.43-.43]
	Poly	9.84 [9.77-9.91]	.30 [.30-.31]	22.80 [22.53-23.07]	.38 [.37-.38]	36.85 [36.59-37.11]	.39 [.38-.39]	71.78 [71.49-72.07]	.39 [.39-.39]
	ULI	9.24 [9.10-9.38]	.31 [.31-.32]	21.81 [21.53-22.09]	.38 [.38-.39]	35.43 [35.17-35.69]	.39 [.39-.39]	69.51 [69.18-69.84]	.40 [.39-.40]
	RIT	9.82 [9.74-9.90]	.30 [.30-.31]	22.65 [22.40-22.90]	.38 [.38-.38]	36.69 [36.44-36.94]	.39 [.38-.39]	71.55 [71.27-71.83]	.39 [.39-.39]
1000	RIR	7.01 [6.99-7.03]	.40 [.39-.40]	21.11 [21.04-21.18]	.40 [.39-.40]	35.27 [35.16-35.38]	.40 [.39-.40]	70.67 [70.50-70.84]	.40 [.39-.40]

SC: Slope coefficient; Poly: Polyserial correlation; ULI: Upper-Lower index; RIT: Item-total correlation; RIR: Item-rest correlation.

Before the exploratory factor analysis, by accepting the item discrimination criterion as .30, the rates of variance explained by a single factor were calculated for 100 replications as a result of removing the items with discrimination below .30 from the analysis. Table 2 provides the mean

number of items included in the analysis, the mean variance explained and the 95% confidence intervals.

Table 2 presents findings similar to the .20 criterion. It was determined that fewer items were included in the analysis when the slope coefficient was used, but the rate of variance explained by a single factor was higher compared to the other methods. On the other hand, it was observed that approximately half of the items were not included in the analysis with the slope coefficient when the .30 criterion was used. For example, an average of 57.85 [57.01-58.69] items were included in the analysis with the slope coefficient in the case where the sample size was 1000 and the item pool size was 100; but when item-rest correlation was used, an average of 75.14 [74.68-75.60] items were included in the analysis. On the other hand, the explained variance rates were found to be .41 and .38, respectively. This shows that although the slope coefficient excluded approximately 20 extra items from the test, it created only a 3% change in the rate of variance that was explained.

Table 3 presents the mean number of items included in the exploratory factor analysis, mean ratio of variance explained and the 95% confidence intervals when the item discrimination index criterion was accepted as .35. The findings in **Table 3** were similar to the findings obtained with the .20 and .30 criteria. On the other hand, the slope coefficient was found to eliminate most of the items in the test when the criterion was .35. For example, for a sample size of 1000 and a item bank of 100, the slope coefficient included an average of 35.82 [34.78-36.86] items in the analysis; while an average of 70.67 [70.50-70.84] items were included in the analysis with the item-rest correlation. This difference which amounted to almost half of the number of items had an effect of only 3% on the variance rate which was explained. The .35 criterion and slope coefficient were far from being ideal choices especially when content validity as considered. On the other hand, similar values obtained with the slope coefficient in the .20 criterion were obtained with other item discrimination indices in the .35 criterion.

4. DISCUSSION and CONCLUSION

A novel item discrimination index was proposed in this study as a new item discrimination index for polytomous items, and this coefficient was compared with the polyserial correlation, item-total, item-rest correlation, and upper-lower 27% groups method. This comparison was done in the context of the coefficient, in the context of the number of items included in the exploratory factor analysis when item discrimination criteria were .20, .30 and .35 before the exploratory factor analysis and in the context of percentages of variance explained by a single factor. Accordingly, it was observed that the slope coefficient generated lower values than the other coefficients in all cases. On the other hand, when the item discrimination criterion was accepted as .20 before the exploratory factor analysis, it was found that more variance was explained with fewer items compared to other indexes. However, it was observed that the values obtained with the slope coefficient in the .20 criterion were obtained with the .35 criterion when other coefficients were used. When the item discrimination criterion was accepted as .35, the slope coefficient eliminated a large number of items, which raised content validity concerns.

Although the current study did not focus on comparing other item discrimination indices within themselves, it was also found that the item discrimination indices obtained specifically by the item-total, item-rest, and upper-lower 27% groups method did not differ significantly from each other. This expected result was also demonstrated by Engelhart (1965), who compared the 10 item discrimination indices. A similar study was conducted by Beuchert and Mendoza (1979) as a Monte-Carlo study, which also reported similar item discrimination indices.

The current study proposed a new method to calculate item discrimination index for Likert-type scales. There are also current studies on calculating discrimination with different methods such as ROC curve (Cum, 2021) and fuzzy logic (Vonglao, 2017). Besides, estimation of item

discrimination parameter based on Item Response Theory is also possible. However, it is difficult for researchers use these methods widely due to the complexity of the calculation of such methods.

In line with all these results, the slope coefficient is recommended as an additional item discrimination index that can be used in scale development studies. However, in cases where the slope coefficient is used, it is recommended to position the item discrimination index criterion around .20. Considering that this study only used simulative data, it is believed that working with real data sets in future studies will improve the research findings.

Acknowledgments

The preliminary results of this study have been presented in 7th Conference of Measurement and Evaluation in Education and Psychology.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Umit Celen: Finding the problem statement and slope coefficient, reporting, data analysis, methodology. **Eren Can Aybek:** Methodology, data generation, data analysis, reporting.

Orcid

Umit Celen  <https://orcid.org/0000-0001-6376-6167>

Eren Can Aybek  <https://orcid.org/0000-0003-3040-2337>

REFERENCES

- Beuchert, A.K., & Mendoza, J.L. (1979). A Monte Carlo comparison of ten item discrimination indices. *Journal of Educational Measurement*, 16(2), 109-117. <http://www.jstor.org/stable/1434454>
- Crocker, L., & Algina, J. (2008). *Introduction to Classical and Modern Test Theory*. Cengage Learning.
- Cum, S. (2021). Examining the discrimination of binary scored test items with ROC analysis. *International Journal of Assessment Tools in Education*, 8(4), 948-958. <https://doi.org/10.21449/ijate.894851>
- Engelhart, M.D. (1965). A comparison of several item discrimination indices. *Journal of Educational Measurement*, 2, 69-76. <https://doi.org/10.1111/j.1745-3984.1965.tb00393.x>
- Johnson, A.P. (1951). Notes on a suggested index of item validity: The U-L index. *Journal of Educational Psychology*, 42(8), 499-504. <https://doi.org/10.1037/h0060855>
- Kelley, T.L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30(1), 17-24. <https://doi.org/10.1037/h0057123>
- Livingston, S.A., & Dorans, N.J. (2004). *A Graphical Approach to Item Analysis*. ETS Research Report.
- Magis, D., & Raiche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, 48(8), 1-31. <https://doi.org/10.18637/jss.v048.i08>
- Martinkova, P., & Drabinova, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, 10(2), 503-515. <https://doi.org/10.32614/RJ-2018-074>
- Moses, T. (2017). A review of developments and applications in item analysis. In *Advancing Human Assessment* (Eds. R. E. Bennett & M. von Davier). Springer Open.

- Oosterhof, A.C. (1976). Similarity of various item discrimination indices. *Journal of Educational Measurement*, 13(2), 145-150. <http://www.jstor.org/stable/1434235>
- Popham, W. J. (2014). *Classroom Assessment: What teachers need to know?* Pearson.
- R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL: <https://www.R-project.org/>
- Revelle, W. (2022) *psych: Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois, USA
URL: <https://CRAN.R-project.org/package=psych>
- Tabachnick, B.G., & Fidell, L.S. (2013). *Using Multivariate Statistics*. Pearson.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.
- Wickham, H., Francois, R., Henry, L., & Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.9
URL: <https://CRAN.R-project.org/package=dplyr>
- Vonglao, P. (2017). Application of fuzzy logic to improve the Likert scale to measure latent variables. *Kasetsart Journal of Social Sciences*, 38(3), 337-344. <https://doi.org/10.1016/j.kjss.2017.01.002>

APPENDIX

Figure 1A. Item discrimination indices for item pool size of 10

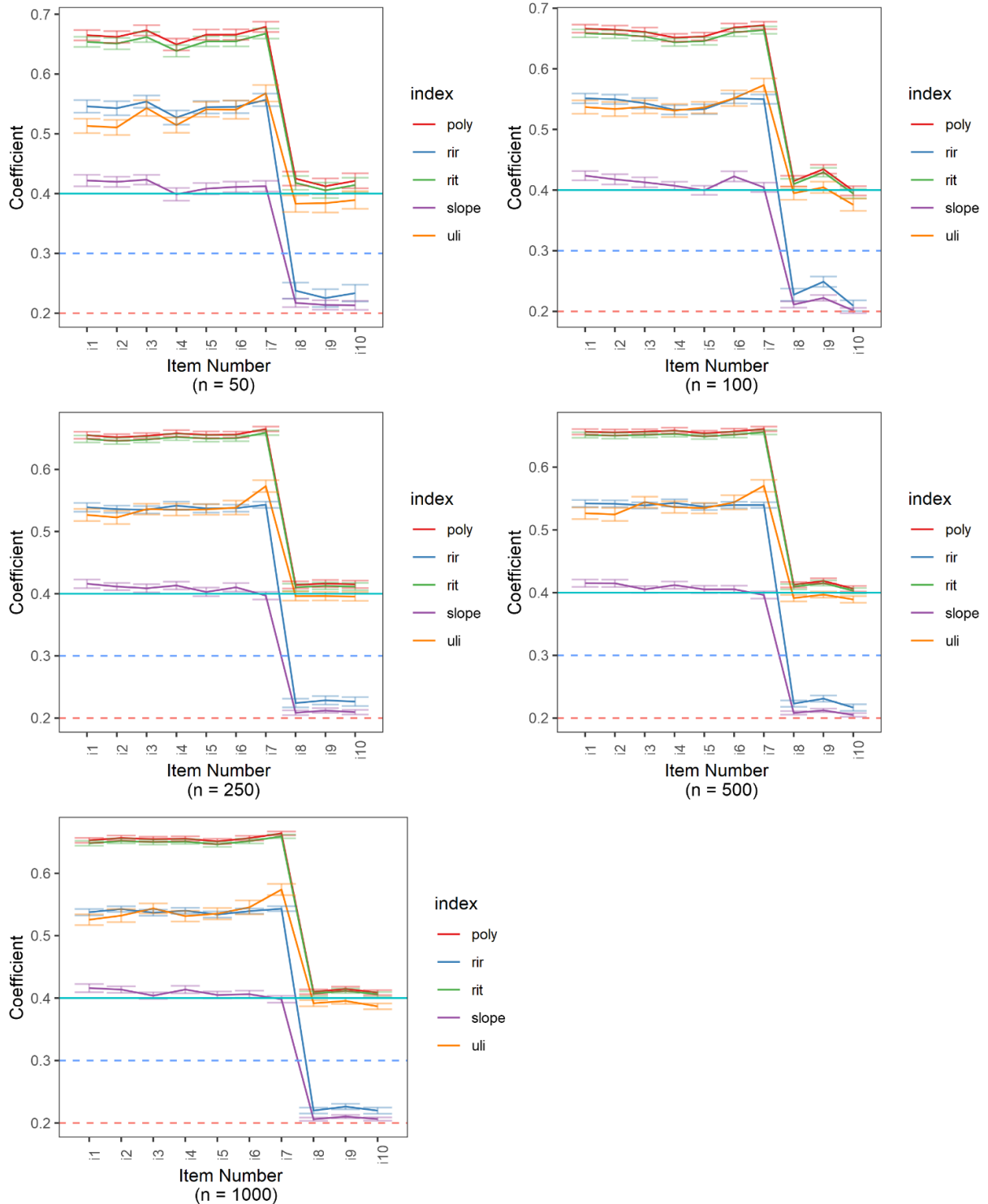


Figure 2A. Item discrimination indices for item pool size of 30

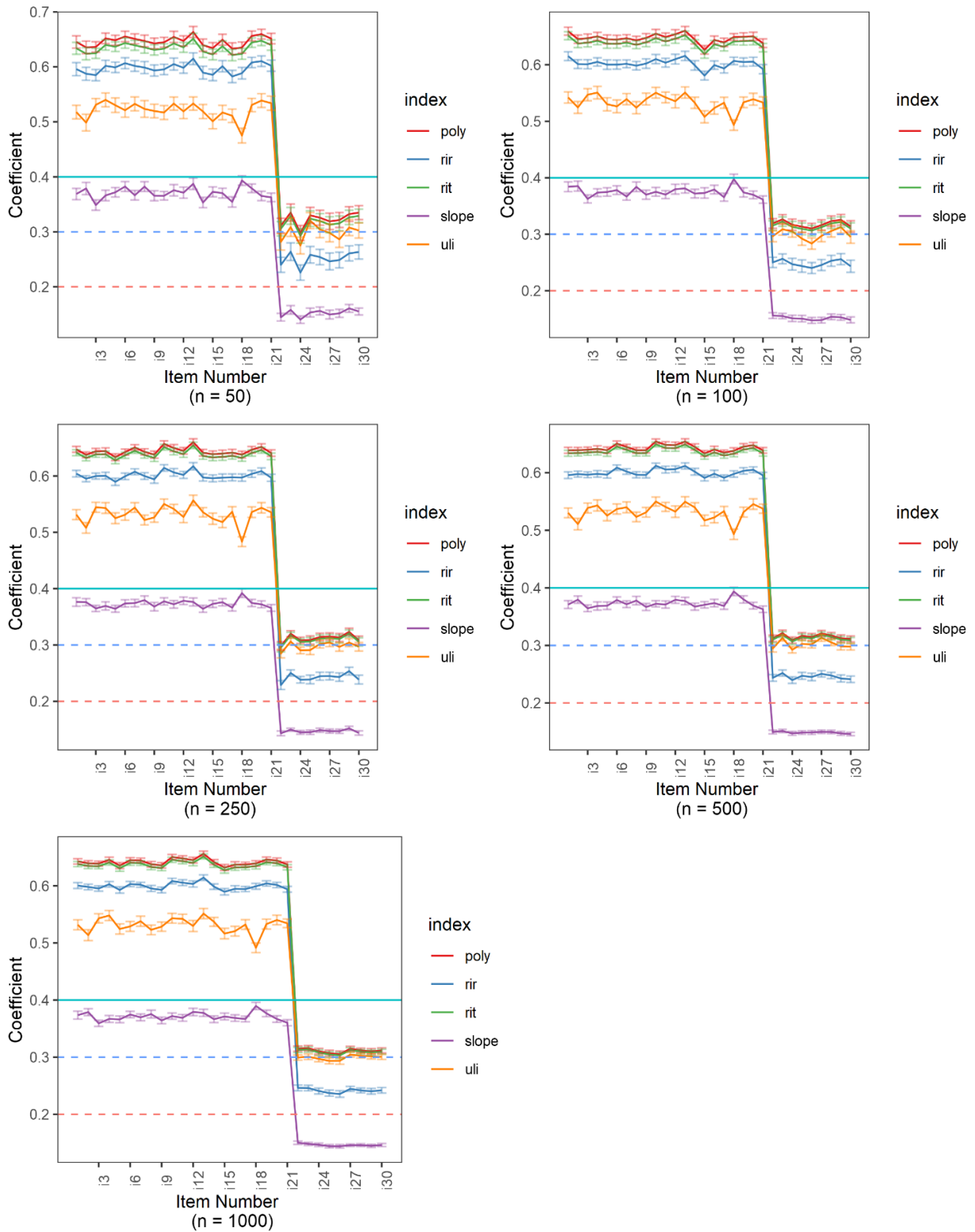


Figure 3A. Item discrimination indices for item pool size of 50

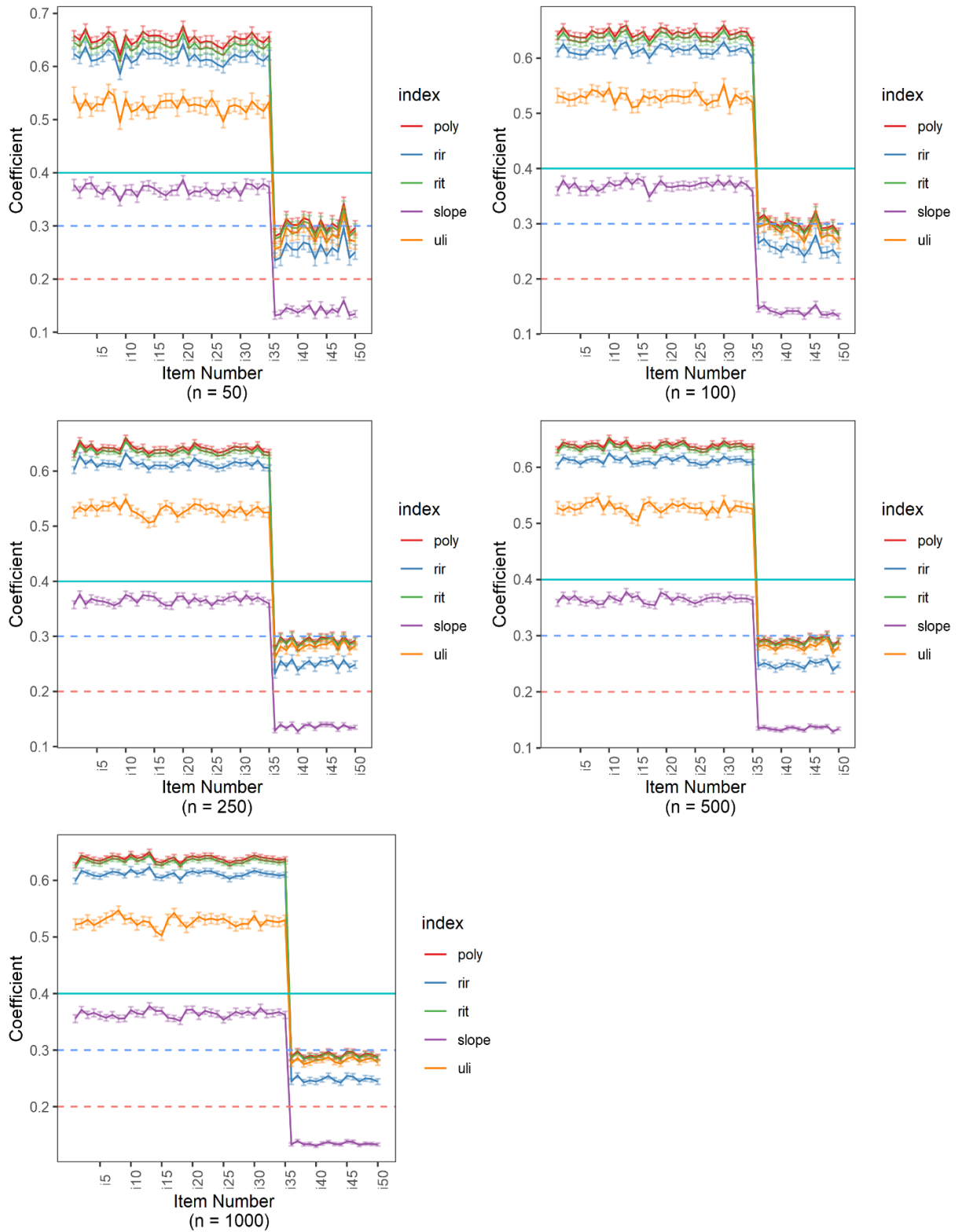
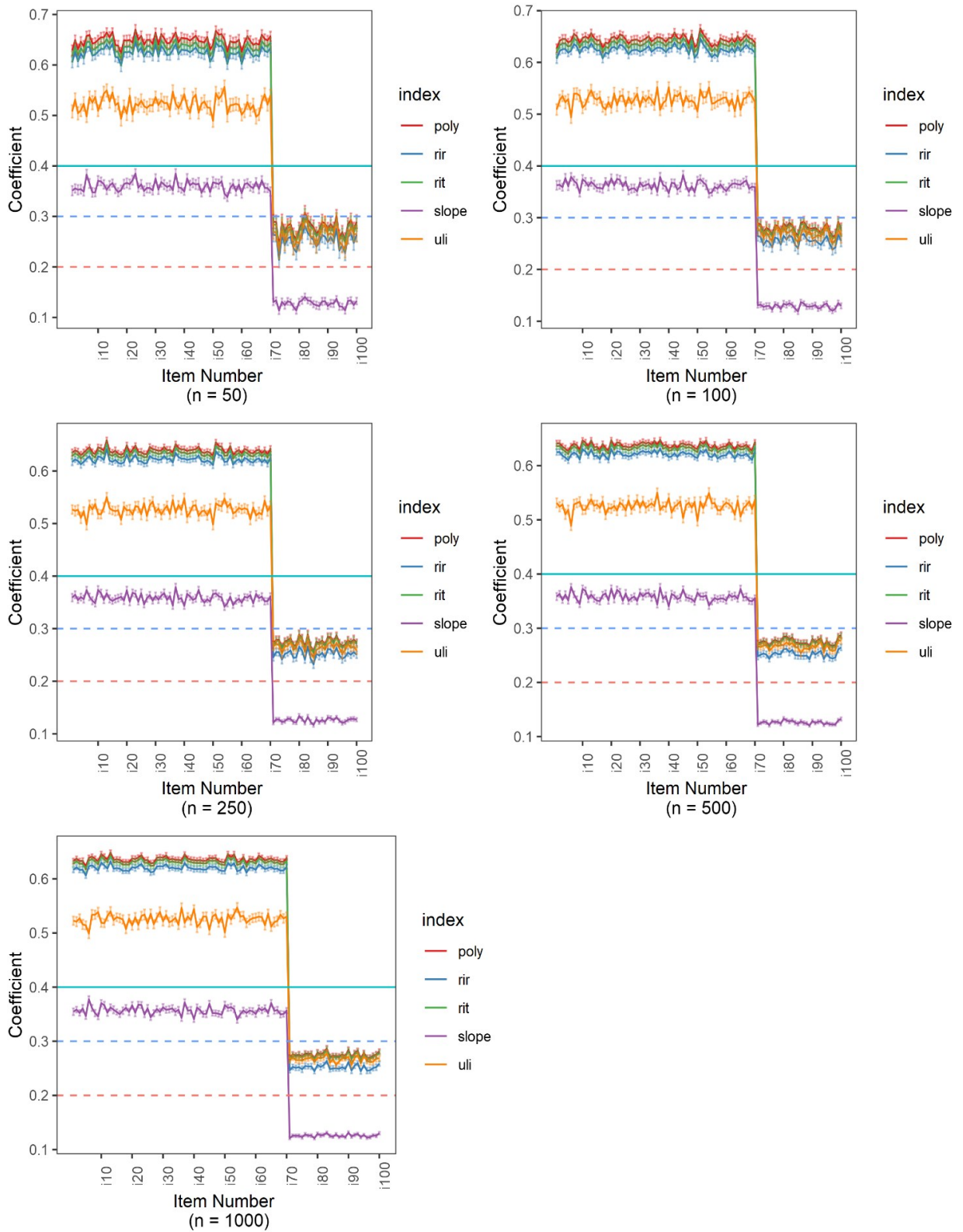


Figure 4A. Item discrimination indices for item pool size of 100



A study of reliability, validity and development of the teacher expectation scale

Hasan Igde^{1,*}, Levent Yakar²

¹Kahramanmaraş Directorate of National Education, Kahramanmaraş, Türkiye

²Kahramanmaraş Sütçü İmam University, Department of Educational Sciences, Kahramanmaraş, Türkiye

ARTICLE HISTORY

Received: July 30, 2021

Revised: Aug. 19, 2022

Accepted: Aug. 21, 2022

Keywords:

Teacher expectation,
Classroom-level teacher
expectations,
Self-fulfilling prophecy.

Abstract: This study aims to develop a 'Teacher Expectation Scale' (TES) to accurately and reliably measure teachers' expectations from their students. The development process of TES has an exploratory mixed method research design. The maximum variety sampling method was used when collecting qualitative data for the study, and the simple random sampling method was used for quantitative data. In the study groups of the research, there are 27 teachers for semi-structured interviews, 423 teachers for Exploratory Factor Analysis (EFA) and 750 teachers for Confirmatory Factor Analysis (CFA). For the content and face validity of the scale, six experts' opinions were obtained. A structure consisting of 36 items and 2 factors was revealed, which explains 73.54% of the total variance as a result of EFA. It has been seen that the items contained in TES show high levels of affiliation to the relevant factors and that all items are discriminative. The explored structure with EFA was evaluated using CFA. The following results were obtained when examining the compliance indices of the obtained model: $\chi^2/df=4.53<5$; CFI=0.99; TLI=0.99; RMSEA=0.07; SRMR=0.05. From the calculated reliability coefficients, McDonald's Omega (0.98) and stratified alpha coefficient (0.96) for the scale overall and Cronbach alpha coefficient (.98) for the dimensions were obtained. Reliability and validity results, obtained from TES showed that it is a valid and reliable measurement tool with two factors and 36 items. The subject of teacher expectation can be examined in terms of many variables using TES developed in the current research.

1. INTRODUCTION

It is possible for individuals to make some predictions about how the phenomena, of which they have an impression, will develop or how others will behave. Individuals often have expectations in accordance with their estimation. When emotions and thoughts turn into actions accordingly with expectations, it means that it is being attempted for the expectations to be realized. A self-fulfilling prophecy process may have been initiated if the source of the expectation being attempted to be realized is based on a false inference. According to the self-fulfilling prophecy theory, while people's beliefs and expectations of what will happen in the future are not actually true, they can have the attitude that will make them a reality (Merton, 1948). When this theory is adapted to the educational context, if a teacher expects a student to succeed, that teacher will

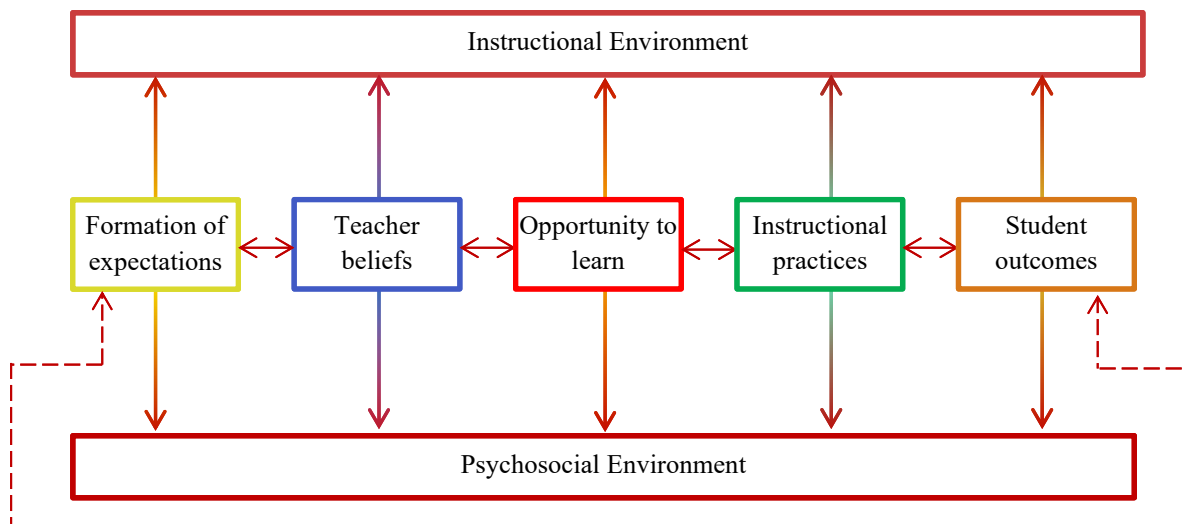
*CONTACT: Hasan IGDE ✉ hasanigde46@gmail.com 📧 Teacher, Kahramanmaraş Directorate of National Education, Kahramanmaraş, Türkiye

probably treat the student in a way that accomplishes his or her expectations. The student, on the other hand, will be likely to fulfil the expectations expected of him or her. Teacher expectation research began with Rosenthal and Jacobson's (1968) seminal work called 'Pygmalion in the Classroom', which paved the way for the concept of self-fulfilling prophecy. Teacher expectation research, which began to be considered as a part of educational psychology with this study, has been an important and developing field of research in terms of the subject area that continues to this day due to its impact on student achievement (de Boer et al., 2010; Rubie-Davies et al., 2020; Wang et al., 2018). The term teacher expectation refers to the inferences teachers make about students' academic and non-academic potential behavior, towards the future based on their experience (Chen et al., 2011; Good, 1987; Riley & Ungerleider, 2012). Teachers work in complex environments where they deal with various events that often develop suddenly and are unpredictable, requiring necessary rapid interpretation and resolution. Research on teacher expectation helps determine how teachers deal with complex processes that can meet the needs of all students. Although successful teaching requires much more than the teacher's expectation developed in the direction that all students can learn, teacher expectations are an important aspect of helping students realize their potential (Good et al., 2018). Teachers plan their lessons and shape learning opportunities under the influence of their expectations (Rubie-Davies et al., 2020). Teachers' approach to their students with their expectations can shape learning outcomes by influencing students' academic beliefs, motivations, and performances (Gershenson et al., 2016; Wang et al., 2018). Understanding teacher expectation is therefore an important element in understanding the nature of teachers' assessment of their students (Dusek & Joseph, 1983).

Teachers' expectations can be individual for each student separately, as well as developing as a whole at the class/group level. In recent years, some works seem to have begun to be conducted that examine teacher expectations at the classroom level as a context, which has not been adequately researched (de Jong et al., 2012; Friedrich et al., 2015; Li & Rubie-Davies, 2017; Timmermans & Rubie-Davies, 2018). Rubie-Davies (2015) has developed a new model, a contextual model of teacher expectation, based on his research on the effects of teacher expectation at the classroom level.

Rubie-Davies' (2015) contextual teacher expectation model is depicted with a series of steps shown in Figure 1.

Figure 1. Rubie-Davies's (2015) contextual model of teacher expectations.



In teacher expectation researches that have been conducted to date, student perceptions of individual teacher expectation are evaluated, and individual interactions between teachers and students, in general, are examined, rather than at the class level (de Boer et al., 2010; Diamond et al., 2004; Friedrich et al., 2015; Hinnant et al., 2009). In contrast with the previous models that focused on individual teacher expectations (see for details. Brophy & Good, 1970; Cooper, 1979; Darley & Fazio, 1980; Rosenthal, 1974) Rubie-Davies's (2015) contextual teacher expectation model, which focuses on classroom-level teacher expectation, offers a broader perspective on the nature of teacher expectation. In this model, the psychosocial environment and teaching environments are defined as the two main tool elements of classroom-level teacher expectation. These intermediary elements affect the social and academic outcomes of students. This model emphasizes that teacher expectation is associated with teacher beliefs and the important role that this relationship plays in influencing teachers' teaching practices and ultimately students' learning opportunities and outcomes. According to Rubie-Davies, just as teacher expectation can affect students' performance, students' behavior can also affect teacher expectation. In this mutual interaction, expectations can be communicated through verbal and non-verbal behaviors. Rubie-Davies' model is designed to illustrate the process of teacher expectation both at individual and class levels.

The concept of 'teacher expectation', whose importance and effects on students have been proven by various international studies, is quite new in the Turkish body of literature, and quite a few studies have been conducted in this field. A small number of studies in the Turkish body of literature (Eryılmaz, 2013; Gökdere, 2013; Kuş & Çelikkaya, 2010; Sazak-Pınar et al., 2012; Tutkun & Dinçer, 2015; Yüksel, 2017) focus more on student characteristics that affect expectations. In teacher expectation research, in which quantitative methods are used quite often (Wang et al., 2018), no scales were found that directly measure teacher expectation according to teacher perception when examined in the Turkish body of literature. Only one scale (Yüksel, 2017) was found to measure indirectly (according to student perception). In the international body of literature, when the scales developed on teacher expectations are examined, it can be said that there are some structural and purposeful differences, although there are partially similar items between them, and the scale developed in the current research. For example, there are significant differences like some scales that have been reached focus only on individual expectations (Szumski & Karwowski, 2019; van den Bergh et al., 2010), and some have a rather small number of items (Auwarter & Aruguete, 2008; Regalla, 2013), some focus only on academic expectation (Barriga et al., 2019; Sweatt, 2000), or some focus on a specific field in academic achievement, such as mathematical achievement (Tiedemann, 2000). On the other hand, as Chen et al. (2011) noted, the teacher expectation phenomenon includes academic and non-academic expectations. No scale has been found in international literature that measures academic expectations as well as non-academic teacher expectations.

The scale of teacher expectations, developed or adapted in accordance with Turkish culture from the point of view of teachers, was not found when scanned in the literature. Considering the mutual cyclical interaction between student behavior and teacher expectations, it is important to examine non-academic teacher expectations as well as academic expectations. This research aims to contribute to a better understanding of the level and direction of teacher expectations for students' academic and non-academic performances. Examining the relationships between various variables and a scale that measures teacher expectations according to teacher perception at the group/class level or school composition can add important insights to the literature. As Rubie-Davies and others (2020) emphasize, although there is now a rich history of teacher expectations, there is still a lot that is unknown. In this context, the aim of this research is to develop a 'Teacher Expectation Scale' that can measure teacher expectations, especially at the group/class level, and to conduct validity and reliability analyses.

2. METHOD

In this study, it is aimed to develop a teacher expectation scale (TES) and conduct validity and reliability analyses. In Turkish culture, the different dimensions of teacher expectations are not exactly known from the point of view of teachers. In this context, it is necessary to first explore the point of view of teachers regarding their expectations. In order to develop a TES based on teachers' points of view and literature, the model of this research is designed in an exploratory sequential design, which is one of the mixed method research types. The goal of the exploratory sequential pattern is to examine the research problem by first discovering it through qualitative data collection and analysis. After this first stage, qualitative data is analyzed, and a new data collection tool is developed from qualitative results. After the scale is developed, newly developed data collection tools are applied for testing (Creswell, 2019, p.41). The qualitative stage, which will meet the requirements of the quantitative stage in studies conducted in the form of scale development, plays a secondary role (Creswell & Plano Clark, 2018, p.98).

2.1. Sample

From the methods of sampling in the qualitative dimension of the research, the maximum variation sampling method was used, and in the quantitative dimension, the simple random sampling method was used. Demographic information of the participating teachers in the study groups that make up the sample of the study is presented in [Table 1](#).

Table 1. Demographic characteristics of teachers in samples.

<i>Data from the sample for Interview</i>			<i>Data from the sample for EFA</i>			<i>Data from the sample for CFA</i>		
Gender	<i>N</i>	<i>(%)</i>	Gender	<i>N</i>	<i>(%)</i>	Gender	<i>N</i>	<i>(%)</i>
Female	14	51.8	Female	239	56.5	Female	403	53.7
Male	13	48.2	Male	184	43.5	Male	347	46.3
Total	27	100	Total	423	100	Total	750	100
Seniority	<i>N</i>	<i>(%)</i>	Seniority	<i>N</i>	<i>(%)</i>	Seniority	<i>N</i>	<i>(%)</i>
1 – 5 Year	8	29.6	1 – 5 Year	57	13.5	1 – 5 Year	191	25.5
6 – 10 Year	8	29.6	6 – 10 Year	103	24.3	6 – 10 Year	199	26.5
11 – 15 Year	1	3.7	11 – 15 Year	90	21.3	11 – 15 Year	119	15.9
16 – 20 Year	5	18.5	16 – 20 Year	57	13.5	16 – 20 Year	98	13.1
21+ Year	5	18.5	21+ Year	116	27.4	21+ Year	143	19.1
Total	27	100	Total	423	100	Total	750	100
School	<i>N</i>	<i>(%)</i>	School	<i>N</i>	<i>(%)</i>	School	<i>N</i>	<i>(%)</i>
Preschool	2	7.4	Preschool	45	10.6	Preschool	65	8.7
Primary School	2	7.4	Primary School	149	35.2	Primary School	241	32.1
Middle School	5	18.5	Middle School	132	31.2	Middle School	286	38.1
High School	18	66.6	High School	97	22.9	High School	158	21.1
Total	27	100	Total	423	100	Total	750	100

As shown in [Table 1](#), there are three groups involved in this study. The priority criteria for maximum diversity at the qualitative stage of the research when determining study groups are the maximum different branches, levels, school types and socioeconomic structures of students in the schools that are assigned to the task, which can be reached in such a way as to best represent the whole. At the quantitative stage, attention was paid to the fact that the study groups reached by simple random sampling method for Explanatory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) consisted of different participants. When research groups

are divided in this direction, all schools in Kahramanmaraş are listed according to the districts through the corporate web page of the Ministry of National Education (MoNE, 2021). The listed schools are classified as school type, school level, provincial, district center and rural schools. Classified schools are divided into two clusters, attempting to make a balanced distribution. Data was collected from the first set for EFA and from the second set for CFA by a simple random sampling method. In this way, in order to generalize the whole, it was attempted to give the possibility of being selected equally to the sample units that could best represent the whole (Büyüköztürk et al., 2019, p.88).

The 27 participants in the first study group of the research constitute the qualitative study group, which is the exploratory aspect of the research in the development of TES. 423 participants in the second study group constitute quantitative study groups in which data is collected for EFA and 750 participants in the third study group for CFA. In order to determine the psychometric properties of the scale to be developed in scale development studies with minimal errors, the minimum number of data ($N=300$) specified by Tabachnick and Fidell (2015) for factor analysis was used as a base. In addition, Kline (1994) suggests that the sample size should be 10 times the number of items (Çokluk et al., 2018, p.206). An attempt was made to reach the number of samples to exceed the specified number of data and the proposal. In this direction, the number of teachers from which the data is collected is seen in [Table 1](#), where the total number of items (36) found on the final scale developed in the current research is more than 10 times for EFA and 20 times for CFA.

2.2. Scale Development Process

During the development process of TES, the subject area was first examined. After the literature review, semi-structured interviews that lasted on average for 20 minutes were conducted face to face with 27 teachers, 15 of which are in a different branch, who work in various schools in terms of socioeconomic levels, so as to form a basis for the item pool of the scale.

Content analysis was performed on responses obtained from teacher opinions. Simultaneously with the content analysis, scale items were written, and the item pool began to be created. The items found in the item pool, during the development of TES, were compared with the information and findings given in the literature and an item pool was attempted to be realized in accordance with the literature (Gökdere, 2013; Kuş & Çelikkaya, 2010; Öztürk et al., 2002; Rubie-Davies, 2004, 2006, 2007, 2010, 2015). During the preparation of the item pool, the findings of expectation studies conducted in Türkiye (Erçetin et al., 2020; Kuş & Çelikkaya, 2010; Yurtal & Yontar, 2006), the general and specific purposes of the Turkish Ministry of National Education, the general competencies of the teaching profession (MoNE, 2017), as well as scale development studies that may include similar items with TES (Barriga et al., 2019; Eden et al., 2000; Sarıtepeci, 2018; Timmermans & Rubie-Davies, 2018) were benefited from. When writing scale items, the opinions of the participating teachers were examined individually, some expressions were changed and turned into a scale item, and consistency with the literature was given importance. In the item pool, firstly 47 items were written. Six experts, two of whom are experts in measurement and evaluation, one is an expert in educational programs and teaching, three are experts in educational sciences in the field of teacher training, and the faculty members were consulted on the written statements. According to expert opinions, some expressions, which are similar to each other, distorted in terms of meaning, or considered not to measure teacher expectations, were removed from the item pool, some items were corrected, and some new items were written. After expert opinions, the number of items was reduced to 42. After that, six teachers, the majority of whom received a master's degree or doctorate in the field of Educational Sciences, studied the items individually in terms of comprehensibility. After the reviews, the opinions of four Turkish language experts were received in terms of language and expression. After the teachers' opinions, it was seen that some

expressions in some items evoked different connotations, and new corrections and item subtractions were made. In the last case, a 37-item draft scale was developed. The scale is developed in five-level Likert-type as; 1-*Strongly disagree* (1.00-1.80), 2-*Mostly disagree* (1.81-2.60), 3-*Moderately agree* (2.61-3.40), 4-*Mostly agree* (3.41-4.20), 5-*Strongly agree* (4.21-5.00).

2.3. Data Analysis

Data analysis was carried out in two stages in the form of qualitative and quantitative data analysis.

2.3.1. Analysis of qualitative data

The content analysis method was used in the analysis of qualitative data collected in the research. Content analysis helps determine the existence of certain words or concepts in texts (Büyüköztürk et al., 2019, p.259) It is an analysis method for defining data, revealing the facts hidden in the data, classifying similar data within a specific concept and theme, and interpreting them by organizing them in a way that the reader can understand (Yıldırım & Şimşek, 2013). In this direction, the available data was analyzed by the researcher and encodings were made with a series of repetitions, including components and operations such as taking Edge notes on data sets, summarizing data, drawing conclusions, creating simple relationship sets, and returning to data sets again. Expert opinion on the coding has been taken. Because some codes may have the same meaning in expert evaluations, they were taken as a single code and new arrangements were made for encodings that did not meet the sub-themes. After the arrangements, the common or similar aspects between the resulting codes were re-examined and the theme and sub-themes were systematic, and the interrelated codes were collected under the relevant theme. An attempt was made to be written by associating the item pool one-on-one with the themes and codes that appeared in the content analysis. Codes and themes reached by content analysis in the research constitute the discoverer aspect of the current research.

2.3.2. Analysis of quantitative data

Lisrel 8.8, IBM SPSS Amos 24, Factor 10.5.03 and IBM SPSS Statistic 26.0 package programs were used in descriptive and structural statistics of quantitative data. The level of significance is designated as .05 in statistical analysis. The validity and reliability of the scale, developed in accordance with qualitative and quantitative data analysis, were examined. Content validity of the scale in accordance with expert opinions and its structure was analyzed by EFA. Kaiser-Meyer-Olkin (KMO) coefficient and Bartlett Sphericity test were used to decide whether the data was suitable for factor analysis. In EFA analysis, the 'Maximum Likelihood Factor' Analysis method, which is a method of removing factors that have high similarities as a factorization technique, was selected, because it is thought that there is a relationship between factors, the 'Direct Oblimin' oblique rotation method was used. After the rotation process, the decision was made by evaluating the results of the eigenvalue slope graph and the parallel analysis method (Timmerman & Lorenza-Seva, 2011) together. The relationship between the score of each item and the total scale score was determined by the Pearson moments product correlation coefficient. Independent samples were analyzed by the *t*-Test to show that items can well distinguish between those with properties they want to measure and those without. CFA analysis was conducted to test whether the defined and bounded structure of TES was verified as a model. In addition, convergent and divergent validity methods with combined reliability have been applied as additional proof of reliability. Combined reliability is used to measure the overall reliability of multiple, heterogeneous, but similar expressions (Raykov, 1998). Convergent validity means that expressions for variables are related to each other and to the factor they form, while divergent validity means that expressions for variables must be less related to the factors they do not belong to than the factors in which they are located (Yaşlıoğlu,

2017). The combined reliability and average variance (AVE) values achieved were calculated in Excel 2010.

3. FINDINGS

This section presents the findings related to the validity and reliability analysis of TES.

3.1. Content Validity

The qualitative findings of the research on exploratory evidence are divided into two themes, academic and non-academic expectations. In the academic expectation theme, 17 codes were reached, while in the non-academic expectation theme, 14 codes were reached. Scale expressions developed with codes reached under generated themes are mapped by association. Expert opinions have been taken on scale expressions mapped to codes. As a result of expert opinions, the scope was validated, and it was determined that there was a semantically close relationship between each developed scale expression and the opinions of the participating teachers.

3.2. Construct Validity

The theoretical basis for the scale developed in the current research is based on Rubie-Davies' (2015) contextual teacher expectation model. It can be said that the theoretical structure of the scale is in accordance with the definitions of academic and non-academic teacher expectations by Chen et al. (2011). In addition, academic and non-academic teacher expectation themes created by content analysis in the qualitative dimension of the research are consistent with the structure discovered with EFA. This consistent structure has been confirmed by the CFA. This shows that the teacher expectation structure developed in qualitative analyses is generalizable with quantitative analyses and provides additional evidence for the structural validity of TES.

3.2.1. Normality analysis

In the process of scale development, normality analysis of the data obtained from the second study group was carried out. The suitability of the data for normal distribution was decided by looking at kurtosis and skewness values from analytical methods, as well as other graphical methods. In the analysis of the data obtained from the second study group, the skewness coefficient was found to be -0.321 and the kurtosis coefficient was 0.252. Accordingly, the fact that the skewness and kurtosis coefficients are between ± 1.5 values (Tabachnick & Fidell, 2015) shows that the data meet the normality assumption.

3.2.2. Exploratory factor analysis

The scale was applied online to 483 participants reached for EFA. Data from the application was examined using Microsoft Office Excel 2010 prior to EFA. 22 data with the same demographic information and responses that appeared to have responded two or more times were extracted from the dataset. In addition, standard Z values were looked at to determine the end values in the dataset prior to EFA, and 38 data that were not in the ± 3.29 range (Tabachnick & Fidell, 2015) were removed from the dataset. Thus, EFA was applied to the data set consisting of the remaining 423 teachers' responses to 37 items.

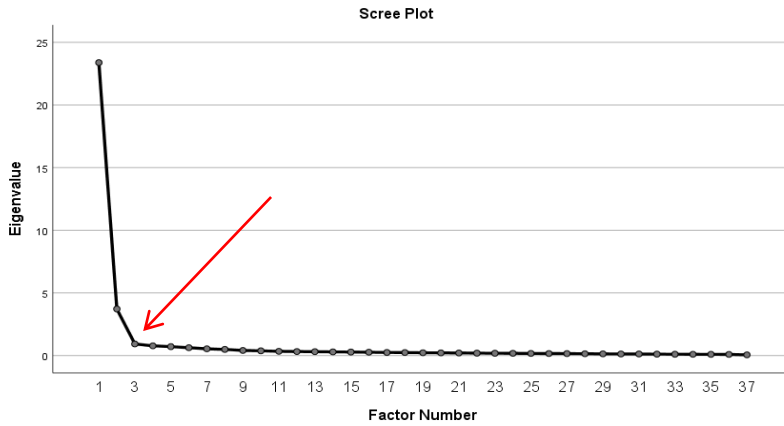
The Kaiser-Meyer-Olkin (KMO) coefficient and Bartlett Sphericity test examined whether the data was suitable for factor analysis. Test results are presented in [Table 2](#).

Table 2. Kaiser-Meyer-Olkin (KMO) test and Bartlett's sphericity test results.

Kaiser-Meyer-Olkin sample suitability measure (KMO)	.98
Bartlett's sphericity test chi-squared value	19461.37
Degree of freedom (Df)	666

Based on the value ranges specified in Table 2, the value of the KMO appears to be .98 in terms of the size of the sample of 423 people. According to this value, the sample is ‘excellent’ and the chi-square value determined by the Bartlett Sphericity Test results is significant ($X^2_{(666)}=19461.37; p<0.01$). The slope chart of the scale is presented in Figure 2.

Figure 2. Scree plot.



As a result of EFA, the eigenvalues of the scale are collected under two factors greater than 1. It is seen in Figure 2 that the eigenvalues are very close to each other below 1 beginning from the third factor.

According to the data analysis, the charge value of the 37th item written as the reverse is designated as .350 in the first factor, and .038 in the second factor. Based on the opinion (Büyüköztürk, 2015) that selecting items with a load value greater than 0.45 would be a good criterion when studying factor loads, so the 37th item has been removed from the scale. Factor analysis has been renewed over the remaining 36 items. The eigenvalues, total variance and parallel analysis proofs explained by the scale after the matter extraction are presented in Table 3.

Table 3. Total variance table explained by the scale.

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Parallel Analysis	
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Real Data Variance	Random Variance
1	23.29	64.69	64.69	22.964	63.79	63.79	66.2	5.7
2	3.70	10.27	74.96	3.51	9.75	73.54	10.5	5.4
3	0.78	2.16	77.12				2.1	5.1

Extraction Method: Maximum Likelihood.

By examining Figure 2 and Table 3, it is seen that the scale elements are collected under factor 2. In the analysis made by the parallel analysis method, the real data variation values in the first two factors are greater than the random variation values. The eigenvalue of the third factor is less than 1. The results of the analysis to determine the number of factors appeared to support each other.

The total variance described by EFA is 73.54%. After the oblique rotation of the Direct oblique, the first sub-dimension contributed to the total variance of 20.85, and the second sub-dimension contributed to the total variance of 19.26. It seems that the cumulative variance finding is above the acceptable level of 60% (Çokluk et al., 2018, p.239).

Since the two-factor structure of TES discovered by EFA corresponds to the definitions of teacher expectations by Chen and others (2011), the items collected in the first factor are titled ‘Academic Expectations (AE)’ and the items collected in the second factor are titled ‘Nonacademic Expectations (NE)’. The item and factor distribution loads of the scale are presented in Table 4.

Table 4. Items and factor loadings (Not proofed English version).

Factors	Item No	Items	Factor Loads	
			1	2
ACADEMIC EXPECTATIONS	4	I think they will achieve their goal.	0.968	0.143
	8	I think their motivation for studying will be high.	0.936	0.067
	3	I think their level of success will be high.	0.928	0.105
	5	I think they will have academic confidence.	0.926	0.071
	13	I think they will gain the learning outcomes included in the curriculum.	0.890	0.040
	9	I think they'll learn the content of the lessons.	0.889	0.006
	6	I think they'll be interested in their lessons.	0.875	0.001
	7	I think they will fulfil their responsibilities for their classes.	0.867	0.015
	16	I think they'll be ready for higher education training.	0.866	0.008
	12	I think they'll ask effective questions in class.	0.850	0.017
	10	I think they will actively participate in the classes.	0.820	0.052
	1	I think they will succeed in the exams they will take.	0.813	0.003
	15	I think they'll have a prepared approach to their development period.	0.806	0.068
	2	I think they will set goals for success.	0.803	0.034
	11	I think they'll give me the right answers to my questions about the lesson.	0.766	0.088
	14	I think they'll reflect on what they've learned in classes in their lives.	0.747	0.077
	17	I think they'll discover their abilities.	0.736	0.109
	18	I think they'll care about their personal development.	0.718	0.131
	20	I think they will use Turkish in accordance with the rules of the language.	0.607	0.232
	19	I think they'll communicate effectively.	0.594	0.271
36	I think they will have high-status professions.	0.593	0.188	
NONACADEMIC EXPECTATIONS	33	I think they will be individuals who respect people around them.	0.071	0.974
	32	I think they will respect values.	0.127	0.968
	31	I think they care about national values.	0.114	0.966
	27	I think they'll be useful people to society.	0.011	0.900
	28	I think they will adopt the behavior that society expects of them.	0.005	0.891
	22	I think they'll be individuals of character.	0.004	0.884
	29	I think they will be sensitive individuals to social events.	0.051	0.867
	34	I think they will build positive relationships with their families.	0.006	0.866
	21	I think they will have moral virtues.	0.013	0.845
	30	I think they'll be sensitive to protecting the natural environment.	0.075	0.834
	25	I think they'll be kind.	0.094	0.826
	26	I think they will show empathic approaches.	0.127	0.765
	24	I think they will show positive behavior appropriate to their developmental period.	0.190	0.736
	23	I think they will take care of their personal hygiene.	0.147	0.704
	35	I think they'll pay attention to their choice of friends.	0.263	0.634

Extraction Method: Maximum Likelihood, Rotation Method: Oblimin with Kaiser Normalization

A closer look at Table 4 suggests factor head values of 21 items (Item No: 1-20 and 36) found in the AE factor of TES change between .593 and .968, and factor head values of 15 items (Item No: 21-35) found in NE factor change between .634 and .974. It is seen that items in the scale are associated with a factor that is close to or above the value of .60. Items binding to over .60 related factors indicate high-level binding (Kline, 1994). In terms of whether the head values of items, boarding and factor meet the acceptance level, the difference between the load values of items are higher than the acceptance level and the load values of items have in two factors greater than .1 (Çokluk et al., 2018, p.233). In this direction, it can be said that TES is a powerful measuring tool.

3.2.2.1. Reliability Study of the Scale. The variance and alpha coefficients explained by each factor are presented in Table 5.

Table 5. Reliability of the scale and sub-factors.

Factors	Item Number	Variance	Cronbach's Alpha	McDonald's Omega
Factor 1 (AE)	21	% 63.79	0.98	0.98
Factor 2 (NE)	15	% 9.75	0.98	0.98
Total	36	% 73.54	0.98	0.98

In Table 5, it is seen that the reliability coefficient values of Cronbach's Alpha and McDonald's Omega are the same values. The Alpha and Omega coefficients of the first and second factors are 0.98. The scale-wide reliability coefficient value obtained with a stratified alpha of TES is calculated as 0.96. After calculating the reliability coefficients of the scale, the internal consistency reliability of the scale was calculated by the Split-half method. The internal consistency coefficient values obtained by analyzing the scale by the Split-Half method are presented in Table 6.

Table 6. Internal consistency coefficients of the scale (Split-Half).

Factor	Cronbach's Alpha		Correlation Between Forms		Spearman-Brown Coefficient		Guttman Split-Half
	Part1	Part2	N of Items	r	Equal Length	Unequal Length	Coefficient
1	.97	.96	21	.91	.96	.96	.95
2	.97	.96	15	.93	.96	.96	.96

In Table 6, it can be said that the internal consistency coefficient values of the two groups formed by analyzing TES separately for each factor by the Split-Half method are close to each other and are very good. These values indicate that items are regulated in a sequential nature (Ocak & Park, 2019). Positive and high levels of linear relationships were found between the groups. When Guttman and Spearman-Brown coefficients are evaluated, it can be said that TES has high reliability.

3.2.2.2. Item Analysis. In order to determine the item discrimination of the scale, the total score of the scale was determined and item analyses were performed on the lower 27% (N:114) and upper 27% (N:114) groups. On the scale, it was found that there was a significant difference between all items compared to the lower and upper groups of 27% compared to the independent samples t-Test. T values for the lower and upper groups range from -16.43 (sd:226, p<.01) to -22.30 (sd:226, p<.01). Adjusted item total test correlation values range from 0.73 to 0.84. Analysis of items by comparing TES' total test correlations with lower and upper groups of 27% is shown in Table 7.

Table 7. Item-Total statistics.

Factors	Item No	Bottom 27% Group (N: 114)		Top 27% Group (N: 114)		t	Corrected Item-Total Correlation
		\bar{X}	S	\bar{X}	S		
ACADEMIC EXPECTATIONS	1	2.40	0.69	3.90	0.69	-16.43	0.76
	2	2.41	0.62	3.90	0.64	-17.88	0.78
	3	2.36	0.64	3.86	0.65	-17.56	0.77
	4	2.46	0.63	3.93	0.61	-18.07	0.77
	5	2.37	0.66	3.99	0.62	-19.25	0.80
	6	2.49	0.63	4.10	0.53	-20.84	0.81
	7	2.51	0.60	4.11	0.50	-21.78	0.82
	8	2.35	0.58	3.96	0.56	-21.32	0.81
	9	2.48	0.57	4.04	0.49	-22.28	0.83
	10	2.51	0.57	4.09	0.56	-21.18	0.81
	11	2.59	0.59	4.03	0.49	-20.02	0.79
	12	2.46	0.57	3.91	0.59	-19.03	0.78
	13	2.54	0.57	4.00	0.58	-19.18	0.79
	14	2.41	0.65	4.01	0.65	-18.63	0.77
	15	2.38	0.62	3.98	0.59	-20.28	0.81
	16	2.30	0.62	4.00	0.59	-21.35	0.80
	17	2.39	0.57	4.00	0.67	-19.64	0.79
	18	2.44	0.57	3.97	0.57	-20.39	0.79
	19	2.50	0.66	4.15	0.60	-19.84	0.80
	20	2.24	0.64	4.02	0.62	-21.24	0.78
36	2.29	0.63	3.82	0.65	-18.00	0.73	
NONACADEMIC EXPECTATIONS	21	2.92	0.71	4.41	0.51	-18.26	0.77
	22	2.96	0.69	4.44	0.52	-18.26	0.79
	23	2.86	0.75	4.39	0.54	-17.61	0.77
	24	2.80	0.58	4.33	0.48	-21.83	0.84
	25	2.75	0.65	4.47	0.52	-22.20	0.83
	26	2.65	0.60	4.32	0.56	-21.96	0.80
	27	2.98	0.65	4.55	0.53	-19.91	0.81
	28	2.89	0.60	4.40	0.54	-19.88	0.80
	29	2.84	0.63	4.52	0.57	-21.05	0.82
	30	2.77	0.63	4.49	0.54	-22.30	0.82
	31	3.06	0.78	4.61	0.51	-17.72	0.75
	32	3.08	0.78	4.61	0.51	-17.54	0.74
	33	2.95	0.70	4.57	0.52	-19.91	0.80
	34	2.88	0.61	4.43	0.56	-19.93	0.78
	35	2.63	0.63	4.27	0.57	-20.67	0.81

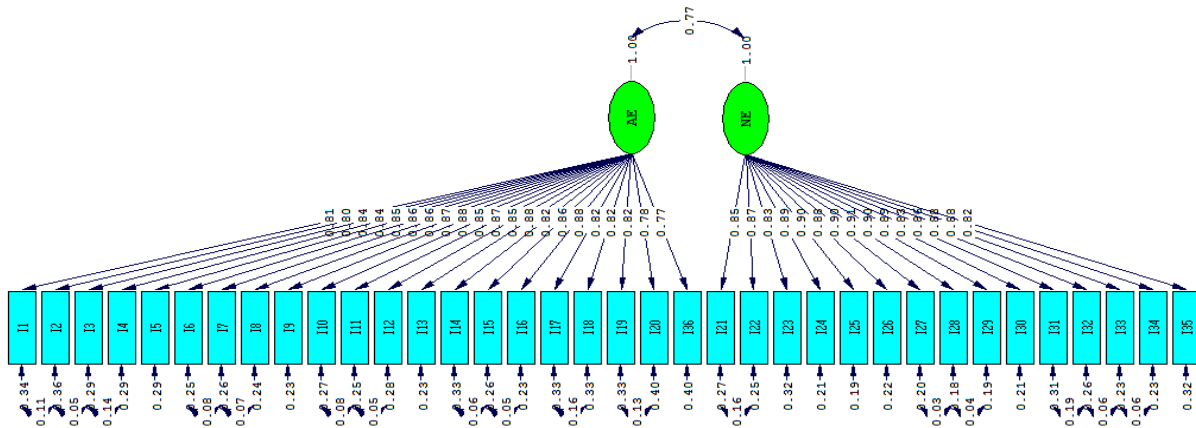
Since the total correlation value of all items contained in the scale is $r \geq 40$, each item found on the scale can be considered ‘a very good item’ (Büyüköztürk, 2015). After EFA, a 2-dimensional structure emerged in which 36 items can take place on the final scale.

3.2.3. Confirmatory factor analysis

The scale was applied online to 875 different participants from the study group reached by EFA. The data from the application was examined using Microsoft Office Excel 2010 before CFA.

59 data with the same demographic information and responses that appeared to have responded two or more times were extracted from the dataset. In addition, standard z values were looked at to determine the end values in the dataset prior to CFA, and 66 data that were not in the ± 3.29 range (Tabachnick & Fidell, 2015) were excluded from the dataset. Thus, CFA analyses were performed on the data set of the remaining 750 teacher responses. A diagram of the model that appeared in accordance with the CFA is presented in Figure 3.

Figure 3. CFA Diagram for TES.



In CFA, t values were examined first. T values which show description states of latent variables to indicator variables exceed 2.56. Its level has been seen to be significant ($p < .01$). T values of all items range from 25.12 to 32.34. In this case, after CFA analysis, items were found to confirm the factors they are related to in the 99% confidence range ($p = .000$). Error variances in the path diagram of the scale have been studied. M28 has the lowest error variance examining the error variances of observed variables. It has a value of 0.18, and the highest error variance is by the M36th item. As it is stated in Figure 3, it has a value of 0.40. When the error variances of the scale are examined, it can be said that there is no item with a high error variance (Çokluk et al., 2018, p.305). In addition, it can be said that there is no incompatible value, and the relationship between hidden variables and observed variables is significant ($p < .05$). Standardized coefficients of 36 items found in TES are between .77 and .91. In this direction, it can be said that there is no item that should be excluded from the analysis.

Modification Indices (MI) for covariance indicate the connection between error terms. This covariance between error terms refers to the measurement error. The most common cause of this error is that the two expressions are understood in the same format, even if they are usually written in different forms (Yaşlıoğlu, 2017). In this direction, when modifying TES, covariance connections were established between indicators that have a closely related meaning in terms of expression, are successive in the order of expressions and are in the same factor. Accordingly, the recommended modification indexes in the CFA analysis were examined. Recommended modification indexes were evaluated in different aspects. First of all, attention was paid to ensure that the modified items were in the same subscale. Secondly, with the thought that consecutive answers may affect each other, only consecutive items were modified. Lastly and more importantly, the criterion of closeness in the meaning of the modified items was used. Totally, 164 covariance connections were proposed. In AE factor, 11 covariance connections (1-2, 2-3, 3-4, 6-7, 7-8, 10-11, 11-12, 14-15, 15-16, 17-18, 19-20) have been established between the indicators. In NE factor, 6 covariance connections (21-22, 27-28, 28-29, 31-32, 32-33, 33-34) have been established between the indicators. It was found that the modifications made a significant contribution to the chi-square value. Compliance indexes for CFA analysis before and after modification are given in Table 8.

Table 8. *Confirmatory factor analysis compliance indexes.*

	χ^2/df	RMSEA	GFI	AGFI	CFI	NFI	TLI	RMR	SRMR
Pre-modification	7.79	0.11	0.69	0.65	0.98	0.98	0.98	0.04	0.05
Post-modification	4.53	0.08	0.81	0.78	0.99	0.99	0.99	0.03	0.05
Compliance Indicator	Acceptable	Acceptable	Acceptable	Weak fit	Excellent	Excellent	Excellent	Good fit	Good fit

Sources: Brown, 2006; Hooper et al., 2008; Jöreskog & Sörbom, 1993; Tabachnick & Fidell, 2001; Ulrich & Lehrmann, 2008; as cited in Ocaik & Park, 2019.

As a result of CFA, χ^2/df , Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), Comparative Fit Index (CFI), and standardized Root Mean Square Residual (SRMR) index values were reported, which were proposed to be examined by Kline (2019) to determine the validity of the model. However, some commonly used harmony indexes are studied in the literature (Çokluk et al., 2018; Tabachnick & Fidell, 2015). In CFA analysis, the value of χ^2 after modification was found to be 2610.72 and the degree of freedom was found to be 576. When these two values are divided into each other, $\chi^2/df(2610.72/576)$ results in a value of 4.53. The threshold value $\chi^2/df \leq 5$ was accepted when interpreting this value (Wheaton et al., 1977). It can be said that $\chi^2/df=4.53$ shows acceptable compliance (Kline, 2019). The value of RMSEA shows a good fit when it is between .05 and .08, and it is thought acceptable between .80 and .10. When CFI and TLI are higher than .90 and when they are close to .95, they are indicative of suitable models (MacCallum et al., 1996; Hu & Bentler, 1999; as cited in Zhu et al., 2018). After modification, it is seen in Table 8 that the RMSEA value is 0.08, the GFI is 0.81, the AGFI is 0.78, and the SRMR compliance index is 0.05.

TES compliance indexes and Criterion compliance indexes were compared. In comparison, it can be said that the TES model developed in the classroom/group-level teacher expectation structure, along with the sub-dimensions of the scale items, has been verified and has generally acceptable compliance indexes.

3.2.4. Composite reliability, convergent validity and divergent validity

Test results for composite reliability and convergent and divergent validity of TES are presented in Table 9.

Table 9. *Composite reliability, convergent and divergent validity test results.*

Factors	CR*	AVE**	AVE SQUARE ROOT
AE	0.97	0.64	0.80
NE	0.97	0.71	0.84

*CR=Composite Reliability, **AVE= Average Variance Extracted

As a result of the analysis applied to the data obtained for CFA, the correlation value between AE and NE factors, which are the sub-dimensions of TES, was found to be 0.75 ($p < .01$). In addition, the fact that CR values for factors are 0.97 provides strong empirical evidence of scale reliability. For Convergent validity, it is seen in Table 9 that the CR values in the sub-dimensions of TES are greater than the average variation Extracted (AVE) values and the AVE values are greater than 0.5 (Raykov, 1998). Fornell and Larcker (1981) state that the fact that AVE square root values are greater than the sub-dimension correlation values is proof of divergent validity. It was found that AVE square root values in factors are greater than the correlation value between factors. As part of the results obtained, it can be said that the desired conditions for composite reliability and convergent and divergent validity are met.

4. DISCUSSION and CONCLUSION

Teacher expectation is the beliefs that teachers have about students' academic abilities and their subsequent success levels (Peterson et al., 2016), and achievements that teachers expect students to gain over time (Rubie-Davies et al., 2020). Many studies have tried to reveal that teacher expectations affect student performances in some ways. When studying this effect, it was found that the research highlighting the characteristics of teachers (Park & Byun, 2020; Peterson et al., 2016; Rubie-Davies et al., 2012; Timmermans & Rubie-Davies, 2018; Watson et al., 2017) seems to have started recently and has become a new focal point (Li, 2016). There is a need for scales that can measure classroom/group-level teacher expectations according to teacher perceptions in Turkish literature (İğde, 2021). In this direction, in the current research, it is aimed to develop a teacher expectation scale that can measure the expectation factors arising from the perceptions and attitudes of teachers at the classroom/group level and test the validity and reliability of the measurement scale.

The opinions of teachers are taken first when preparing the teacher expectation scale (TES). Codes and categories are organized by content analysis of teachers' opinions. Scale items are written in a way that is related to the specified codes and categories and in accordance with the teacher expectation literature (Chen et al., 2011; Rubie-Davies, 2004, 2006, 2007, 2010, 2015; Szumski & Karwowski, 2019). Expert opinions are taken to ensure the validity of the scope and outlook of the scale items. In accordance with the expert opinions, the content, size and description of the items are revised. The 47-item draft scale written before is organized as 37 items after expert opinions. The number of organized items differs from most studies in the teacher expectation literature. In teacher expectation studies, scales comprising of one item (Gregory & Huang, 2013; Papageorge et al., 2019; Peterson et al., 2016; Rubie-Davies et al., 2020; Watson et al., 2017; Zhu et al., 2018) or a couple of items, (Archambault et al., 2012; Denessen et al., 2020; Friedrich et al., 2015; Gentrup et al., 2020; Rubie-Davies & Peterson, 2016; Timmermans & Rubie-Davies, 2018) are widely used (Friedrich et al., 2015). The first studies on the subject (Babad et al., 1982; Rosenthal & Jacobson, 1968) and other studies conducted since (de Boer et al., 2010; Gentrup et al., 2020; Papageorge et al., 2019; Szumski & Karwowski, 2019; Zhu et al., 2018) mostly focus on the individual teacher expectation effect. The studies that examine teacher expectations on the basis of classroom/group level (de Jong et al., 2012; Demanet & van Houtte, 2012; Friedrich et al., 2015; Li & Rubie-Davies, 2017, Park & Byun, 2020; Rubie-Davies, 2006; Rubie-Davies et al., 2020; Timmermans & Rubie-Davies, 2018) have started to become widespread in recent years. Accordingly, the scale developed in the current study focuses on classroom/group-level teacher expectations. Group-level teacher expectation is measured by the general perception of teachers about the academic abilities of students in a group (Park & Byun, 2020). The TES developed in the current study can measure this general perception of classroom/group-level teacher expectation in a comprehensive and useful way. In addition, among teacher expectation scales in the international literature (Auwarter & Aruguete, 2008; Barriga et al., 2019; Regalla, 2013; Sweatt, 2000; Szumski & Karwowski, 2019; Tiedemann, 2000; van den Bergh et al., 2010), non-academic teacher expectation is often ignored. The TES developed in the current study takes into account the expectation of non-academic teachers as well as academic teacher expectations.

EFA and CFA are used to test the structural validity of TES. As a result of EFA, an item with a low factor load is excluded from the scale. Thus, a two-factor structure consisting of 36 items is obtained. The item content and factors show similarities with the studies of Chen et al. (2011). In this specified study, teacher expectation is defined in the form of teacher impression in schools regarding the potential academic and non-academic behavior of students. Especially with this study, and also with other teacher expectation studies (Barriga et al., 2019; Rubie-

Davies, 2015; Wang et al., 2021), consistently, the first factor of the scale is named ‘academic expectation’ and the second factor is named ‘non-academic expectation’.

Each item contained in TES shows a high level of connection to the corresponding factor. In order to determine whether the theoretically designed model has been verified with data, CFA has been conducted. The data obtained from CFA showed that the compliance indices of the two-factor structure in TES are sufficient. In order to determine the total score predictive power of the items in TES and to determine their level of distinctiveness, item analyses are performed. In the lower and upper groups of 27% of the scale items within the scope of item analyses, a statistically significant difference was found between the groups and the *t* value is found to be significant. The adjusted item total test correlation values of the items indicate that the scale has high item distinctiveness and high validity. The CR and AVE values of TES provide the desired conditions for convergent and divergent validity with combined reliability. In addition, TES's Cronbach's Alpha, McDonald's Omega and stratified Alpha coefficient results are also confirmed by combined reliability coefficients.

The data collected with the scale have internal consistency. It is concluded that the correlation value between the two factors is high with the entire TES score and that there is a suggestive relationship between them. This high correlation between academic and non-academic teacher expectation factors is in accordance with the teacher expectation literature. The results show that all factors and the scale measure a similar structure. The final form of the 36-items TES including only positive wordings is provided in [Appendix](#). As a result, TES reliability and validity proofs are presented, and TES is brought to the literature.

4.1. Implications

It can measure teacher expectations at the individual level with TES, and it can be more useful to measure teacher expectations at the group/class/school level as a whole. Along with TES, research can be conducted through standardized tests that measure the socio-psychological characteristics of students. Research examining the relationships between teacher expectations and various teacher qualities (such as self-esteem, teacher judgment, teacher enthusiasm, dedication, burnout, stereotypical thinking, prejudice, etc.) can be done using TES. Revealing the current state of TES, teacher expectation can also be used to determine which variables that uniform teacher expectation affects and which variables are affected. Conducting research, in which TES will be used, will be important in terms of contributing to the measurement power and purpose of the use of the scale.

Acknowledgments

This paper was produced from part of the first author's master's thesis prepared under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Kahramanmaraş Sütçü İmam University/ Directorate of Social Sciences Institute, 07.12.2020/49

Authorship Contribution Statement

Hasan Igde: Literature review, Investigation, Development of Data Collection Tool, Analysis, Visualization, Resources, and Writing the original draft. **Levent Yakar:** Methodology, Supervision, and Validation.

Orcid

Hasan Igde  <https://orcid.org/0000-0001-8857-6319>

Levent Yakar  <https://orcid.org/0000-0001-7856-6926>

REFERENCES

- Archambault, I., Janosz, M., & Chouinard, R. (2012). Teacher beliefs as predictors of adolescents' cognitive engagement and achievement in mathematics. *The Journal of Educational Research*, 105(5), 319-328. <https://doi.org/10.1080/00220671.2011.629694>
- Auwarter, A.E., & Aruguete, M.S. (2008). Effects of student gender and socioeconomic status on teacher perceptions. *The Journal of Educational Research*, 101(4), 242-246. <https://doi.org/10.3200/JOER.101.4.243-246>
- Babad, E., Inbar, J., & Rosenthal, R. (1982). Pygmalion, galatea, and the golem: Investigations of biased and unbiased teachers. *Journal of Educational Psychology*, 74(4), 459-474. <https://doi.org/10.1037/0022-0663.74.4.459>
- Barriga, C.A., Rodríguez, C., & Ferreira, R.A. (2019). Factors that bias teacher expectations: Findings from Chile. *Revista Latinoamericana de Psicología*, 51(3), 171-180. <https://doi.org/10.14349/rlp.2019.v51.n3.4>
- Brophy, J.E., & Good, T.L. (1970). Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology*, 61(5), 365-374. <https://doi.org/10.1037/h0029908>
- Büyüköztürk, Ş. (2015). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum [Data analysis handbook for social sciences, statistics, research design, SPSS applications and interpretation]* (28th ed.). Pegem Academy.
- Büyüköztürk, Ş., Kılıç-Çakmak, Erkan-Akgün, O., Karadeniz, Ş., & Demirel, F. (2019). *Eğitimde bilimsel araştırma yöntemleri [Scientific research methods in education]* (26th ed.). Pegem Academy.
- Chen, Y.-H., Thompson, M.S., Kromrey, J.D., & Chang, G.H. (2011). Relations of student perceptions of teacher oral feedback with teacher expectancies and student self-concept. *The Journal of Experimental Education*, 79, 452-477. <https://doi.org/10.1080/00220973.2010.547888>
- Cooper, H.M. (1979). Pygmalion grows up: A model for teacher expectation communication and performance influence. *Review of Educational Research*, 49(3), 389-410. <https://doi.org/10.3102/00346543049003389>
- Creswell, J.W. (2019). *A concise introduction to mixed methods research*. M. Sözbilir (Ed.). Pegem Academy.
- Creswell, J.W., & Plano Clark, V.L. (2018). *Karma yöntem araştırmaları, tasarımı ve yürütülmesi [Designing and conducting mixed methods research]*. Y. Dede, & S.B. Demir (Eds.). Anı Publication.
- Çokluk, O., Şekercioğlu, G., & Büyüköztürk, S. (2018). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları [Multivariate statistics SPSS and LISREL applications for social sciences]*. Pegem Academy.
- Darley, J.M., & Fazio, R.H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist*, 35(10), 867-881. <https://doi.org/10.1037/0003-066X.35.10.867>
- de Boer, H., Bosker, R.J., & van der Werf, M.P.C. (2010). Sustainability of teacher expectation bias effects on long-term student performance. *Journal of Educational Psychology*, 102(1), 168-179. <https://doi.org/10.1037/a0017289>
- de Jong, R.J., van Tartwijk, J., Verloop, N., Veldman, I., & Wubbels, T. (2012). Teachers' expectations of teacher-student interaction: Complementary and distinctive expectancy patterns. *Teaching and Teacher Education*, 28(7), 948-956. <https://doi.org/10.1016/j.tate.2012.04.009>

- Demanet, J., & van Houtte, M. (2012). Teachers' attitudes and students' opposition to school misconduct as a reaction to teachers' diminished effort and effect. *Teaching and Teacher Education*, 28, 860-869. <https://doi.org/10.1016/j.tate.2012.03.008>
- Denessen, E., Keller, A., van den Bergh, L., & van den Broek, P. (2020). Do teachers treat their students differently? An observational study on teacher-student interactions as a function of teacher expectations and student achievement. *Hindawi Education Research International*, 2020(2471956), 1-18. <https://doi.org/10.1155/2020/2471956>
- Diamond, J.B., Randolph, A., & Spillane, J.P. (2004). Teachers' expectations and sense of responsibility for student learning: The importance of race, class, and organizational habitus. *Anthropology & Education Quarterly*, 35(1), 75-98. <https://doi.org/10.1525/aeq.2004.35.1.75>
- Dusek, J.B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology*, 75(3), 327-346. <https://doi.org/10.1037/00220663.75.3.327>
- Eden, D., Geller, D., Gewirtz, A., Tenner, G.N., Inbar, I., Liberman, M., Pass, Y., Segev, I.S., & Shalit, M. (2000). Implanting pygmalion leadership style through workshop training: Seven field experiments. *Leadership Quarterly*, 11(2), 171-210. [https://doi.org/10.1016/S1048-9843\(00\)00042-4](https://doi.org/10.1016/S1048-9843(00)00042-4)
- Erçetin, Ş.Ş., Akbaşlı, S., & Baysülen, E. (2020). Managers' and teachers' expectations from students and students' perception of these expectations. *International Journal of Society Researches*, 16 (Education and Society Special Issue), 5941-5954. <https://doi.org/10.26466/opus.705266>
- Eryılmaz, A. (2013). Motivation and amotivation at school: Developing the scale of expectations from the teacher about class engagement. *Mehmet Akif Ersoy University Journal of Education Faculty*, 13(25), 1-18.
- Friedrich, A., Flunger, B., Nagengast, B., Jonkmann, K., & Trautwein, U. (2015). Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement. *Contemporary Educational Psychology*, 41, 1-12. <https://doi.org/10.1016/j.cedpsych.2014.10.006>
- Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50. <https://doi.org/10.1177/002224378101800104>
- Gentrup, S., Lorenz, G., Kristen, C., & Kogan, I. (2020). Self-fulfilling prophecies in the classroom: Teacher expectations, teacher feedback and student achievement. *Learning and Instruction*, 66, 1-17. <https://doi.org/10.1016/j.learninstruc.2019.101296>
- Gershenson, S., Holt, S.B., & Papageorge, N.W. (2016). Who believes in me? The effect of student-teacher demographic match on teacher expectations. *Economics of Education Review*, 52, 209-224. <https://doi.org/10.1016/j.econedurev.2016.03.002>
- Good, T.L. (1987). Two decades of research on teacher expectations: Findings and future directions. *Journal of Teacher Education*, 38(4), 32-47. <https://doi.org/10.1177/002248718703800406>
- Good, T.L., Sterzinger, N., & Lavigne, A. (2018). Expectation effects: Pygmalion and the initial 20 years of research. *Educational Research and Evaluation*, 24(3-5), 99-123. <https://doi.org/10.1080/13803611.2018.1548817>
- Gökdere, E. (2013). Management of teacher expectations. *Education and Society in The 21st Century*, 2(5), 179-183.
- Gregory, A., & Huang, F. (2013). It takes a village: The effects of 10th-grade college-going expectations of students, parents, and teachers four years later. *American Journal of Community Psychology*, 52, 41-55. <https://doi.org/10.1007/s10464-013-9575-5>

- Hinnant, J.B., O'Brien, M., & Ghazarian, S.R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*(3), 662-670. <https://doi.org/10.1037/a0014306>
- İgde, H. (2021). A review study on the phenomenon of teacher expectation. *The Journal of Turkish Educational Sciences (JTES), 19*(2), 1366-1390. <https://doi.org/10.37217/tebd.988678>
- Kline, P. (1994). *An easy guide to factor analysis*. Guilford Publications.
- Kline, R.B. (2019). Exploratory and confirmatory factor analysis. Y. Petscher, & C. Schatsschneider (Eds.), *Applied quantitative analysis in the social sciences* (pp.171-207). Routledge.
- Kuş, Z., & Çelikkaya, T. (2010). Social sciences teachers' expectations for effective social sciences teaching. *Van Yüzüncü Yıl University Journal of Education, 7*(2), 26-51.
- Li, Z. (2016). The magnitude of teacher expectation effects: Differences in students, teachers and contexts. *International Journal of Learning Teaching and Educational Research, 15*(2), 76-93. Access address: <http://ijlter.org/index.php/ijlter/article/view/612>
- Li, Z., & Rubie-Davies, C.M. (2017). Teachers matter: Expectancy effects in Chinese university English-as-a-foreign-language classrooms. *Studies in Higher Education, 42*, 2042-2060. <https://doi.org/10.1080/03075079.2015.1130692>
- Merton, R.K. (1948). The self-fulfilling prophecy. *The Antioch Review, 8*(2), 193-210. <https://doi.org/10.2307/4609267>
- Ministry of National Education, (2017). *Öğretmenlik mesleği genel yeterlikleri [General competencies for teaching profession]*. Directorate General for Teacher Training and Development. https://oygm.meb.gov.tr/meb_iys_dosyalar/2018_06/29111119_TeachersGeneralCompetencies.pdf
- Ministry of National Education, (2021, 21 January). *Okullar ve diğer kurumlar [Schools and other institutions]*. <http://www.meb.gov.tr/baglantilar/okullar/index.php?ILKODU=46>
- Ocak, G., & Park, F. (2019). Developing analytical thinking scale for high school students. *Afyon Kocatepe University Journal of Social Sciences, 22*(1), 49-68. <https://doi.org/10.32709/akusosbil.565699>
- Öztürk, B., Koç, G., & Tezel-Şahin, F. (2002). Behavioral differences of the classroom teachers about high and low expected students. *The Journal of Educational Sciences and Practice, 1*(2), 161-181.
- Papageorge, N.W., Gershenson, S., & Kang, K.M. (2019). Teacher expectations matter. *Review of Economics and Statistics, 102*(2), 234-251. https://doi.org/10.1162/rest_a_00838
- Park, J.H., & Byun, S.Y. (2020). Principal support, professional learning community, and group-level teacher expectations. *School Effectiveness and School Improvement, 32*(1), 1-23. <https://doi.org/10.1080/09243453.2020.1764061>
- Peterson, E.R., Rubie-Davies, C., Osborne, D., & Sibley, C. (2016). Teachers' explicit expectations and implicit prejudiced attitudes to educational achievement: relations with student achievement and the ethnic achievement gap. *Learning and Instruction, 42*, 123-140. <https://doi.org/10.1016/j.learninstruc.2016.01.010>
- Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22*(4), 375-385. <https://doi.org/10.1177/014662169802200407>
- Regalla, M. (2013). *Teacher expectations and students from low socioeconomic background: A perspective from Costa Rica* (ED540254). ERIC. <https://eric.ed.gov/?id=ED540254>
- Riley, T., & Ungerleider, C. (2012). Self-fulfilling prophecy: How teachers' attributions, expectations, and stereotypes influence the learning opportunities afforded Aboriginal students. *Canadian Journal of Education, 35*(2), 303-333. <http://hdl.handle.net/10072/47128>

- Rosenthal, R. (1974). *On the social psychology of the self-fulfilling prophecy: Further evidence for pygmalion effects and their mediating mechanisms*. MSS Modular Publications.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. Holt, Rinehart and Winston.
- Rubie, C.M. (2004). *Expecting the best: Instructional practices, teacher beliefs and student outcomes* (Publication No. 3129406) [Doctoral dissertation, Auckland University]. Auckland Campus Repository. <https://researchspace.auckland.ac.nz/bitstream/handle/2292/28/02whole.pdf>
- Rubie-Davies, C.M. (2015). *Becoming a high expectation teacher: Raising the bar*. Routledge. <https://doi.org/10.4324/9781315761251>
- Rubie-Davies, C.M. (2007). Classroom interactions: Exploring the practices of high and low expectation teachers. *British Journal of Educational Psychology*, 77(2), 289-306. <https://doi.org/10.1348/000709906X101601>
- Rubie-Davies, C.M. (2010). Teacher expectations and perceptions of student attributes: Is there a relationship? *British Journal of Educational Psychology*, 80(1), 121-135. <https://doi.org/10.1348/000709909X466334>
- Rubie-Davies, C.M. (2006). Teacher expectations and student self-perceptions: Exploring relationships. *Psychology in the Schools*, 43(5), 537-552. <https://doi.org/10.1002/pits.20169>
- Rubie-Davies, C.M., Flint, A., & McDonald, L.G. (2012). Teacher beliefs, teacher characteristics, and school contextual factors: What are the relationships? *British Journal of Educational Psychology*, 82, 270-288. <https://doi.org/10.1111/j.2044-8279.2011.02025.x>
- Rubie-Davies, C.M., Meissel, K., Alansari, M., Watson, P., Flint, A., & McDonald, L. (2020). Achievement and beliefs outcomes of students with high and low expectation teachers. *Social Psychology of Education*, 23(5), 1173-1201. <https://doi.org/10.1007/s11218-020-09574-y>
- Rubie-Davies, C.M., & Peterson, E.R. (2016). Relations between teachers' achievement over and underestimation, and students' beliefs for Maori and Pakeha students. *Contemporary Educational Psychology*, 47, 72-83. <https://doi.org/10.1016/j.cedpsych.2016.01.001>
- Saritepeci, M. (2018). Adaptation study of the achievement motivation scale based on value-expectancy theory. *International Journal of Education Science and Technology*, 4(1), 28-40.
- Sazak-Pınar, E., Çifçi-Tekinarslan, I., & Sucuoğlu, B. (2012). Assessing teachers' and mothers' expectancies for social skills of children with mental retardation. *Elementary Education Online*, 11(2), 353-368.
- Sweatt, S. (2000). *The relationship among teacher expectations, teacher attitudes toward the TAAS, and student achievement* (Publication No. 49945119) [Doctoral dissertation, University of North Texas]. UNT Digital Library. <https://digital.library.unt.edu/ark:/67531/metadc2691/>
- Szumski, G., & Karwowski, M. (2019). Exploring the pygmalion effect: The role of teacher expectations, academic self-concept, and class context in students' math achievement. *Contemporary Educational Psychology*, 59, 1-10. <https://doi.org/10.1016/j.cedpsych.2019.101787>
- Tabachnick, B.G., & Fidell, L.S. (2015). *Using multivariate analysis*. Harper Collins College Publishers.
- Tiedemann, J. (2000). Gender-related beliefs of teachers in elementary school mathematics. *Educational Studies in Mathematics*, 41(2), 191-207. <http://dx.doi.org/10.1023/A:1003953801526>

- Timmerman, M.E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220. <https://doi.org/10.1037/a0023353>
- Timmermans, A.C., & Rubie-Davies, C.M. (2018). Do teachers differ in the level of expectations or in the extent to which they differentiate in expectations? Relations between teacher-level expectations, teacher background and beliefs, and subsequent student performance. *Educational Research and Evaluation, 24*(3-5), 241-263. <https://doi.org/10.1080/13803611.2018.1550837>
- Tutkun, C., & Dinçer, C. (2015). An examination on expectations regarding the social skills considered critical for preschoolers' success. *Ankara University Journal of Faculty of Educational Sciences, 48*(1), 65-85.
- van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R.W. (2010). The implicit prejudice attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal, 47*(2), 497-527. <https://doi.org/10.3102/0002831209353594>
- Wang, S., Meissel, K., & Rubie-Davies, C.M. (2021). Teacher expectation effects in Chinese junior high schools: Exploring links between teacher expectations and student achievement using a hierarchical linear modelling approach. *Social Psychology of Education, 24*(5), 1305-1333. <https://doi.org/10.1007/s11218-021-09654-7>
- Wang, S., Rubie-Davies, C.M., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educational Research and Evaluation, 24*(3-5), 124-179. <https://doi.org/10.1080/13803611.2018.1548798>
- Watson, P.W.S.J., Rubie-Davies, C.M., Meissel, K., Peterson, E.R., Flint, A., Garrett, L., & McDonald, L. (2017). Teacher gender, and expectation of reading achievement in New Zealand elementary school students: Essentially a barrier? *Gender and Education, 31*(8), 1-20. <https://doi.org/10.1080/09540253.2017.1410108>
- Wheaton, B., Muthen, B., Alwin, D.F., & Summers, G.F. (1977). Assessing reliability and stability in panel models. *Sociological Methodology, 8*, 84-136. <https://doi.org/10.2307/270754>
- Yaşlıoğlu, M.M. (2017). Factor analysis and validity in social sciences: Application of exploratory and confirmatory factor analyses, *Istanbul University Journal of the School of Business, 46*(Special Issue), 74-85.
- Yıldırım, A., & Şimşek, H. (2013). *Sosyal bilimlerde nitel araştırma yöntemleri [Qualitative research methods in the social sciences]*. (9th Extended Ed.). Seçkin Publisher.
- Yurtal, F., & Yontar, A. (2006). Sınıf öğretmenlerinin öğrencilerinden bekledikleri sorumluluklar ve sorumluluk kazandırmada kullandıkları yöntemler [The responsibilities that classroom teachers expect from their students and the methods they use to gain responsibility]. *Çukurova University Journal of Social Sciences Institute, 15*(2), 411-424.
- Yüksel, S. (2017). *The effect of improving teacher expectancy strategies on the attitudes and academic achievement of students in English course (An example of action research)* [Unpublished master's thesis]. University of Düzce.
- Zhu, M., Urhahne, D., & Rubie-Davies, C.M. (2018). The longitudinal effects of teacher judgement and different teacher treatment on students' academic outcomes. *Educational Psychology, 38*, 648-668. <https://doi.org/10.1080/01443410.2017.1412399>

APPENDIX

Teacher Expectation Scale (TES)'s Turkish version

Öğretmen Beklentisi Ölçeği (ÖBÖ)						
Madde No	MADDELER Derslerine girdiğim öğrencilerin,	Kesinlikle katılmıyorum	Çoğunlukla katılmıyorum	Orta düzeyde katılıyorum	Çoğunlukla Katılıyorum	Kesinlikle katılıyorum
2	Başarılı olmak için hedefler belirleyeceklerini düşünüyorum.	1	2	3	4	5
3	Başarı düzeylerinin yüksek olacağını düşünüyorum.	1	2	3	4	5
4	Başarı hedeflerine ulaşacaklarını düşünüyorum.	1	2	3	4	5
5	Akademik özgüvene sahip olacaklarını düşünüyorum.	1	2	3	4	5
6	Derslerine karşı ilgili olacaklarını düşünüyorum.	1	2	3	4	5
7	Derslerle ilgili sorumluluklarını yerine getireceklerini düşünüyorum.	1	2	3	4	5
8	Ders çalışma motivasyonlarının yüksek olacağını düşünüyorum.	1	2	3	4	5
9	Ders içeriklerini öğreneceklerini düşünüyorum.	1	2	3	4	5
10	Derslere aktif katılım göstereceklerini düşünüyorum.	1	2	3	4	5
11	Ders konusunda sorularına doğru yanıtlar vereceklerini düşünüyorum.	1	2	3	4	5
12	Derslerde etkili sorular soracaklarını düşünüyorum.	1	2	3	4	5
13	Ders programında yer alan öğrenme kazanımlarını edineceklerini düşünüyorum.	1	2	3	4	5
14	Derslerde öğrendiklerini hayatlarına yansıtacaklarını düşünüyorum.	1	2	3	4	5
15	Gelişim dönemlerine uygun hazırbuluşluğa sahip olacaklarını düşünüyorum.	1	2	3	4	5
16	Üst kademe öğrenimlerine hazır olacaklarını düşünüyorum.	1	2	3	4	5
17	Yeteneklerini keşfedeceklerini düşünüyorum.	1	2	3	4	5
18	Kişisel gelişimlerine önem vereceklerini düşünüyorum.	1	2	3	4	5
19	Etkili iletişim kuracaklarını düşünüyorum.	1	2	3	4	5
20	Türkçeyi dil kurallarına uygun kullanacaklarını düşünüyorum.	1	2	3	4	5
21	Ahlaki erdemlere sahip olacaklarını düşünüyorum.	1	2	3	4	5
22	Karakterli bireyler olacaklarını düşünüyorum.	1	2	3	4	5
23	Kişisel bakımlarına özen göstereceklerini düşünüyorum.	1	2	3	4	5
24	Gelişim dönemlerine uygun olumlu davranışlar göstereceklerini düşünüyorum.	1	2	3	4	5
25	Nezaket kurallarına uyacaklarını düşünüyorum.	1	2	3	4	5
26	Empatik yaklaşımlar göstereceklerini düşünüyorum.	1	2	3	4	5
27	Topluma faydalı bireyler olacaklarını düşünüyorum.	1	2	3	4	5
28	Toplumun kendisinden beklediği davranışları benimseyeceklerini düşünüyorum.	1	2	3	4	5
29	Toplumsal olaylara duyarlı bireyler olacaklarını düşünüyorum.	1	2	3	4	5
30	Doğal çevreyi korumaya duyarlı olacaklarını düşünüyorum.	1	2	3	4	5
31	Milli değerlere önem vereceklerini düşünüyorum.	1	2	3	4	5
32	Manevi değerlere saygılı olacaklarını düşünüyorum.	1	2	3	4	5
33	Çevrelerine saygılı bireyler olacaklarını düşünüyorum.	1	2	3	4	5
34	Aileleriyle olumlu ilişkiler kuracaklarını düşünüyorum.	1	2	3	4	5
35	Arkadaş seçimlerine dikkat edeceklerini düşünüyorum.	1	2	3	4	5
36	Yüksek statülü mesleklere sahip olacaklarını düşünüyorum.	1	2	3	4	5