



Volume 8

Issue 3

2021

***International Journal of
Assessment Tools in Education***

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

<http://www.ijate.net>

e-ISSN: 2148-7456

© IJATE 2021





e-ISSN 2148-7456

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

Volume 8

Issue 3

2021

Publisher : Dr. İzzet KARA
International Journal of Assessment Tools in Education
&
Pamukkale University,
Education Faculty,
Department of Mathematic and Science Education,
20070, Denizli, Turkey

Phone : +90 258 296 1036
Fax : +90 258 296 1200
E-mail : ijate.editor@gmail.com

Frequency : 4 issues per year (March, June, September, December)

Online ISSN: 2148-7456

Website : <http://www.ijate.net/>

Website : <http://dergipark.org.tr/en/pub/ijate>

Cover Design: IJATE

Support Contact: Dr. İzzet KARA
(Journal Manager & Founding Editor)

Phone : +90 258 296 1036
Fax : +90 258 296 1200
E-mail : ikara@pau.edu.tr

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) is an international, peer-reviewed online journal. IJATE is aimed to receive manuscripts focusing on evaluation and assessment in education. It is expected that submitted manuscripts could direct national and international argumentations in the area. Both qualitative and quantitative studies can be accepted, however, it should be considered that all manuscripts need to focus on assessment and evaluation in education.

IJATE as an online journal is sponsored and hosted by **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.

Starting from this issue, the abbreviation for *International Journal of Assessment Tools in Education* is "***Int. J. Assess. Tools Educ.***" has been changed.

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. Eren Can AYBEK](#), *Pamukkale University, Turkey*

Editorial Board

[Dr. Beyza AKSU DUNYA](#), *Bartın University, Turkey*

[Dr. Selahattin GELBAL](#), *Hacettepe University, Turkey*

[Dr. Stanislav AVSEC](#), *University of Ljubljana, Slovenia*

[Dr. Murat BALKIS](#), *Pamukkale University, Turkey*

[Dr. Gulsah BASOL](#), *Gaziosmanpaşa University, Turkey*

[Dr. Bengü BORKAN](#), *Boğaziçi University, Turkey*

[Dr. Kelly D. BRADLEY](#), *University of Kentucky, United States*

[Dr. Okan BULUT](#), *University of Alberta, Canada*

[Dr. Javier Fombona CADAUIECO](#), *University of Oviedo, Spain*

[Dr. William W. COBERN](#), *Western Michigan University, United States*

[Dr. R. Nukhet CIKRIKCI](#), *İstanbul Aydın University, Turkey*

[Dr. Safiye Bilican DEMİR](#), *Kocaeli University, Turkey*

[Dr. Nuri DOGAN](#), *Hacettepe University, Turkey*

[Dr. R. Sahin ARSLAN](#), *Pamukkale University, Turkey*

[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, Turkey*

[Dr. Yasemin KAYA](#), *Atatürk University, Turkey*

[Dr. Hulya KELECIOGLU](#), *Hacettepe University, Turkey*

[Dr. Hakan KOGAR](#), *Akdeniz University, Turkey*

[Dr. Omer KUTLU](#), *Ankara University, Turkey*

[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 Universit, China*

[Dr. Nesrin OZTURK](#), *Izmir Democracy University, Turkey*

[Dr. Turan PAKER](#), *Pamukkale University, Turkey*

[Dr. Murat Dogan SAHIN](#), *Anadolu University, Turkey*

[Dr. Hossein SALARIAN](#), *University of Tehran, Iran*

[Dr. Ragıp TERZI](#), *Harran University, Turkey*

[Dr. Hakan TURKMEN](#), *Ege University, Turkey*

[Dr. Ozen YILDIRIM](#), *Pamukkale University, Turkey*

English Language Editors

[Dr. Hatice ALTUN](#) - *Pamukkale University, Turkey*

[Dr. Arzu KANAT MUTLUOGLU](#) - *Ted University, Turkey*

Editorial Assistant

[Anil KANDEMİR](#) - *Middle East Technical University, Turkey*

CONTENTS

Research Articles

[Development of the Hostility in Pandemic Scale \(HPS\): A Validity and Reliability Study /](#)
Pages: 475-486 [PDF](#)

Emine Burcu TUNÇ, Simel PARLAK, Müge ULUMAN, Derya ERYİĞİT

[Examining the Invariance of a Measurement Model of Teachers' Awareness and Exposure Levels to Nanoscience by Using the Covariance Structure Approach /](#) Pages: 487-508 [PDF](#)

Şeref TAN, Zeki IPEK, Ali Derya ATİK, Figen ERKOÇ

[The Learning Effect of Corpora on Strong and Weak Collocations: Implications for Corpus-Based Assessment of Collocation Competence /](#) Pages: 509-526 [PDF](#)

Hatice ALTUN

[The Continuity of Students' Disengaged Responding in Low-stakes Assessments: Evidence from Response Times /](#) Pages: 527-541 [PDF](#)

Hatice Cigdem BULUT

[The Use of Exploratory Graph Analysis to Validate Trust in Relationships Scale /](#) Pages: 542-552 [PDF](#)

Akif AVCU

[Determination of cyber accessibility of teacher made tests/exams /](#) Pages: 553-569 [PDF](#)

Gülden ÖZDEMİR, Atilla ÖZDEMİR, Selahattin GELBAL

[Investigation of Measurement Invariance of State Test Anxiety Scale /](#) Pages: 570-582 [PDF](#)

Hüseyin SELVİ

[Comparison of G and Phi coefficients estimated in generalizability theory with real cases /](#)
Pages: 583-595 [PDF](#)

Kaan Zulfikar DENİZ, Emel ILICAN

[Uncovering the Reasons of EFL Teachers' Unwillingness and Demotivation towards Being More Assessment Literate /](#) Pages: 596-612 [PDF](#)

Elçin ÖLMEZER ÖZTÜRK

[The Unit Testlet Dilemma: PISA Sample /](#) Pages: 613-632 [PDF](#)

Cansu AYAN, Fulya BARIŞ PEKMEZCİ

[Investigation of Measurement Invariance According to Home Resources: TIMSS 2015 Mathematical Affective Characteristics Questionnaire /](#) Pages: 633-648 [PDF](#)

Derya ÇAKICI ESER

[Examination of Common Exams Held by Measurement and Assessment Centers: Many Facet Rasch Analysis /](#) Pages: 649-666 [PDF](#)

Gülden KAYA UYANIK, Tuğba DEMİRTAŞ TOLAMAN, Duygu GÜR ERDOĞAN

[Adaptation of Statistics Anxiety Scale to Turkish: Validity and Reliability Study /](#) Pages: 667-683 [PDF](#)

İsmail DURAK, Yalçın KARAGÖZ

[Validity Evidence for the Perceptions of Secondary School Students of 'What Research is' Scale and Measurement Invariance /](#) Pages: 684-703 [PDF](#)

Nurullah ERYILMAZ

[MonteCarloSEM: An R Package to Simulate Data for SEM](#) / Pages: 704-713 [PDF](#)

Fatih ORÇAN

[Investigating Invariant Item Ordering in Intelligence Tests: Mokken Scale Analysis of KBIT-2](#)

/ Pages: 714-728 [PDF](#)

Eren Halil ÖZBERK, Elif Bengi ÜNSAL ÖZBERK, Sait ULUÇ, Ferhunde ÖKTEM

Development of the Hostility in Pandemic Scale (HPS): A Validity and Reliability Study

Emine Burcu Tunc^{1,*}, Simel Parlak², Muge Uluman¹, Derya Eryigit¹

¹Department of Educational Sciences, Faculty of Education, Marmara University, Istanbul, Turkey

²Department of Educational Sciences, Faculty of Education, Istanbul Okan University, Istanbul, Turkey

ARTICLE HISTORY

Received: Dec. 08, 2020

Revised: Jan. 13, 2021

Accepted: May 16, 2021

Keywords:

Hostility in pandemic,
Pandemic,
Hostility,
Scale development,
Validity.

Abstract: The aim of this research is to develop Hostility in Pandemic Scale (HPS) for Turkey Population to determine the hostility levels of individuals, which is a factor affecting the mental well-being of the society during the pandemic. The study group consists of 855 individuals between the ages of 18-65 from different genders, and have experienced the pandemic process. For the construct validity of the scale results, exploratory factor analysis was conducted and a one-dimensional structure consisting of 22 items was revealed. It was determined that the variance explained by the scale showing a one-dimensional structure was 41.5%. As a result of the confirmatory factor analysis performed through a separate study group, it was revealed that all items have significant t values, and the model established according to model fit indexes has meaningful and acceptable fit values. Buss-Perry Aggression Scale was applied with HPS for the criterion validity. As a result of the criterion validity analysis, a significant relationship was found between the scale scores. The Cronbach Alpha was calculated to analyses internal consistency of the scale and a reliability level of 0.93 was obtained. The test-retest reliability results were found as 0.89. In addition, item statistics revealed that all of the scale items can discriminate well among the respondents. Results of the analysis revealed that, the Hostility Scale in Pandemic Process provides valid and reliable results.

1. INTRODUCTION

Pandemic triggers changes in the psychological and sociological structure of the society. Therefore, understanding the epidemiology of the pandemic and defining the changes occurring in the societies undergoing the pandemic process is necessary to guide not only the current pandemic, but also the repetitive waves of the same virus and public health responses in future pandemics (Trauer et al., 2011). During the pandemic, individuals might face post-traumatic stress disorder (Lee et al., 2018), stress, anxiety, depressive symptoms, rejection, fear and anger (Jones et al., 2017). Negative effects on psychological well-being in the society may lead to the development of hostile feelings and actions regarding the emergence and spread of the virus.

Hostility is a complex set of tendencies that includes negative beliefs, angry emotions, and aggressive interactions (Spilberger et al., 1983), but can also be seen as a transient state (Rosenman, 1991) or a stable personality trait (Miller et al., 1996). Although closely related,

*CONTACT: Emine Burcu TUNÇ ✉ burcu.tunc@marmara.edu.tr 📧 Department of Educational Sciences, Faculty of Education, Marmara University, Istanbul, Turkey

hostility appears to be more precisely differentiated from both anger and aggression, as it combines attitudinal and cognitive characteristics (Gambone, 1999). Considering the literature on hostility, it can be stated that the focus is primarily on the link between personality and negative health outcomes (Becker & Lesiak, 1977; Faay et al., 2020; Keith et al., 2017; Ranchor et al., 1997). It is also recognized that social-environmental conditions and genetics are important dynamics in the formation and maintenance of hostility (Contrada, 1994). Hostility, which is a negative attitude, often causes people to experience anger. The individual, who has a hostile attitude, experiences a negative and pessimistic view of the world, distrust towards other people and a desire to harm. These individuals generally worry about problems and cannot cope with uncertainty (Eckhardt, Bradley & Deffenbacher, 2004). In addition, these individuals who have difficulty reading social cues display aggressive behavior in their social interactions (Suls & Bunde 2005). Individuals who have a hostile attitude experience stress due to having a pessimistic perspective, anger and aggression. This stress negatively affects the mental and social lives of individuals, who have a hostile attitude, and causes health problems (Maan Diong et al., 2005).

Hostility is related with many outcomes in the social health and it is important that these relationships could be analyzed statistically. Thus, there are several assessment tools in the literature to be used in the studies to assess hostility with different variables. Buss and Durke (1957) developed an inventory in order to assess different kinds of hostility, such as: Assault, Indirect Hostility, Irritability, Negativism, Resentment, Suspicion, Verbal Hostility, and Guilt. In their scale, Cook and Medley (1957) associated hostility with enhanced risk for physical disorders, psychological dysfunction, and problems in interpersonal relationships. Xenophobia Scale by Veer et al., (2013) is developed to assess the xenophobia; hostility towards people from a different country, ethnic or cultural group. This scale is adapted to Turkish by Özmete, Yıldırım and Duru (2012). Bussy and Perry (1992) developed Hostility scale which is the most common used hostility scales and is adapted to Turkish by Madran. (2012). Although there are several hostility scales developed in the literature, there is not a scale developed about hostility in pandemic, which still changes the society.

In the study carried out by Becerra-García et al., (2020) with 151 participants between the ages of 18-76, it was found that individuals between the ages of 18-35 have higher rates of hostility. Similarly, in the study conducted by Pérez-Fuentes et al., (2020) with 1004 participants, it was found that the perceived threat from Covid-19 has a direct positive effect on sadness-depression, anxiety and anger-hostility moods, and that anxiety and anger-hostility directly affect the perception of threat from the virus.

Considering that Covid-19 creates many psychological and sociological problems in the individual such as panic, anxiety and hostility; it is clear that revealing the psychological aspects of the fight against Covid-19 will contribute to the social mental health (Jakovljevic et al., 2020). Hostility can manifest in emerging and invisible ways in behavior, and the presence of hostility is a variable that affects how the society will go through the pandemic process. For the aforementioned reasons, in this study, it was aimed to develop the Hostility Scale in the Pandemic Process to determine the hostility levels exhibited by individuals during the pandemic.

2. METHOD

In this study, it was aimed to develop a measurement tool for determining the hostility levels of individuals during the pandemic. This research is a scale development study. The information about the study group and the processes followed in the development of the Hostility in the Pandemic Scale (HPS) are given below.

2.1. Study Group

The study group of this research consists of 855 individuals between the ages of 18-65. In this context, the values in Table 1 were reached by removing outliers at each stage, and analyzes were carried out on a total of 855 individuals. The necessary ethical approval is obtained before the study and the informed consent of the participants were obtained before the application of the scale.

Table 1. Working groups included in the study.

Study groups	Applied scale / scales	Performed statistical transactions	Number of individuals
First Study Group	HPS	Application of EFA for construct validity and testing internal consistency	370 individuals
Second Study Group	HPS	Application of CFA for construct validity	353 individuals
Third Study Group	Aggression Scale with HPS	Calculating the relationship between the scores of two scales for criterion validity	75 individuals
Fourth Study Group	HPS	Calculation of the relationship between the first and second applications for test-retest reliability	57 individuals

The first study group consisted of 370 individuals after the outliers were removed. Individuals between the ages of 18-65 were reached. 29% ($n= 108$) of the group are men and 71% ($n= 262$) are women. Among the main study groups, 353 individuals were reached for the second study group. 20% of the group are men ($n= 71$) and 80% ($n= 282$) are women. Another 75 individuals were part of the study group for criterion validity and 57 individuals for test-retest reliability.

2.2. Development Process of the Scale

Firstly, the related literature on the pandemic process and the concept of hostility was reviewed. Based on the literature review, the general framework of the concept of hostility, considering the points that the pandemic and hostility are compatible with, the expressions that can be included in the scale have been examined and discussed. As a result, 39 items in total were written by the researchers. The scale items were evaluated by a total of six experts, two assessment and evaluation experts, two guidance and psychological counseling experts, and two Turkish language experts. In accordance with the opinions of the experts, items that are difficult to understand, that are irrelevant with the subject, and that contain more than one jurisdiction have been revised or removed from the scale in line with the recommendations. In this direction, the final trial form consisting of 35 items was obtained. The scale is a five-point Likert-type, rated as 1. *Strongly disagree*, 2. *Disagree*, 3. *Undecided*, 4. *Agree*, and 5. *Strongly agree*.

In the first step, the Explanatory Factor Analysis (EFA) was applied to reveal the construct validity of the scale. At this stage, factor loadings were mainly taken into consideration while deciding on the items that should be included in the scale. According to Tabachnick & Fidell (2007), and Kline (2011), factor loadings should be at least 0.32. Therefore, 0.32 was accepted as the criterion value for the factor loadings in the current study. The Cronbach Alpha reliability coefficient was used to test the internal consistency of the results obtained from the scale whose construct validity was proven. Crocker & Algina (1986) and Tan (2009) stated that the reliability coefficients in the range of 0.70-0.80 are acceptable. In this study, this criterion was taken into account for internal consistency.

In the second stage, the Confirmatory Factor Analysis (CFA) was conducted to test the accuracy of the structure revealed in the first stage. The CFA is used in examinations to test a model

developed by the researcher in line with this theory (Tavşancıl, 2009). Therefore, the one-dimensional structure revealed by the EFA was examined based on the CFA. The fact that all t-values in a measurement model are meaningful, indicates that the items in the model are compatible with the model and should be included in the scale (Byrne, 2010). However, as a criterion of whether the measurement model is an acceptable model as a whole, fit index values should also be examined (Şimşek, 2007). In this study, after conducting the EFA, the fit index values provided by the CFA were examined emphasized and the construct validity of the scale was tried to be proven.

In the third stage, Spearman's rho correlation coefficient between HPS and the results obtained from the Buss & Perry Aggression Scale was examined within the scope of criterion-based validity. Normality assumption could not be achieved therefore spearman correlation was used (for two scale, Kolmogorov-Smirnov test, $p < 0.05$; $p = 0.00$). The Buss & Perry Aggression Scale was developed by Buss and Perry (1992) and adapted to Turkish culture by Demirtaş Madran (2012). The scale consists consisting of 29 items with a five-point Likert type and four sub-dimensions: physical aggression, verbal aggression, hostility and anger. As a result of the validity and reliability analysis of the Turkish form, it was revealed that the scale provides reliable and valid results. In the fourth stage, for test and re-test reliability, HPS was administreted on the same group at two weeks interval and significant correlation between two sets of results were found after Pearson correlation analysis. Normality assumption could be achieved therefore Pearson correlation was used (for two sets of results, Kolmogorov-Smirnov test, $p > 0.05$; $p = 0.22$).

3. FINDINGS

In this section, findings obtained from the validity and reliability studies of the Hostility in Pandemic Scale (HPS) have been included.

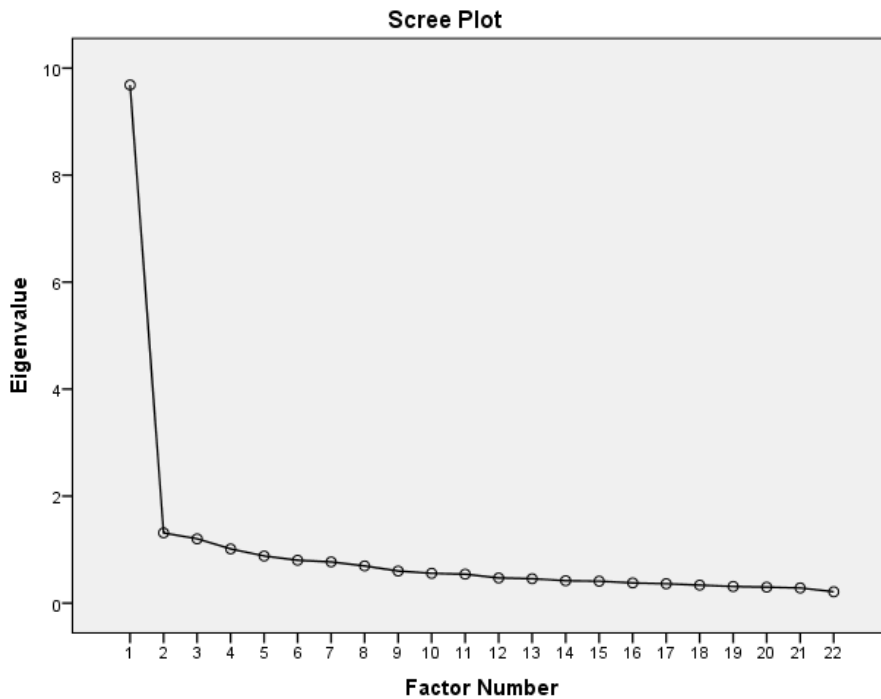
3.1. Structural Validity

To analyze the construct validity of the scale, the EFA and CFA were conducted on the data obtained from applying the scale on the study group.

3.1.1. *The exploratory factor analysis (EFA)*

After removing the outliers clearing the extreme values, analyzes were carried out on 370 individuals. The fact that the KMO value is 0.95 and the Barlett Sphericity Test result is significant ($\chi^2 = 3806.79$, $df = 231$) shows that the data is suitable for factor analysis. As a result of the Principal Axis Factoring technique in the EFA, items with a factor loading of less than 0.32 were removed from the initial 35 items. With the remaining 22 items, it was determined that a single-factor structure that explains 41.5% of the total variance emerged and this single-factor structure was also suitable for theoretical explanations. As seen in Figure 1 scree plot is the proof of unidimensionality. Çokluk et al., (2012) state that the variance explained 30% in one-dimensional structures in social sciences is sufficient. Therefore, it has been revealed that the variance explained by the developed scale is also quite sufficient.

Figure 1. Scree plot.



Findings obtained from the EFA are presented in Table 2. According to the results in Table 2, it can be seen that all scale items have factor loadings above the lower limit of 0.32. It was also revealed that the scale items met criterion value of 0.20 for the explained common variance.

Table 2. Factor structure of the scale and factor loadings.

Item no	Factor loading	Common variance
I1	0.77	0.59
I2	0.75	0.56
I3	0.74	0.54
I4	0.73	0.52
I5	0.72	0.52
I6	0.71	0.50
I7	0.69	0.48
I8	0.68	0.47
I9	0.68	0.47
I10	0.66	0.44
I11	0.66	0.44
I12	0.66	0.43
I13	0.65	0.42
I14	0.60	0.36
I15	0.57	0.32
I16	0.57	0.32
I17	0.56	0.31
I18	0.56	0.31
I19	0.55	0.30
I20	0.55	0.30
I21	0.54	0.29
I22	0.45	0.20

3.1.2. The confirmatory factor analysis (CFA)

The study group for CFA consists of 353 individuals. It was tested whether the data collected from the second study group confirmed the structure consisting of 22 items and one factor obtained as a result of the EFA. Some of the modifications recommended by the CFA were made to achieve better fit indices. The modifications that were applied include the identification of error covariances among items I1-I4, MI-I13, I17-I18, I20-I21, I13-I11, I15-I3, I8-I2 and I20-I22. Table 3 shows perfect and acceptable fit criteria for fit indices.

Table 3. Perfect and acceptable fit values for fit indices and fit index values obtained from CFA.

Reviewed indices of fit	Perfect fit criteria	Acceptable fit criteria	Achieved fit indexes	Conclusion
χ^2/sd	$0 \leq \chi^2 / sd \leq 2$	$2 \leq \chi^2 / sd \leq 3$	2.87	Acceptable
GFI	$.95 \leq GFI \leq 1.00$	$.90 \leq GFI \leq .95$	0.90	Acceptable
AGFI	$.90 \leq AGFI \leq 1.00$	$.85 \leq AGFI \leq .90$	0.84	Acceptable
CFI	$.95 \leq CFI \leq 1.00$	$.90 \leq CFI \leq .95$	0.92	Acceptable
NFI	$.95 \leq NFI \leq 1.00$	$.90 \leq NFI \leq .95$	0.88	Acceptable
NNFI	$.95 \leq NNFI \leq 1.00$	$.90 \leq NNFI \leq .95$	0.90	Acceptable
IFI	$.95 \leq IFI \leq 1.00$	$.90 \leq IFI \leq .95$	0.92	Acceptable
RMSEA	$.00 \leq RMSEA \leq .05$	$.05 \leq RMSEA \leq .08$	0.073	Acceptable
SRMR	$.00 \leq SRMR \leq .05$	$.05 \leq SRMR \leq .10$	0.050	Perfect
PNFI	$.95 \leq PNFI \leq 1.00$	$.50 \leq PNFI \leq .95$	0.75	Acceptable
PGFI	$.95 \leq PGFI \leq 1.00$	$.50 \leq PGFI \leq .95$	0.68	Acceptable

$\chi^2_{nd} = 569.89$, $df = 198$, 90% Probability Confidence Interval for RMSEA = (0.066; 0.080)

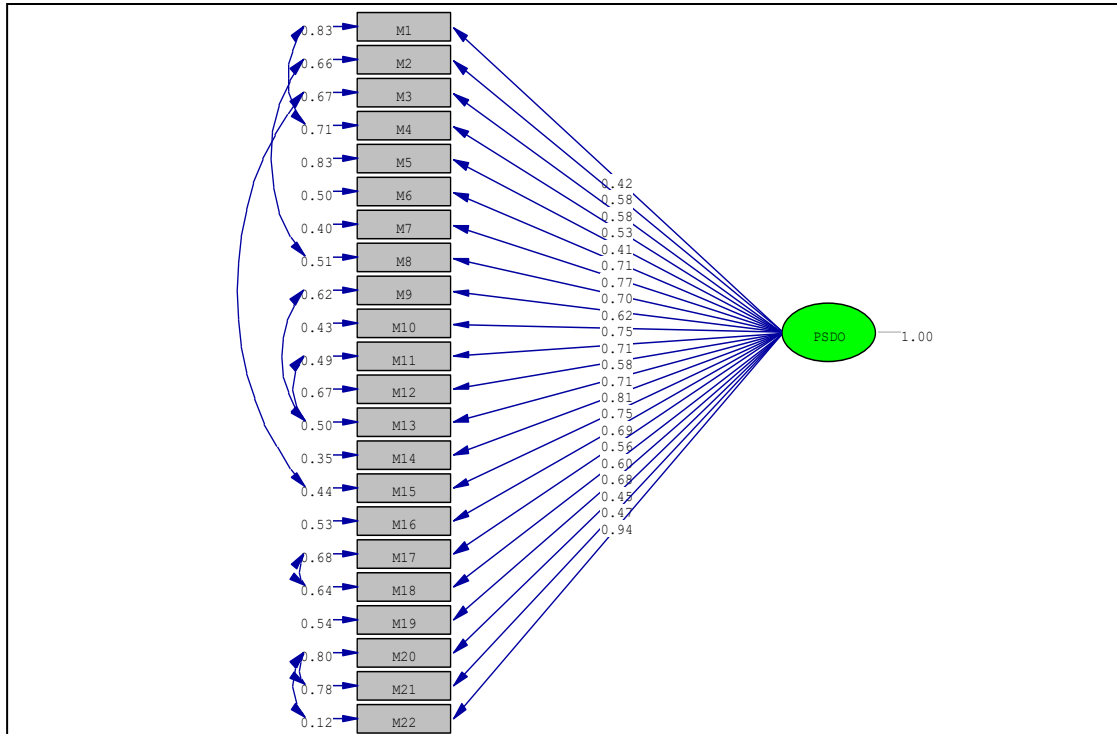
It was demonstrated with these values that the level of fit of the model obtained from the CFA is sufficient. The t values provided by the CFA are given in Table 4. It was determined that the t-values for the items were between 7.93 and 22.73. The t values greater than 1.96 and 2.58 are meaningful at the .05 and .01 levels, respectively (Kline, 2011).

Table 4. t values obtained from CFA.

Item no	t value
I1	8.03
I2	11.77
I3	11.64
I4	10.72
I5	7.93
I6	15.24
I7	17.14
I8	15.08
I9	12.81
I10	16.57
I11	14.91
I12	11.79
I13	15.27
I14	18.20
I15	16.20
I16	13.78
I17	11.33
I18	12.27
I19	14.23
I20	8.27
I21	8.24
I22	22.73

Therefore, it was determined that all of the t values are meaningful, and all items should be included in the scale. The factor loads for the one-dimensional model obtained are given in Figure 2. As seen in Figure 2, factor loadings vary between 0.41 and 0.94.

Figure 2. Measurement model for the scale.



3.2. The Criterion Validity

For the criterion-based validity, the HPS and Buss& Perry Aggression scale was applied to 75 participants. Spearman's rho correlation analysis was performed to determine the relationship between the results obtained from the two scales and a significant relationship was found ($r=0.41$, $p=0.00$, $p<0.01$). This result shows that there is a positive relationship between the results obtained from the two scales. This result evidences that the scale can provide valid results.

3.3. Reliability

The reliability of the scale was examined based on Cronbach Alpha and test-retest methods. Considering that the measurement results with a reliability coefficient of 0.70 and above are reliable (Crocker & Algina, 1986), it has been revealed that the calculated 0.93 Cronbach Alpha reliability coefficient is quite high.

The test-retest method was used as the second proof of the reliability of the scale results. The scale, consisting of 22 items, was applied twice with an interval of two weeks and the consistency between the two applications was examined. There was a high level and significant relationship between the two applications with $r=0.89$ ($p=0.00$, $p<0.01$). This result shows that there is agreement between the results obtained from the two applications and there is evidence that the second reliability condition is met.

3.4. Item Statistics

In order to determine the discrimination levels of the items and to determine the predictive power of the total score, corrected item-total correlations were calculated and 27% sub-upper group comparisons were included. The findings obtained as a result of item analysis are shown in Table 5.

Table 5. Results of item analysis.

Item no	Average	Standard deviation	Corrected item-total correlation	When the item is removed scale Alpha	t
I1	2.25	1.25	.698	.927	16.32
I2	2.14	1.24	.698	.927	17.10
I3	1.77	1.03	.693	.928	13.94
I4	1.85	1.03	.691	.928	13.40
I5	2.49	1.34	.693	.927	17.38
I6	2.69	1.37	.680	.928	19.96
I7	2.64	1.36	.661	.928	15.81
I8	2.29	1.31	.662	.928	14.69
I9	2.65	1.28	.656	.928	15.34
I10	1.75	1.10	.635	.929	12.40
I11	2.14	1.17	.624	.929	14.36
I12	3.09	1.27	.630	.929	14.92
I13	2.91	1.38	.586	.929	15.87
I14	1.85	1.11	.560	.930	11.16
I15	1.83	1.11	.554	.930	11.55
I16	2.08	1.25	.538	.930	13.97
I17	3.01	1.34	.551	.930	10.73
I18	2.18	1.29	.531	.930	11.46
I19	3.36	1.26	.514	.931	10.34
I20	1.66	1.01	.511	.930	10.40
I21	2.05	1.39	.460	.932	7.23
I22	1.65	1.06	.380	.932	10.54

When the table is examined, it was determined that the t values ($df=198, p<0.01$) regarding the differences in item scores of the 27% lower and upper groups were significant. Item-total score correlations vary between 0.38 and 0.70. Items with item-total score correlations over 0.30 are considered discriminating. All of these findings reveal that the items are discriminatory.

3.5. Evaluation of Scores Obtained from the Scale

There are 22 items in the scale and there is no reverse item. The scale is a five-point Likert-type as; "Strongly Agree (5), Agree (4), Undecided (3), Disagree (2), and Strongly Disagree (1)". The scale has a one-dimensional structure. The total score is obtained from the scale, and the higher the scores mean the higher the hostility perceptions of the individuals during the pandemic.

4. DISCUSSION and CONCLUSION

Covid 19 outbreak threatens mental health as well as physical health. Mental health deterioration and the stress experienced increase the feelings of hostility in the individual. According to Siegman & Smith (1994), hostility is defined as a negative attitude towards others and especially the feeling of anger. With the Covid-19 outbreak, it is observed that there is social insecurity among people, and this increases hostility (Kim, 2020). This hostility may also be against foreigners or some ethnic groups (Bartos et al., 2020). In the statement published by the World Health Organization (2020) on January 30, 2020 regarding this negative change in the social sense, it was emphasized that countries should be careful against stigmatization and discrimination in the fight against Covid-19. In addition to its social impact, it is seen that hostility in interpersonal relationships increases during the pandemic (Pietromonaco & Overall2020). Research conducted with 3233 participants in China reveals that individuals with higher stress levels and using negative coping strategies and show more hostility (Duan et al.,

2020). Thus, in this study, it was aimed to develop a measurement tool for determining the hostility levels of individuals during the pandemic.

EFA and CFA were applied to test the construct validity of the scale results. According to EFA results, the factor loads of the items in the scale should be at least 0.32 (Kline 2011, Tabachnick & Fidell 2007). As a result of EFA, items with insufficient factor loading were removed from the scale and a 22-item scale was created. Tabachnick & Fidell (2001) and Şencan (2005) stated that the common variance is at least 0.20. It has been revealed that all items in the scale contribute more than 0.20 to the common variance.

As a result of EFA, a single factor structure that explains 41.5% of the total variance with 22 items emerged. Çokluk et al., (2012) stated that the variance explained 30% in one-dimensional structures in social sciences is sufficient. Therefore, it has been revealed that the variance explained by the developed scale is also quite sufficient.

The findings obtained from CFA applied to test whether the structure consisting of 22 items and a single factor obtained as a result of EFA was verified or not, showed that the fit indices of the model were sufficient. In addition, it was revealed that all t values obtained as a result of CFA are meaningful. Byrne (2010) and Şimşek (2007) stated that all t-values in a measurement model are meaningful, the items in the model are compatible with the model and should be included in the scale. Therefore, CFA revealed that all items are necessary for the scale.

Buss& Perry Aggression Scale were used for Criterion Validity. A significant relationship was found in the Spearman's rho correlation analysis conducted to determine the relationship between the results obtained from the two scales. This result shows that there is a positive relationship between the results obtained from the two scales and there is evidence that the criterion validity is provided.

Cronbach Alpha and test-retest method were used to test the reliability of HPS. Cronbach Alpha reliability coefficient was found as 0.93 and test-retest reliability was found 0.89 as the second proof of the reliability of the scale results. Crocker & Algina (1986) and Tan (2009) stated that the reliability coefficients in the range of 0.70-0.80 are acceptable. The results obtained prove that the reliability of the scale results is high.

In order to determine the distinctiveness levels of the items in HPS, and to determine the predictive power of the total score, corrected item-total correlations were calculated, and 27% sub-upper group comparisons were included. When interpreting the item-total score correlation, items with a value above 0.30 are considered sufficient to distinguish the feature to be measured. The significance of the t values for the differences between the 27% sub-upper group is also considered as evidence for the distinctiveness of the items (Erkuş, 2012). As a result of the analysis, it was found that item-total score correlations were ranked between 0.38 and 0.70, and t-values are significant for all items. These findings reveal that the items are distinctive. As a result of all these analyzes, it was determined that HPS is a valid and reliable measurement tool in revealing hostility.

In a study conducted with 1014 people in Spain, it is revealed that threat perception originating from Covid-19 causes negative emotions such as depression, anxiety, anger, and hostility (Pérez et al., 2020). Similarly, when the items were examined in the study conducted, it was observed that anger and negative attitude towards other people are significant and item loads are high. This draws attention to the importance of taking into account the hostility felt during the pandemic in terms of both society and individual health and reveals the necessity of conducting studies to reduce the feelings of stress and hostility by strengthening social support networks. Hence it is thought, this scale will be very useful in terms of examining the factors affecting the mental well-being of the society and increasing the studies supporting well-being.

When the literature is examined, it is seen that there are measurement tools such as The Cook-Medley Inventory, The Buss-Durkee Inventory, The Hostility and Direction of Hostility Questionnaire, The NEO-Personality Inventory-Revised, The Rorschach Inkblot Test. However, no scale was found regarding the hostility experienced during the Covid-19 outbreak and thus this study is important in regard of this. The strength of the study is that there is more than one evidence for the validity and reliability of the scale and at the same time a large study group of 855 individuals has been reached. However, the fact that the majority of the individuals in the study group consisted of women is considered as a limitation of the study.

Acknowledgments

We would like to thank all of the participants who participated in this study for their support.

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 author(s). For the study, 56665618-204.01.07 numbered Ethical Board approval was taken from İstanbul Okan University Ethical Board.

Authorship Contribution Statement

Emine Burcu Tunc: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing - original draft. **Simel Parlak:** Investigation, Resources, Visualization, Software, Formal Analysis, and Writing -original draft. **Muge Uluman:** Methodology, Supervision, and Validation. Authors may edit this part based on their case. **Derya Eryigit:** Methodology, Supervision, and Validation. Authors may edit this part based on their case.

ORCID

Emine Burcu Tunc  <https://orcid.org/0000-0002-8225-9299>

Simel Parlak  <https://orcid.org/0000-0002-8651-2693>

Muge Uluman  <https://orcid.org/0000-0003-4155-3114>

Derya Eryigit  <https://orcid.org/0000-0002-3708-7176>

5. REFERENCES

- Bartos, V., Bauer, M., Cahlikova, J., Chytilová, J. (2020). Covid-19 Crisis Fuels Hostility Against Foreigners. *CESifo Working Paper* No. 8309, Available at SSRN: <https://ssrn.com/abstract=3618833>
- Becerra-García, J.A., Giménez Ballesta, G., Sánchez-Gutiérrez, T., Barbeito Resa, S., Calvo Calvo, A. (2020). Psychopathological symptoms during Covid-19 quarantine in spanish general population: a preliminary analysis based on sociodemographic and occupational-contextual factors. *Revista Espanola de Salud Publica*, 94:e202006059. <https://doi.org/10.1093/geronb/gbaa074>
- Becker, E. W., & Lesiak, W. J. (1977). Feelings of hostility and personal control as related to depression. *Journal of Clinical Psychology*, 33(3), 654-657. <https://doi.org/10.1002/1097-4679>
- Buss, A. H., & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*, 21(4), 343-349. <https://doi.org/10.1037/h0046900>
- Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of personality and social psychology*, 63(3), 452. <https://doi.org/10.1037/0022-3514.63.3.452>
- Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications and programming*. Taylor and Francis Group.
- Contrada, R. J. (1994). Personality and anger in cardiovascular disease: Toward a psychological model. In A. Siegman, & T. Smith (Eds.), *Anger, hostility, and the heart* (pp. 149-170). Erlbaum.

- Cook, W. W. & Medley, D. M. (1954) Proposed hostility and Pharisaeic-virtue scales for the MMPI. *Journal of Applied Psychology*, 38, 414-418.
- Crocker, L., and Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth: Holt, Rinehart and Winston Inc.
- Çokluk, Ö., Şekercioğlu, G., Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları [Multivariate statistics for social sciences: SPSS and LISREL applications]*. Pegem Akademi Yayıncılık.
- Demirtaş-Madran, H. A. (2012). Buss-Perry saldırganlık ölçeği'nin Türkçe formunun geçerlik ve güvenilirlik çalışması [Reliability and Validity of the Buss-Perry Aggression Questionnaire-Turkish Version]. *Türk Psikoloji Dergisi*, 24(2), 1-6.
- Duan, H., Yan, L., Ding, X., Gan, Y., Kohn, N., & Wu, J. (2020). Impact of the COVID-19 pandemic on mental health in the general Chinese population: Changes, predictors and psychosocial correlates. *Psychiatry Research*, 293, 113396. <https://doi.org/10.1016/j.psychres.2020.113396>
- World Health Organization (2020). *Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)*. [https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-\(2005\)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-\(2019-ncov\)](https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov))
- Eckhardt, C., Norlander, B., & Deffenbacher, J. (2004). The assessment of anger and hostility: A critical review. *Aggression and Violent Behavior*, 9(1), 17-43. [https://doi.org/10.1016/S1359-1789\(02\)00116-7](https://doi.org/10.1016/S1359-1789(02)00116-7)
- Erkuş, A. (2012). *Psikolojide ölçme ve ölçek geliştirme [Measurement and scale development in psychology]*. Pegem Akademi Yayınları.
- Faay, M. D., Van Baal, G. C. M., Arango, C., Díaz-Caneja, C. M., Berger, G., Leucht, S., ... & Petter, J. (2020). Hostility and aggressive behaviour in first episode psychosis: Results from the OPTiMiSE trial. *Schizophrenia Research*, 223, 271-278. <https://doi.org/10.1016/j.schres.2020.08.021>
- Fraenkel, J R, Wallend, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. McGraw Hill.
- Gambone, G. C. (1998). *Cognitive patterns of self- and other-representation as indicators of hostility* [Unpublished doctoral dissertation]. The State University of New Jersey-New Brunswick.
- Jakovljevic, M., Bjedov, S., Jaksic, N., & Jakovljevic, I. (2020). COVID-19 pandemia and public and global mental health from the perspective of global health security. *Psychiatria Danubina*, 32(1), 6-14. <https://doi.org/10.24869/psyd.2020.6>
- Jones, N. M., Thompson, R. R., Schetter, C. D., & Silver, R. C. (2017). Distress and rumor exposure on social media during a campus lockdown. *Proceedings of the National Academy of Sciences*, 114(44), 11663-11668. <https://doi.org/10.1073/pnas.1708518114>
- Keith, F., Krantz, D. S., Chen, R., Harris, K. M., Ware, C. M., Lee, A. K., ... & Gottlieb, S. S. (2017). Anger, hostility, and hospitalizations in patients with heart failure. *Health Psychology*, 36(9), 829. <https://doi.org/10.1037/hea0000519>
- Kim, B. (2020). Effects of social grooming on incivility in COVID-19. *Cyberpsychology, Behavior, and Social Networking*, 23(8), 519-525. <https://doi.org/10.1089/cyber.2020.0201>
- Kline, R. B. (2011). *Principles and practice of structural equation modeling*. The Guilford Press.
- Lee, S. M., Kang, W. S., Cho, A. R., Kim, T., & Park, J. K. (2018). Psychological impact of the 2015 MERS outbreak on hospital workers and quarantined hemodialysis patients.

- Comprehensive Psychiatry*, 87, 123-127. <https://doi.org/10.1016/j.comppsy.2018.10.003>
- Maan Diong, S., Bishop, G. D., Enkelmann, H. C., Tong, E. M., Why, Y. P., Ang, J. C., & Khader, M. (2005). Anger, stress, coping, social support and health: Modelling the relationships. *Psychology & Health*, 20(4), 467-495. <https://doi.org/10.1080/0887044040512331333960>
- Miller, T. Q., Smith, T. W., Turner, C. W., Guijarro, M. L., & Hallet, A. J. (1996). Meta-analytic review of research on hostility and physical health. *Psychological Bulletin*, 119(2), 322. <https://doi.org/10.1037/0033-2909.119.2.322>
- Özmete, E., Yildirim, H., & Duru, S. (2018). Yabancı düşmanlığı (zenofobi) ölçeğinin Türk kültürüne uyarlanması: Geçerlik ve güvenilirlik çalışması [Adaptation of the scale of xenophobia to Turkish culture: Validity and reliability study]. *Sosyal Politika Çalışmaları Dergisi*, 18, 191-209. <https://doi.org/10.21560/spcd.v18i39974.451063>
- Pérez-Fuentes, M. D. C., Molero Jurado, M. D. M., Martos Martínez, Á. & Gázquez Linares, J. J. (2020). Threat of COVID-19 and emotional state during quarantine: Positive and negative affect as mediators in a cross-sectional study of the Spanish population. *Plos one, Public Library of Science* 15(6), 1-11. <https://doi.org/10.1371/journal.pone.0235305>
- Pietromonaco, P. R., & Overall, N. C. (2020). Applying relationship science to evaluate how the COVID-19 pandemic may impact couples' relationships. *American Psychologist*. Advance online publication. <http://dx.doi.org/10.1037/amp0000714>
- Ranchor, A. V., Sanderman, R., Bauma, J., Buunk, B. P., & van den Heuvel, W. J. (1997). An exploration of the relation between hostility and disease. *Journal of Behavioral Medicine*, 20(3), 223-240. <https://doi.org/10.1023/A:1025538926879>
- Rosenman, R. H. (1991) Type A behavior pattern and coronary heart disease: The hostility factor?. *Stress Medicine*, 7(4), 245-253. <https://doi.org/10.1002/smi.2460070407>
- Siegmán, A. W. & Smith, T. W. (1994). *Anger, hostility, and the heart*. Lawrence Erlbaum Associates, Inc.
- Spielberger, C. D., Jacobs, G., Russell, S. & Crane, R. S. (1983). Assessment of anger: The state-trait anger scale. *Advances in Personality Assessment*, 2, 161-189.
- Suls, J. & Bunde, J. (2005). Anger, anxiety, and depression as risk factors for cardiovascular disease: the problems and implications of overlapping affective dispositions. *Psychological bulletin*, 131(2), 260.
- Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik [Reliability and validity in social and behavioral measures]*. Seçkin Yayıncılık.
- Şimşek, Ö. F. (2007). *Yapısal Eşitlik Modellemesine Giriş: Temel İlkeler ve LISREL Uygulamaları [Reliability and validity in social and behavioral measures]*. Ekinoks Yayıncılık.
- Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics*. Pearson Education, Inc.
- Tan, Ş. (2009). Misuses of KR-20 and Cronbach's alpha reliability coefficients. *Eğitim ve Bilim Dergisi*, 34, 152, 101-112.
- Tavşancıl, E. (2009). *Tutumların ölçülmesi ve SPSS ile veri analizi [Measuring attitudes and data analysis with SPSS]*. Nobel Yayın Dağıtım.
- Trauer, J. M., Laurie, K. L., McDonnell, J., Kelso, A., & Markey, P. G. (2011). Differential effects of pandemic (H1N1) 2009 on remote and indigenous groups, Northern Territory, Australia, 2009. *Emerging Infectious Diseases*, 17(9), 1615.
- Van der Veer, K., Ommundsen, R., Yakushko, O., Higler, L., Woelders, S., & Hagen, K. A. (2013). Psychometrically and qualitatively validating a cross-national cumulative measure of fear-based xenophobia. *Quality & Quantity*, 47(3), 1429-1444. <https://doi.org/10.1007/s11135-011-9599-6>

Examining the Invariance of a Measurement Model of Teachers' Awareness and Exposure Levels to Nanoscience by Using the Covariance Structure Approach

Seref Tan^{1,*}, Zeki Ipek², Ali Derya Atik³, Figen Erkoç⁴

¹Gazi University, Gazi Faculty of Education, Department of Measurement and Evaluation in Education, Ankara

²Republic of Turkey, Ministry of National Education, Antalya

³Kilis 7 Aralık University, Faculty of Education, Department of Mathematics and Science Education, Kilis

⁴Gazi University, Gazi Faculty of Education, Department of Biology Education, Ankara

ARTICLE HISTORY

Received: Nov. 19, 2020

Revised: Mar. 23, 2021

Accepted: May. 16, 2021

Keywords:

Configural invariance,
Factorial structure
invariance,
Structural covariance
invariance,
Measurement residual
invariance,
Invariance of
nanotechnology scale.

Abstract: The main aim of this study is to examine the measurement invariance of the structural equating model constructed on the Awareness and Exposure subscales of Nanoscience and Nanotechnology Awareness Scale (NSTAS) test for three teacher branches, three school types, and two genders by using the covariance structural analysis to test configural and metric invariances. The other aim of this study is showing how to use the IBM AMOS-24 software package with examples to address the issue of measurement invariance using the covariance structural analysis approach. Study sample was 1039 complete records gathered from science teachers with convenience sampling. Research data were collected in two stages. In the first stage, data were obtained from 624 teachers who participated to the study in the 2015-16 academic year. In the second stage, data were obtained in 2019 from 415 teachers via a link to access to the scale and all the instructions for the NSTAS in 2019. The covariance structures analysis was used to examine the measurement invariance of the scale. The comparative fit index was used to compare the measurement invariance in the measurement model. The study revealed that configural, measurement weight and structural covariance invariances were ensured for branches, school types and genders. Residual invariance was ensured only for gender. As a result, it was concluded that the NSTAS scale was not biased for teacher branches, school types or gender. NSTAS scale is recommended for the purposes of comparing branch, school type and gender groups.

1. INTRODUCTION

Nanoscience and nanotechnology (NSNT) are an abstract and complex topic with various applications resulting from the manipulation of atoms and molecules. Nanotechnology, one of the most promising technologies of the 21st century, utilizes devices, structures, and molecules on the scales of nanometers ranging between 1 and 100 nm (Bayda et al., 2020). The responsible development of nanotechnology that addresses the ethical, legal, and societal issues together with research, commercialization, worker education, and public engagement is assumed to

*CONTACT: Seref TAN ✉ sereftan4@yahoo.com 📧 Gazi University, Gazi Faculty of Education, Department of Measurement and Evaluation in Education, Ankara, Turkey

determine public trust and the future of innovation driven by NSNT. However, describing a world people cannot see and physically interact needs enhancement of understanding these emerging technologies using science communication/citizen-science to reach its full revolutionary potential. Public attitudes, and reflexive governance are essential to public acceptance of NSNT innovation (Boholm & Larsson, 2019).

It is one of the most rapidly growing/broad multidisciplinary fields in science, technology, life sciences and engineering research/innovation and is founded on the convergence of traditional disciplines to create, study, and apply materials at the nanoscale (Holland et al., 2018). Nanotechnology generates great opportunities for cutting-edge research in science and for innovation in industrial production and affects the everyday lives. Presently science teachers typically have insignificant exposure to NSNT, and few opportunities to understand the basic concepts. Developing countries must take their positions in the world nanotechnology market and industry, so planning for good NSNT training is especially important for developing countries. Depending on new information and how it is presented public attitudes toward NSNT may become unstable at times, show rapid change potential since attitudes depend on values, beliefs, and worldviews rather than on facts (Boholm & Larsson, 2019).

Developments and economic impact on commerce and society have brought nanotechnology education to the forefront. Along this line, developed countries have made NSNT education a priority, with intensive education planning and research at primary level being launched. The significance of awareness should be emphasized as an initial step in all nano education processes. The rapid development and impact of NSNT on economy has led policy makers and educators to focus on nanotechnology education (Laherto, 2010). Integrating a new multidisciplinary science at the interface of different scientific and engineering disciplines into the secondary school is a significant endeavor; however, it can be spread throughout a well-designed secondary science education curriculum. Furthermore, factors affecting awareness and knowledge level of teachers/teacher trainees in NSNT should be determined and analyzed before implementing education programs (Hingant & Albe, 2010; Jones et al., 2013). Communicating NSNT to different levels of students places the teacher at the center of learning and teaching activities for NSNT; a significant responsibility (Hingant & Able, 2010). If teachers are not familiar with NSNT, teaching these topics will be a major challenge for them (Greenberg, 2009). Therefore, teachers need to develop their own knowledge and awareness of NSNT to understand and be able to communicate these issues to their students (Blonder et al., 2014). The responsible development of NSNT to safeguard the environment, human health, and safety, and to ensure that the new technology benefits society, requires citizen involvement, dialog, and participation. These cannot be achieved without teacher education and training in NSNT.

It is provided in AERA, APA, and NCME (2014) as standards for evidence regarding internal structure, “if the rationale for a test score interpretation for a given use depends on premises about the relationship among test items or among parts of the test, evidence concerning the internal structure of the test should be provided.” Theoretical structure of a measuring tool raises the concern whether it works the same in different groups, when the differences between the groups are tested. Ensuring the measurement invariance of measuring tools is neglected in almost all research. As Millsap and Yun-Tein (2004) pointed out, the extension of the analysis to the multiple-population case is less well-known, especially for ordered-categorical data in the literature on factor analysis. As Camilli (2006) pointed out that measurement invariance contributes to validity evidence in that scores from a tool are subject to issues of bias and lack of fairness if invariance does not hold.

Whether the Nanoscience and Nanotechnology Awareness Scale (NSTAS), (İpek et al., 2020) measures the same characteristics for three different teacher branches, three school types, and

two genders are determined as sub-groups to test the measurement invariances. When different groups are to be compared, the obtained scores from the scale should not be biased (Tan & Pektaş, 2020). Further investigations are necessary to explain/justify the question of whether the scale items perform similarly across subgroups, and one way to examine this question is through assessing the measurement invariance of a scale (Chung et al., 2016). There are several studies in the literature on measurement invariance for test scores (Arana et al., 2018; Camerota et al., 2018). For a measurement model to have the same structure across different groups, the factor loadings of the items in a scale, and the correlations and variances among the identified factors, should be the same (Tan & Pektaş, 2020). While examining the measurement invariance of a measurement model between groups, the model created at each stage is built on the model created in the previous stage, i.e., the models are nested.

As stated by Byrne (2016, pp. 227-228), “In seeking evidence of multigroup equivalence, researchers are typically interested in finding the answer to one of five questions. First, do the items comprising a particular measuring instrument operate equivalently across different populations? In other words, is the measurement model group-invariant? Second is the factorial structure of a single instrument or of a theoretical construct equivalent across populations? Third, are certain paths in a specified causal structure equivalent across populations? Fourth are the latent means of constructs in a model different across populations? Finally, does the factorial structure of a measuring instrument replicate across independent samples drawn from the same population? This latter question addresses the issue of cross-validation.”

As Chung et al. (2016) stated, configural invariance is the fact that factor structures between groups are equivalent. In other words, configural invariance tests that the same pattern of item-factor loadings exists across groups compared, which requires that the same items have nonzero loadings on the same factors. To observe whether the other steps of invariance are ensured, comparisons are made based on the configural invariance values (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). On the other hand, metric invariance refers to equivalence among factor loadings. Chung et al. (2016) emphasized metric invariance, in addition to configural invariance, requires that unstandardized factor loadings be the same across groups. The scalar invariance is based on the equivalence of factor covariances across groups. Therefore, scalar invariance, addition to configural and metric invariance, factor variances and factor covariances are the same across groups. It is a kind of invariance where factor covariances are equalized across the groups after configural and metric invariances are ensured (Cheung & Rensvold, 2002; Meredith, 1993). Strict invariance requires proof that errors do not vary by group. Strict invariance, addition to configural, metric and scalar invariance, the error variances are the same across groups. It is a type of invariance where all factor loadings, factor variances, factor covariances and error variances are constrained (Cheung & Rensvold, 2002).

In the present study, the stages of identifying configural and metric measurement invariances of NSTAS were realized by using the covariance structural analysis (COVS) approach. In COVS approach of testing measurement invariances, only the variances and the covariances between paired observed variables are used as observed variables.

1.1. Aim of the Study

The very first step in nano education at any level is ensuring the awareness of the teachers (Bryan et al., 2012; Enil & Köseoğlu, 2016). The present study aimed to examine the measurement invariance of the structural equating model constructed on the Awareness and Exposure subscales of NSTAS test for three science teacher branches, three school types, and two genders by using the covariance structural analysis (COVS). In this study we also use the IBM AMOS-24 software package as illustrated with examples to address measurement invariances using the covariance structural analysis approach. This is a significant contribution

to the field of science education measurement and assessing since most of the measurement invariance studies are confined purely to the measurement field.

2. METHOD

2.1. The Research Model

This study is a descriptive study, as it is intended to present the present situation in terms of measurement invariance of NSTAS structural model and no variable is manipulated. Details of the scale have been published elsewhere (İpek et al., 2020).

2.2. The Study Group

The sample of the study consists of 1039 complete records (without any missing records) gathered from science teachers. Research data were collected in two stages. The data in the first stage were obtained from 624 teachers in the 2015-16 academic year, used in İpek's (2017) doctoral thesis.

Data in the second stage were obtained during 2019 by using a link to access the NSTAS scale and all instructions. In rare cases the scale was administered face-to-face to the respondents. Convenience sampling approach was used to form the study group. The distribution of the 1039 science teachers to the branches, school types and gender were as follows: Biology 38.5%; physics 31.5%, and chemistry 30.0%; science high school 16.3%, Anatolian high school 56.4% and vocational high school 27.3%; and male 45.4% and female 54.6%.

2.3. Data Collection Instruments

The *Nanotechnology Awareness Instrument* (NAI, Dyehouse et al., 2008, refer to [Appendix](#) for the instrument) was adapted into a Turkish version and named *Nanoscience and Nanotechnology Awareness Scale* (NSTAS, refer to [Appendix](#) for the scale); validity and reliability of the Turkish version were tested by the authors. The original scale (NAI) assessed changes in higher education student awareness, exposure, and motivation for nanotechnology, as well as factual knowledge about nanotechnology. The nanotechnology awareness subscale measures whether respondents “know something about nanotechnology” and whether they “have heard about nanotechnology and its applications”. Awareness is supported by exposure, where respondents’ previous exposure to nanotechnology may enhance their awareness and knowledge. NAI consisted of two parts: Items in Part A regarding awareness, exposure, and motivation subscales, and Part B regarding factual knowledge about nanotechnology (Dyehouse et al., 2008). Our version, the NSTAS, has three subscales, the *Awareness* (8 items) and *Exposure* (6 items) subscales adopted from NAI (total of 14 items), and the subscale *Knowledge* developed by the authors. The Awareness (8 items) and Exposure (6 items) subscales were used to perform measurement invariance analysis. The Cronbach alpha internal consistency coefficient of the Awareness (8 items) subscale was found to be .934 and Exposure subscale .845. Stratified alpha reliability coefficient for whole scale (with Awareness and Exposure, 14 items) was found to be .945.

2.4. Data Analysis

The covariance structural analysis approach was utilized to examine the measurement model invariances by sub-groups, explained above. The multivariate normal distribution assumption was tested for each subgroup. The multivariate normal distribution assumption was not met for any subgroup. Therefore, bootstrap estimation with 500 bootstrap samples was used to estimate the model parameters. In testing measurement invariances between the .01 reduction criterion the CFI value (Δ CFI) was used. Based on the conditions for ensuring measurement invariance, this has been accepted as proof for the presence of measurement invariance (Cheung &

Rensvold, 2002). Also, a difference of less than .01 in the Δ CFI index supports the less parameterized model (Chung et al., 2016).

During the analyses, the operations were done via the IBM AMOS-24 package program and explained as follows (Byrne, 2016):

IBM AMOS-24 operations for configural invariance.

1. The groups are defined by selecting the *Manage Groups* function from the *Analyze* menu in the *AMOS program*.
2. Subsequently, the data files are assigned to the defined groups by selecting the *Data Files* function from the *File* menu.
3. The *Emulisrel6* box is ticked by selecting *Estimation* from *Analysis Properties* in the *View* menu.
4. Finally, the analysis is run by selecting *Calculate Estimates* from the *Analyze* menu.

IBM AMOS-24 operations for configural, factor loading, structural variances and measurement residual invariances.

Until the stage of making the predictions, as an addition to the operations mentioned above, the parameters to be predicted in the model are labelled manually or automatically. For automatic labelling,

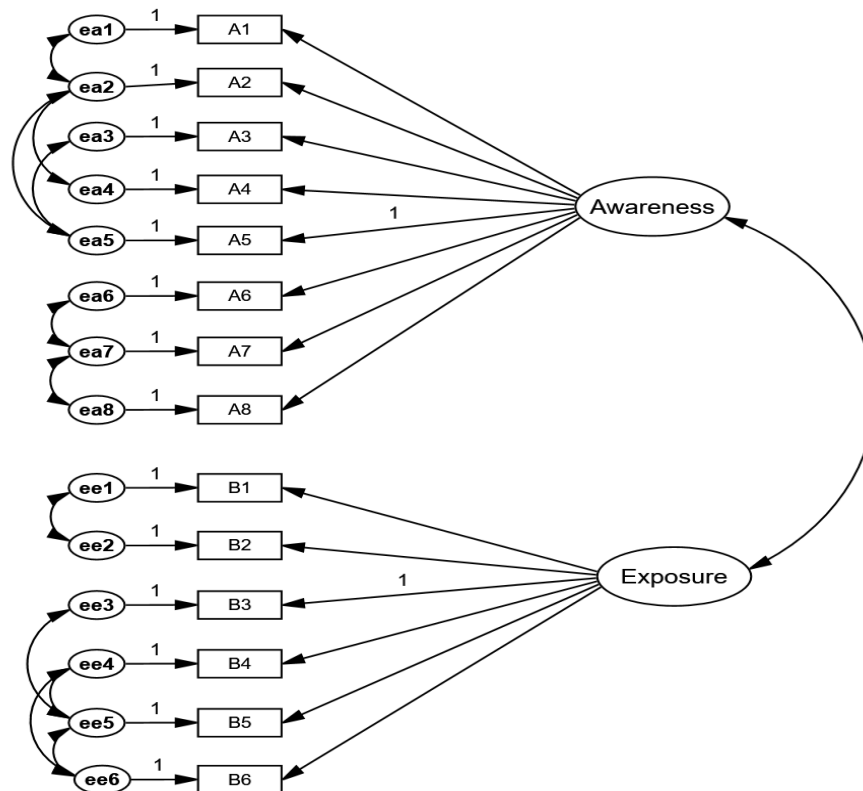
1. *Multiple Group Analysis* function is selected from the *Analyze* menu.
2. The parameters to be constrained are selected in the *Multiple-Group Analysis* dialog box.
3. The analysis is run by selecting *Calculate Estimates* from the *Analyze* menu.

3. RESULT / FINDINGS

3.1. Measurement Model

The baseline measurement model, which is used for eight subgroups, is presented in Figure 1, below.

Figure 1. The baseline measurement model for the multiple-group invariance of the NSTAS.



Chi square= \Cmin Df= \Df, GFI= \Gfi CFI= \Cfi, RMSEA= \RMSEA

As it seen in Figure 1, Awareness latent variable is measured with 8 items (A1 to A8) and Exposure latent variable is measured with 6 items (B1 to B6). There are covariance connections between the Awareness and Exposure latent variable and 11 covariance connections between some measurement residual variables in the baseline model. Item A5 was taken as reference for the scale of Awareness latent variable and item B3 for the scale of Exposure latent variable.

3.2. Measurement Invariance by Branch

The goodness of fit indices of the baseline measurement model used for all subgroups created within the scope of the study are presented below. Having good model fit indexes in all subgroups for the baseline measurement model is a prerequisite for invariance analysis.

Step 1: Goodness of Fit Indexes of the Baseline Measurement Model for Branch

The baseline model is presented in Figure 1. In the baseline measurement model based on the branches of teachers, the goodness of fit indexes (Schermelleh-Engel et al., 2003) were found as follows:

- ✓ for Physics teachers $\chi^2_{65}=215.097$; $\chi^2/sd=3.309$; GFI=0.916; CFI=0.959 and RMSEA=.084;
- ✓ for Chemistry teachers $\chi^2_{65}=175.102$; $\chi^2/sd=2.694$; GFI=0.927; CFI=0.964 and RMSEA=.074; and
- ✓ for Biology teachers $\chi^2_{65}=216.704$; $\chi^2/sd=3.334$; GFI=0.931; CFI=0.961 and RMSEA=.076.

In conclusion, the baseline measurement model in Figure 1 displayed a high level of model fit for Physics, Chemistry, and Biology teachers.

Step 2: Configural invariance of the Measurement Model for Branch

As stated by Byrne (2016), to ensure configural invariance, factor loading patterns and the number of factors should be similar for each group. The measurement model based on teachers' branch has provided configural invariance with $\chi^2_{195}=606.903$; $\chi^2/df=3.112$; GFI=.925; CFI=.961 and RMSEA=.045. That is, in this unconstrained measurement model, the factor structure for Physics, Chemistry, and Biology Teacher groups was found to be similar. These results show that the model in Figure 1 is a valid measurement model for all subgroups. The unstandardized estimated parameters (regression weights, covariances, and variances) of three branches for configural invariance are given for each group in Tables 1a, 1b and 1c, below.

Table 1a. Regression weight estimates for configural model.

Regression Weights			Estimates		
			Physics	Chemistry	Biology
A7	<---	Awareness	.812**	.624**	1.009**
A6	<---	Awareness	.857**	.710**	.924**
A5	<---	Awareness	1.000	1.000	1.000
A4	<---	Awareness	.984**	1.000**	1.033**
A3	<---	Awareness	.936**	.868**	1.068**
A2	<---	Awareness	.888**	.877**	.941**
A1	<---	Awareness	.959**	.973**	1.101**
B6	<---	Exposure	.402**	.385**	.221**
B5	<---	Exposure	.455**	.473**	.313**
B4	<---	Exposure	.620**	.619**	.450**
B3	<---	Exposure	1.000	1.000	1.000
B2	<---	Exposure	.830**	.820**	.875**
B1	<---	Exposure	.430**	.481**	.437**
A8	<---	Awareness	.921**	.911**	1.093**

*: $p < .05$; **: $p < .01$

Table 1b. Covariance estimates for configural model.

Covariance			Estimates		
			Physics	Chemistry	Biology
Awareness	<-->	Exposure	.897**	.650**	.670**
ea7	<-->	ea6	.412**	.438**	.310**
ee6	<-->	ee4	.564**	.565**	.608**
ea5	<-->	ea3	.039	.058	.203**
ee5	<-->	ee3	.093**	-.037	.003
ea4	<-->	ea2	.033	.096**	.113**
ea2	<-->	ea1	.073*	.080**	.114**
ea7	<-->	ea8	.004	.040	.040
ee6	<-->	ee5	.806**	.506**	.611**
ee5	<-->	ee4	.706**	.608**	.725**
ee2	<-->	ee1	.139**	.201**	.050
ea5	<-->	ea2	.046	.036	.030

*: $p < .05$; **: $p < .01$

Table 1c. Variance estimates for configural model.

Variances	Estimates		
	Physics	Chemistry	Biology
Awareness	1.172**	1.021**	.749**
Exposure	1.497**	1.570**	1.617**
ea7	.730**	.805**	.696**
ea6	.560**	.636**	.551**
ea5	.422**	.393**	.575**
ea4	.339**	.321**	.383**
ea3	.550**	.527**	.494**
ea2	.372**	.386**	.430**
ea1	.554**	.449**	.534**
ee6	1.056**	1.001**	.825**
ee5	1.070**	1.084**	.911**
ee4	1.193**	1.093**	1.142**
ee3	.477**	.450**	.469**
ee2	.336**	.415**	.254**
ee1	.530**	.442**	.426**
ea8	.522**	.432**	.514**

*: $p < .05$; **: $p < .01$

Step 3: Configural and Measurement Weights Invariance of the Measurement Model for Branch

As Byrne (2016) notes, in testing the measurement, structural and measurement error invariance, the focus is on the parameters, related to the measurement model, structural components and measurement errors, being equal in all groups. The measurement model based on teachers' branch has provided configural and *measurement weights invariance* with $X^2_{219}=654.437$; $X^2/df=2.988$; $GFI=.919$; $CFI=.959$ and $RMSEA=.044$. For testing the significant model differences, the CFI change value that we take the criteria was found to be less than .01 ($\Delta CFI=.002$). So, difference between configural invariance model and configural and measurement weights invariance model is not significant. In other words, the measurement model with restricted regression weights for Physics, Chemistry and Biology Teacher groups have been found to have good fit indexes with no significant CFI changes. So, measurement weights are equal for Physics, Chemistry, and Biology Teacher groups in the population.

The unstandardized estimated parameters (constrained regression weights, covariances, and variances) of three branches for configural and measurement weights invariance are given for each group in [Tables 2a](#), [2b](#), and [2c](#) below.

Table 2a. Regression weight estimates for configural and constrained measurement weights model.

Constrained Regression Weights			Estimates		
			Physics	Chemistry	Biology
A7	<---	Awareness	.815**		
A6	<---	Awareness	.839**		
A5	<---	Awareness	1.000		
A4	<---	Awareness	1.004**		
A3	<---	Awareness	.967**		
A2	<---	Awareness	.900**		
A1	<---	Awareness	1.010**		
B6	<---	Exposure	.318**		
B5	<---	Exposure	.397**		
B4	<---	Exposure	.549**		
B3	<---	Exposure	1.000		
B2	<---	Exposure	.845**		
B1	<---	Exposure	.454**		
A8	<---	Awareness	.968**		

*: $p < .05$; **: $p < .01$

Table 2b. Covariance estimates for configural and constrained measurement weights model.

Covariance			Estimates		
			Physics	Chemistry	Biology
Awareness	<-->	Exposure	.873**	.622**	.726**
ea7	<-->	ea6	.417**	.431**	.328**
ee6	<-->	ee4	.590**	.588**	.607**
ea5	<-->	ea3	.040	.053	.197**
ee5	<-->	ee3	.089**	-.025	.000
ea4	<-->	ea2	.034	.106**	.106**
ea2	<-->	ea1	.070*	.082**	.110**
ea7	<-->	ea8	.002	.031	.059*
ee6	<-->	ee5	.821**	.529**	.612**
ee5	<-->	ee4	.724**	.639**	.721**
ee2	<-->	ee1	.126**	.194**	.056
ea5	<-->	ea2	.050	.040	.026

*: $p < .05$; **: $p < .01$

Table 2c. Variance estimates for configural and constrained measurement weights model.

Variances	Estimates		
	Physics	Chemistry	Biology
Awareness	1.121**	.925**	.869**
Exposure	1.502**	1.563**	1.623**
ea7	.732**	.804**	.734**
ea6	.567**	.627**	.559**
ea5	.429**	.407**	.567**
ea4	.340**	.339**	.374**
ea3	.547**	.514**	.497**
ea2	.373**	.393**	.422**
ea1	.548**	.454**	.530**
ee6	1.080**	1.019**	.829**
ee5	1.079**	1.113**	.907**
ee4	1.221**	1.123**	1.137**
ee3	.481**	.470**	.454**
ee2	.318**	.387**	.285**
ee1	.521**	.445**	.425**
ea8	.520**	.428**	.528**

*: $p < .05$; **: $p < .01$

Step 4: Configural, Measurement Weight and Structural Covariance Invariance of the Measurement Model for Branch

The measurement model based on teachers' branch has provided configural, measurement weight, and *structural covariance invariance* with $X^2_{225}=667.589$; $X^2/df=2.967$; $GFI=.918$; $CFI=.958$ and $RMSEA=.044$. For testing the significant model differences, the CFI change value that we take the criteria was found to be less than .01 ($\Delta CFI=.003$). So, difference between configural invariance model and configural, measurement weight and structural covariance invariance model is not significant. In other words, the measurement model with constrained regression weights and structural covariances for Physics, Chemistry and Biology Teacher groups have good fit indexes with no significant CFI changes. So, measurement weights and structural covariances are equal for Physics, Chemistry, and Biology Teacher groups in the population.

The unstandardized estimated parameters (constrained regression weights, constrained structural covariances, other covariances and variances) of three branches for *Configural, Measurement Weights, and Structural Covariance Invariance* model are given for each group in [Tables 3a](#), [3b](#), and [3c](#) below.

In this model, since we have two structural variables (*Awareness* and *Exposure*), there is one structural covariance and two structural variances to be constrained additionally.

Table 3a. Regression weight estimates for configural, constrained measurement weights, and constrained structural covariances model.

Constrained Regression Weights			Estimates
			Physics Chemistry Biology
A7	<---	Awareness	.815**
A6	<---	Awareness	.838**
A5	<---	Awareness	1.000
A4	<---	Awareness	1.004**
A3	<---	Awareness	.967**
A2	<---	Awareness	.901**
A1	<---	Awareness	1.009**
B6	<---	Exposure	.318**
B5	<---	Exposure	.399**
B4	<---	Exposure	.550**
B3	<---	Exposure	1.000
B2	<---	Exposure	.846**
B1	<---	Exposure	.454**
A8	<---	Awareness	.968**

*: $p < .05$; **: $p < .01$

Table 3b. Covariance estimates for configural, constrained measurement weights, and constrained structural covariances model.

Covariance			Estimates		
			Physics	Chemistry	Biology
<i>Awareness</i>	<-->	<i>Exposure</i>	.742**	.742**	.742**
ea7	<-->	ea6	.420**	.429**	.326**
ee6	<-->	ee4	.594**	.585**	.607**
ea5	<-->	ea3	.041	.056	.195**
ee5	<-->	ee3	.089**	-.025	.001
ea4	<-->	ea2	.033	.105**	.105**
ea2	<-->	ea1	.069*	.082**	.109**
ea7	<-->	ea8	.003	.030	.058
ee6	<-->	ee5	.824**	.526**	.612**
ee5	<-->	ee4	.728**	.635**	.721**
ee2	<-->	ee1	.123**	.195**	.057
ea5	<-->	ea2	.049	.041	.026

*: $p < .05$; **: $p < .01$

Table 3c. Variance estimates for configural, constrained measurement weights, and constrained structural covariances model.

Variances	Estimates		
	Physics	Chemistry	Biology
<i>Awareness</i>	.967**	.967**	.967**
<i>Exposure</i>	1.562**	1.562**	1.562**
ea7	.735**	.802**	.731**
ea6	.570**	.624**	.558**
ea5	.429**	.412**	.565**
ea4	.339**	.339**	.374**
ea3	.548**	.515**	.494**
ea2	.371**	.392**	.421**
ea1	.546**	.457**	.528**
ee6	1.082**	1.016**	.829**
ee5	1.082**	1.110**	.908**
ee4	1.226**	1.119**	1.137**
ee3	.465**	.500**	.456**
ee2	.313**	.384**	.287**
ee1	.520**	.448**	.425**
ea8	.522**	.427**	.527**

*: $p < .05$; **: $p < .01$

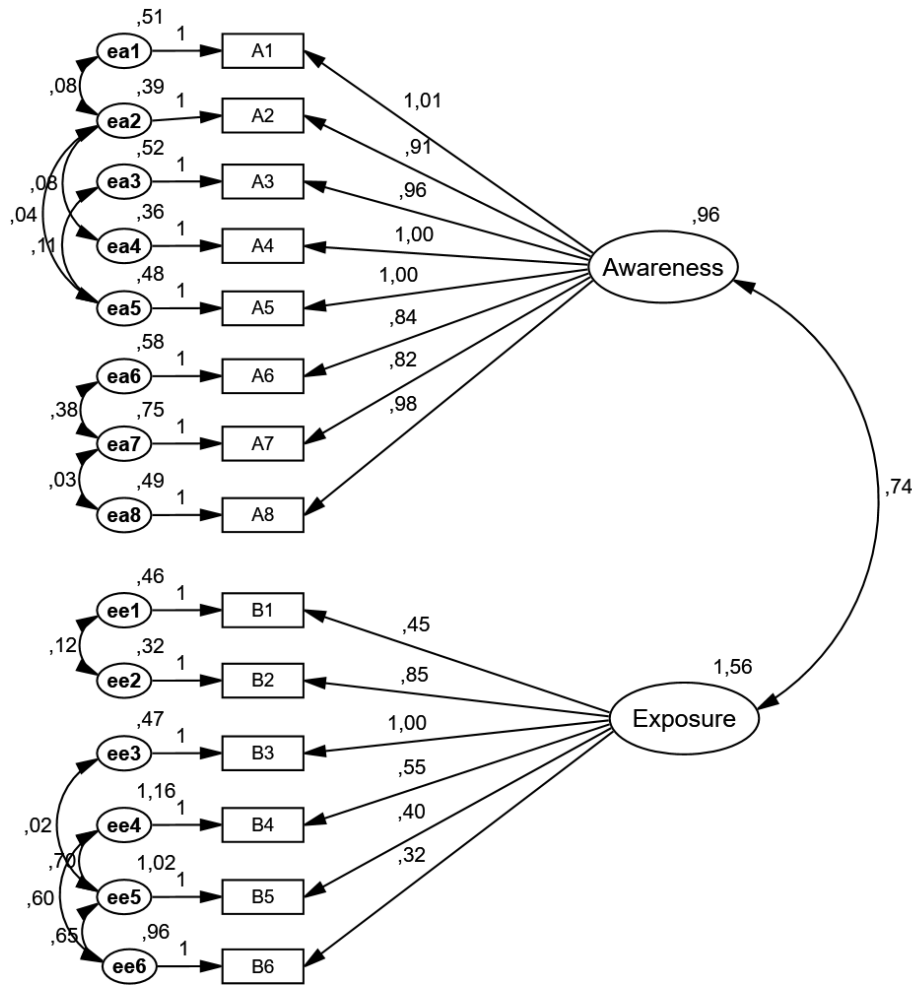
Step 5: Configural, Measurement Weight, Structural Covariance, and Measurement Residual Invariance of the Measurement Model for Branch

The goodness of fit indexes for this model were found to be good with $X^2_{275}=846.863$; $X^2/df=3.080$; $GFI=.895$; $CFI=.946$ and $RMSEA=.045$. However, for testing the significant model differences, the CFI change value was higher than .01 ($\Delta CFI=.015$). It is clear that, difference between configural invariance model and configural, measurement weight, structural covariance, and measurement residual invariance model is significant. Therefore, measurement residual estimates are not identical for Physics, Chemistry, and Biology Teacher groups in the population.

Because all the model parameters are constrained equal, the unstandardized estimated parameters of the model are given in the path diagram, [Figure 2](#), below.

The main findings regarding the measurement invariance according to the branches are presented in [Table 4](#) below. As can be observed in [Table 4](#), according to the unconstrained (configural) model, the changes in CFI in the models obtained by constraining, in sequence, measurement weights, and structural covariances were less than .01. However, when error residuals constrained the changes, CFI was found to be more than .01. Hence, it was concluded that the measurement model has provided configural, measurement weight, and structural covariance invariance; but did not provide measurement residual invariance across three branches.

Figure 2. Path diagram for configural, measurement weight, structural covariance, and measurement residual invariance of the measurement model for branch.



Chi square= 846,863 Df= 275, GFI= ,895 CFI= ,946, RMSEA= ,045

Note: Only 3 covariance estimates (ee5 <--> ee3=.017 with $p=.389$; ea7 <--> ea8=.032 with $p=.062$; and ea5 <--> ea2=.037 with $p=.013$) were not significant, all the other parameters were significant.

Table 4. Configural, measurement weight, structural covariance, and measurement residual invariance results by branch.

Model	Number of parameters	χ^2	df	χ^2/df	CFI	ΔCFI	RMSEA
1. Unconstrained (Configural)	120	606.903	195	3.112	.961		.045
2. Measurement Weights	96	654.437	219	2.988	.959	.002	.044
3. Structural Covariances	90	667.589	225	2.967	.958	.003	.044
4. Measurement Residuals	40	846.863	275	3.080	.946	.015	.045

Note: Unconstrained Model: All the parameters are predicted freely.
 Measurement Weights Model = All *Factor loadings* are constrained (equated).
 Structural Covariances Model = All *Factor loadings + factor variances and covariances* are constrained (equated).
 Measurement Errors Model = All *Factor loadings + factor variances + factor covariances + error variances* are constrained (equated).

3.3. Measurement Invariance by School Types

Goodness of Fit Indexes of the Baseline Measurement Model for School Type

In the baseline measurement model based on the school types, the goodness of fit indexes were found to be as follows:

- ✓ for science high school teachers $X^2_{65}=163.060$; $X^2/sd=2.509$; GFI=0.885; CFI=0.937 and RMSEA=.095;
- ✓ for Anatolian high school teachers $X^2_{65}=328.329$; $X^2/sd=5.051$; GFI=0.927; CFI=0.953 and RMSEA=.083; and
- ✓ for vocational high school teachers $X^2_{65}=224.257$; $X^2/sd=3.45$; GFI=0.906; CFI=0.947 and RMSEA=.093.

In conclusion, the baseline measurement model in [Figure 1](#) displayed a high level of model fit for three school types.

3.4. Configural, Measurement Weight, Structural Covariance, and Measurement Residual Invariance of the Measurement Model for School Type

The unstandardized estimated parameters of the model are given with path diagram for school types in [Figure 3](#), below, and the main findings regarding the measurement invariance according to the school types are presented in [Table 5](#) below.

Table 5. *Configural, measurement weight, structural covariance, and measurement residual invariance results by branch.*

Model	Number of parameters	X^2	df	X^2/df	CFI	ΔCFI	RMSEA
1. Unconstrained (Configural)	120	715.646	195	3.670	.949		.051
2. Measurement Weights	96	794.010	219	3.626	.943	.003	.050
3. Structural Covariances	90	820.660	225	3.647	.941	.005	.051
4. Measurement Residuals	40	1143.416	275	4.158	.914	.035	.055

Note: Unconstrained Model: All the parameters are predicted freely.

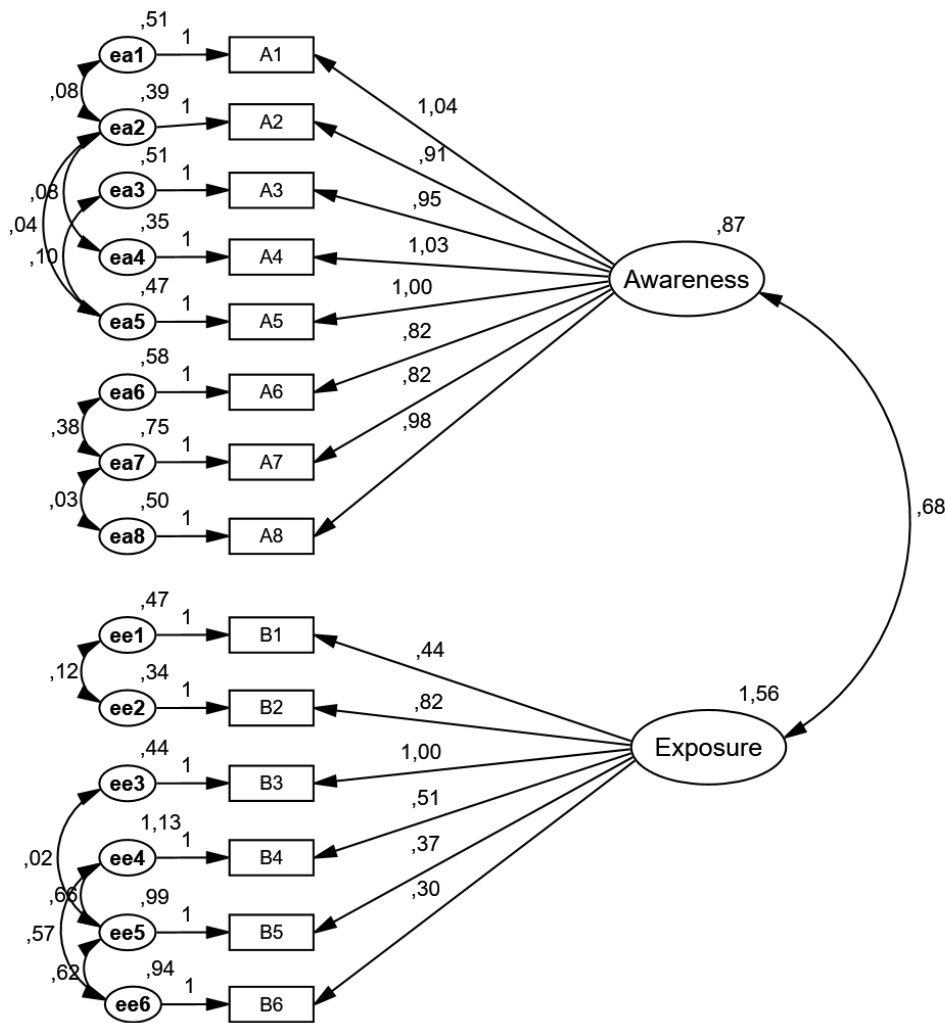
Measurement Weights Model = All *Factor loadings* are constrained (equated).

Structural Covariances Model = All *Factor loadings + factor variances and covariances* are constrained (equated).

Measurement Errors Model = All *Factor loadings + factor variances + factor covariances + error variances* are constrained (equated).

As it seen in [Table 5](#), according to the unconstrained (configural) model, the changes in CFI in the models obtained by constraining, in sequence, measurement weights, and structural covariances were less than .01. However, when error residuals constrained the changes in CFI was found to be more than .01. Hence, the measurement model has provided configural, measurement weight, and structural covariance invariance; but, not provided for measurement residual invariance across three school types.

Figure 3. Path diagram for configural, measurement weight, structural covariance, and measurement residual invariance of the measurement model for school type.



Chi square= 1143,416 Df= 275, GFI= ,858 CFI= ,914, RMSEA= ,055

Note: Only 3 covariance estimates (ee5 < -- > ee3=.024 with p=.226; ea7 < -- > ea8=.033 with p=.054; and ea5 < -- > ea2=.036 with p=.015) were not found to be significant, all the other parameters were found to be significant.

3.5. Measurement Invariance by Genders

Goodness of Fit Indexes of the Baseline Measurement Model for Gender

In the baseline measurement model based on the gender, the goodness of fit indexes were found to be as follows:

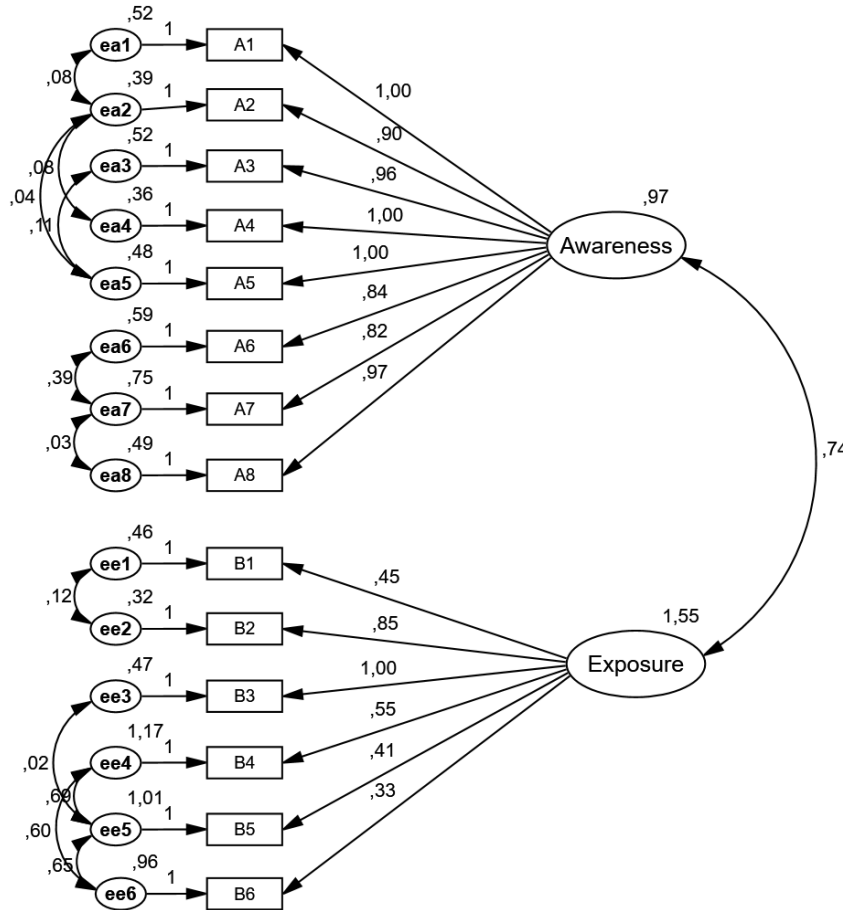
- ✓ for male teachers $X^2_{65}=164.122$; $X^2/sd=2.525$; GFI=0.953; CFI=0.978 and RMSEA=.057; and
- ✓ for female teachers $X^2_{65}=324.513$; $X^2/sd=4.993$; GFI=0.927; CFI=0.957 and RMSEA=.084.

In conclusion, the baseline measurement model in Figure 1 displayed a high level of model fit for the two genders.

3.6. Configural, Measurement Weight, Structural Covariance, and Measurement Residual Invariance of the Measurement Model for Gender

The unstandardized estimated parameters of the model are given with path diagram for genders in Figure 4, below, and the main findings regarding the measurement invariance according to the genders are presented in Table 6 below.

Figure 4. Path diagram for configural, measurement weight, structural covariance, and measurement residual invariance of the measurement model for gender.



Chi square= 610,502 Df= 170, GFI= ,924 CFI= ,958, RMSEA= ,050

Note: Only 3 covariance estimates (ee5 < -- > ee3=.020 with $p=.303$; ea7 < -- > ea8=.032 with $p=.061$; and ea5 < -- > ea2=.035 with $p=.017$) were not significant, all other parameters were found to be significant.

Table 6. Configural, measurement weight, structural covariance, and measurement residual invariance results by branch.

Model	Number of parameters	χ^2	df	χ^2/df	CFI	ΔCFI	RMSEA
1. Unconstrained (Configural)	80	488.635	130	3.759	.966		.052
2. Measurement Weights	68	505.893	142	3.563	.965	.001	.050
3. Structural Covariances	65	507.348	145	3.499	.966	.000	.049
4. Measurement Residuals	40	610.502	170	3.591	.958	.008	.050

Note: Unconstrained Model: All the parameters are predicted freely.

Measurement Weights Model = All Factor loadings are constrained (equated).

Structural Covariances Model = All Factor loadings + factor variances and covariances are constrained (equated).

Measurement Errors Model = All Factor loadings + factor variances + factor covariances + error variances are constrained (equated).

As it seen in Table 6, according to the unconstrained (configural) model, the changes in CFI in the models obtained by constraining, in sequence, measurement weights, structural covariances, and measurement residuals were less than .01. Hence, the measurement model has provided configural, measurement weight, structural covariance, and measurement residual invariance across two genders.

4. DISCUSSION and CONCLUSION

This study investigates the measurement invariance of the Nanoscience and Nanotechnology Awareness Scale (NSTAS) for three teacher branches, three school types, and two genders by using the covariance structural analysis to test configural and metric invariances.

There is need to plan and implement NSNT education at primary, secondary, undergraduate, and graduate levels, since teachers' knowledge and competences are the key to education. Factors affecting awareness and knowledge level of teachers/teacher trainees in NSNT should be determined and analyzed before implementing education programs (Hingant & Able, 2010; Jones et al., 2013). The NSTAS instrument was originally developed by Dyehouse et al. (2008) to promote awareness and factual knowledge among higher education students in the USA about nanotechnology uses, so students become acquainted with nanotechnology as a new field of research and innovation affecting society. The greater objective was to motivate university students to academic and career options in the field.

Braeken and Blömeke (2016) pointed out, "to allow for making group comparisons in terms of correlations with external variables, the stricter requirement of equal factor loadings" across groups (i.e., metric or 'weak' invariance) needs to hold. They also pointed out that "if we wish to directly compare observed scale sum scores between groups, then additionally, the residual item variances would be required to be equal across groups, such that every item can be considered equally reliable across groups". There are some group comparisons and some educational decisions based on these comparisons regarding nanotechnology and nanoscience using NSTAS scores. In terms of objectivity features of scientific research, to test whether the structural validity or the measurement model of the NSTAS scale works in different subgroups in the same way. In other words, it is extremely important to determine whether the measurement tool provides biased group results using the measurement invariance approach. Wicherts (2016) emphasized that measurement invariance is very important for the validity of tests. In the literature, we could not find any study about measurement invariance in the field of nanotechnology. Very few studies have been found in the literature on measurement instruments used in hard sciences. Some of them are given below.

Rocabado et al. (2019) performed measurement invariance testing for the configural, metric, and scalar models comparing black female students and all other students within the traditional and flipped courses for the two-factor model prescribed for the pre and posttests. Their analysis results showed that configural, metric, and scalar invariance was ensured. Maier et al. (2013) developed a preschool teachers' attitudes and beliefs toward science teaching scale. They used teacher ethnicity, education level, and experience level as subgroups. They conclude that the three factors remained invariant across each subgroup. Luo et al. (2019) presented validity evidence of scores produced from the S-STEM measurement tool, and they concluded that measurement invariance results showed that the instrument items in the surveys measured the same constructs in the same ways across gender, age groups, and races/ethnicities. Braeken and Blömeke (2016) investigated the measurement equivalence of teachers' beliefs across countries for the case of 'mathematics as-a fixed-ability'. They concluded that data provided configural and metric invariance but did not provide scalar invariance across countries. Clearly none of the measurement invariance studies cited provide indisputable explanation about the steps of invariance measurement. It is obvious that there is a deficiency in the hard science literature in

terms of emphasizing the importance of measurement invariance and elaborating step by step instructions and guidance.

Having examined the measurement model invariance with respect to configural, measurement weight, and structural covariance invariance for three groups of branches, three group of school types and two groups of genders, the present study arrived at the conclusion that configural, measurement weight and structural covariance invariances were ensured for branches, school types and genders. Also, residual invariance was ensured for genders. Residual invariances are not provided for branches, and school types leading us to conclude that not every item can be considered equally reliable across those groups.

In conclusion, the results of this study provide evidence that the measurement invariance requirement for valid group comparisons for the Nanoscience and Nanotechnology Awareness Scale has been satisfied; measurement invariance can be successfully implemented in science and technology education. Casas and Blanco-Blanco (2017) acknowledged using the method for Social Cognitive Career Theory (SCCT) models in predicting mathematical/scientific interests and occupational aspirations among Colombian secondary students. Another successful application was by Caputo (2017) in science and mathematics education of 7th grade secondary students in Italy. The Measure of Acceptance of the Theory of Evolution (MATE, a single-factor instrument that assesses an individual's overall acceptance of evolutionary theory) was tested to assess how it operates differently when administered to a population of non-science major preservice elementary teachers when compared with the reference population of in-service high school biology teachers and found to be reliable with the measurement invariance approach (Wagler & Wagler, 2013). As a result, it has been proved that the NSTAS scale will not generate biased measurements in comparing groups by teacher branches, school types and gender. Since the internal structure of NSTAS holds for different groups, NSTAS scale can be safely used to compare branch, school type and gender groups. Testing and interpreting the measurement invariance with the covariance structure approach using IBM AMOS-24, implemented with cases in this study, can be applied to all scales aimed at comparing different groups.

Acknowledgments

Presented orally at the “6th International Congress on Measurement and Evaluation in Education and Psychology” held between 05-08 September 2018 in Prizren-Kosovo.

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: 81576613/605/1955049 Gazi University, Ankara.

Authorship Contribution Statement

Şeref Tan: Conceptualization, Data Analysis, Methodology, Software, Resources, Discussion, Writing, Supervision and Validation. **Zeki Ipek:** Investigation, Methodology, Resources, Writing. **Ali Derya Atik:** Conceptualization, Investigation, Data Analysis, Resources, Discussion, Writing. **Figen Erkoc:** Investigation, Data Analysis, Resources, Discussion, Writing, Supervision and Validation.

ORCID

Seref Tan  <https://orcid.org/0000-0002-9892-3369>

Zeki Ipek  <https://orcid.org/0000-0002-8097-5849>

Ali Derya Atik  <https://orcid.org/0000-0002-5841-6004>

Figen Erkoc  <https://orcid.org/0000-0003-0658-2243>

5. REFERENCES

- AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Arana, F. G., Rice, K. G., & Ashby, J. S. (2018). Perfectionism in Argentina and the United States: Measurement structure, invariance, and implications for depression. *Journal of Personality Assessment*, *100*(2), 219-230. <https://doi.org/10.1080/00223891.2017.1296845>
- Bayda, S., Adeel, M., Tuccinardi, T., Cordani, M., & Flavio Rizzolio, F. (2020). The history of nanoscience and nanotechnology: From chemical–physical applications to nanomedicine. *Molecules*, *25*(1), 112. <https://doi.org/10.3390/molecules25010112>
- Blonder, R., Parchmann, I., Akaygun, S., & Albe, V. (2014). Nanoeducation: Zooming into teacher professional development programmes in nanoscience and technology. In C. Bruguère., A. Tiberghien., & P. Clément. (Eds.), *Topics and Trends in Current Science Education* (pp. 159–174). 9th ESERA Conference Selected Contributions. New York: Springer.
- Braeken, J., & Blömeke, S. (2016). Comparing future teachers' beliefs across countries: Approximate measurement invariance with Bayesian elastic constraints for local item dependence and differential item functioning. *Assessment & Evaluation in Higher Education*, *41*(5), 733–749. <http://dx.doi.org/10.1080/02602938.2016.1161005>
- Bryan, L. A., Sederberg, D., Daly, S., Sears, D., & Giordano, N. (2012). Facilitating teachers' development of nanoscale science, engineering, and technology content knowledge. *Nanotechnology Reviews*, *1*(1), 85-95. <https://doi.org/10.1515/ntrev-2011-0015>
- Boholm, A., & Larsson, S. (2019). What is the problem? A literature review on challenges facing the communication of nanotechnology to the public. *Journal of Nanoparticle Research*, *21*(86), 1-21. <https://doi.org/10.1007/s11051-019-4524-3>
- Byrne, B. M. (2013). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Psychology Press.
- Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Routledge.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (pp. 221–256). Praeger.
- Camerota, M., Willoughby, M. T., Kuhn, L. J., & Blair, C. B. (2018). The childhood executive functioning inventory (CHEXI): Factor structure, measurement invariance, and correlates in US preschoolers. *Child Neuropsychology*, *24*(3), 322-337. <http://doi:10.1080/09297049.2016.1247795>
- Caputo, A. (2017). A brief scale on attitude toward learning of scientific subjects (ATLoSS) for middle school students. *Journal of Educational, Cultural and Psychological Studies*, *16*, 56-76. <http://dx.doi.org/10.7358/ecps-2017-016-capu>
- Casas, Y., & Blanco-Blanco, A. (2017). Testing Social Cognitive Career Theory in Colombian adolescent secondary students: a study in the field of mathematics and science. *Revista Complutense de Educación*, *28*(4) 1173-1192. <http://dx.doi.org/10.5209/RCED.52572>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5
- Chung H., Kim, J., Park R., Bamer A. M., Bocell, F. D., & Amtmann D. (2016). Testing the measurement invariance of the University of Washington Self-Efficacy Scale short form across four diagnostic subgroups. *Qual Life Res*, *25*(10), 2559-2564. <http://doi:10.1007/s11136-016-1300-z>
- Dyehouse, M. A., Diefes-Dux, H. A., Bennett, D. E., & Imbrie, P. K. (2008). Development of an instrument to measure undergraduates' nanotechnology awareness, exposure,

- motivation and knowledge. *Journal of Science Education and Technology*, 17(5), 500-510. <https://doi.org/10.1007/s10956-008-9117-3>
- Enil, G., & Köseoğlu, Y. (2016). Investigation of nanotechnology awareness, interests, and attitudes of pre-service science (Physics, Chemistry and Biology) teachers. *International Journal of Social Sciences and Education Research*, 2(1), 50-63. <https://doi.org/10.24289/ijsser.279084>
- Greenberg, A. (2009). Integrating nanoscience into the classroom: Perspectives on nanoscience education projects. *ACS Nano*, 3(4), 762-769. [https://doi: 10.1021/nn900335r](https://doi:10.1021/nn900335r)
- Hingant, B., & Albe, V. (2010). Nanosciences and nanotechnologies learning and teaching in secondary education: A review of literature. *Studies in Science Education*, 46(2), 121-152. <https://doi.org/10.1080/03057267.2010.504543>
- Holland, L. A., Carver, J. S., Veltri, L. M., Henderson, R. J., & Quedado, K. D. (2018). Enhancing research for undergraduates through a nanotechnology training program that utilizes analytical and bioanalytical tools. *Analytical and Bioanalytical Chemistry*, 410, 6041-6050. [http://doi: 10.1007/s00216-018-1274-5](http://doi:10.1007/s00216-018-1274-5)
- İpek, Z. (2017). *Research on awareness levels of physics, chemistry, and biology teachers about nanoscience and nanotechnology*. [Doctoral Dissertation, Gazi University, Ankara]. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp>
- İpek, Z., Atik, A. D., Tan, Ş., & Erkoç, F. (2020). Study of the validity and reliability of Nanotechnology Awareness Scale in Turkish Culture. *International Journal of Assessment Tools in Education*, 7(4), 674-689. <https://doi.org/10.21449/ijate.708169>
- Jones, M. G., Blonder, R., Gardner, G. E., Albe, V., Falvo, M., & Chevrier, J. (2013). Nanotechnology and nanoscale science: Educational challenges. *International Journal of Science Education*, 35(9), 1490–1512. [http://doi: 10.1080/09500693.2013.771828](http://doi:10.1080/09500693.2013.771828)
- Laherto, A. (2010). An analysis of the educational significance of nanoscience and nanotechnology in scientific and technological literacy. *Science Education International*, 21(3), 160-175.
- Luo, W., Wei, H.-R., Ritzhaupt, A. D., Huggins-Manley, A. C., & Gardner-McCune, C. (2019). Using the S-STEM survey to evaluate a middle school robotics learning environment: validity evidence in a different context. *Journal of Science Education and Technology*, 28, 429-443. <https://doi.org/10.1007/s10956-019-09773-z>
- Maier, M. F., Greenfield D. B., & Bulotsky-Shearer R. J. (2013). Development and validation of a preschool teachers' attitudes and beliefs toward science teaching questionnaire. *Early Childhood Research Quarterly* 28, 366– 378. <https://doi.org/10.1016/j.ecresq.2012.09.003>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543. <http://dx.doi.org/10.1007/BF02294825>
- Millsap, R. E., & Yun-Tein, J. (2004) Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39(3), 479-515. http://doi:10.1207/S15327906MBR3903_4
- Rocabado, G. A., Kilpatrick, N. A., Mooring, S. R., & Lewis J. E. (2019). Can we compare attitude scores among diverse populations? An exploration of measurement invariance testing to support valid comparisons between black female students and their peers in an organic chemistry course. *Journal of Chemical Education*, 96, 2371-2382. <http://doi:10.1021/acs.jchemed.9b00516>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.

- Tan, Ş., & Pektaş, S. (2020). Examining the invariance of a measurement model by using the covariance structure approach. *International Journal of Contemporary Educational Research*, 7(2), 27-39. <https://doi.org/10.33200/ijcer.756865>
- 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), 4-70. <http://doi:10.1177/109442810031002>
- Wagler, A., & Wagler, R. (2013). Addressing the lack of measurement invariance for the measure of acceptance of the theory of evolution. *International Journal of Science Education*, 35(13), 2278-2298. <http://dx.doi.org/10.1080/09500693.2013.808779>.
- Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability test-ing. *The Clinical Neuropsychologist*, 30(7), 1006-1016. <https://doi.org/10.1080/13854046.2016.1205136>

6. APPENDIX

Table A1. Nanotechnology Awareness Instrument (Dyehouse et al., 2008)

For the following items, please indicate the extent to which you agree or disagree using the following scale: Strongly disagree, disagree, neutral, agree, or strongly agree.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
What is your awareness of nanotechnology? I can:					
1. Name a nanoscale-sized object.					
2. Describe one way nanotechnology directly impacts my life.					
3. Name a field of study that currently conducts nanotechnology research.					
4. Describe one way nanotechnology may benefit society/humankind.					
5. Name an application of nanotechnology.					
6. Describe a process to manufacture objects at the nanoscale.					
7. Name an instrument used to make measurements at the nanoscale.					
8. Describe one way nanotechnology may directly impact my life in the future.					

For the following items, please indicate the extent to which you have participated in each activity using the following scale: Not at all/never, very little, sometimes/ occasionally, a fair amount, or a great deal.	Not at all/never	Very little	Sometimes/ occasionally	A fair amount	A great deal
What is your exposure to nanotechnology? I have:					
1. Heard the term nanotechnology.					
2. Read [something] about nanotechnology.					
3. Watched a program about nanotechnology.					
4. Had one [or more] instructors/teachers talk about nanotechnology in class.					
5. Participated in an activity involving nanotechnology [lab, project,...].					
6. Taken a class about nanotechnology.					

Table A2. Nanoscience and Nanotechnology Awareness Scale (NSTAS) - Turkish Version.

Farkındalık Alt Ölçeği (Awareness Subscale)	Kesinlikle Katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle Katılıyorum
1. Nanoölçek boyutunda bir nesne adı söyleyebilirim.					
2. Nanoteknolojinin hayatımı doğrudan etkileyen bir yöntemini söyleyebilirim.					
3. Bugünlerde nanoteknoloji araştırması yürüten bir çalışma alanı ismi söyleyebilirim.					
4. Nanoteknolojinin topluma/insanlığa faydalı olabilecek bir yöntemini tanımlayabilirim.					
5. Bir nanoteknoloji uygulamasının adı söyleyebilirim.					
6. Nanoölçekte nesnelere üretmek için kullanılan bir yöntemi tanımlayabilirim.					
7. Nanoölçekte ölçüm yapmakta kullanılan bir araç ismi söyleyebilirim.					
8. Gelecekte nanoteknolojinin hayatımı doğrudan etkileyebilecek bir yöntemini söyleyebilirim.					

Deneyim (etkileşim) Alt Ölçeği (Exposure Subscale)	Hiçbir zaman	Nadiren	Ara sıra	Çok sık	Her zaman
7. Nanoteknoloji terimini duydum.					
8. Nanoteknoloji hakkında bir şeyler okudum.					
9. Nanoteknoloji hakkında bir program izledim.					
10. Sınıfta bir (veya daha fazla) öğretmen/öğretim elemanının nanoteknoloji hakkındaki konuşmalarını dinledim.					
11. Nanoteknoloji konusunun işlendiği bir etkinliğe katıldım (laboratuvar çalışması, proje, seminer, konferans).					
12. Nanoteknoloji hakkında bir ders aldım.					

The Learning Effect of Corpora on Strong and Weak Collocations: Implications for Corpus-Based Assessment of Collocation Competence

Hatice Altun ^{1,*}

¹Pamukkale University, The School of Foreign Languages, Denizli, Turkey

ARTICLE HISTORY

Received: Dec. 22, 2020

Revised: Apr. 26, 2021

Accepted: May 20, 2021

Keywords:

Strong and weak collocations,
Concordancers,
Traditional learning,
Lexical assessment.

Abstract: Although corpora and corpus linguistics have been applied for quite long in foreign and second language settings, there is still limited understanding about how EFL learners use corpus tools along with dictionaries to enhance their collocation knowledge. This study aims to gain insight into the effectiveness of corpus-based pedagogy in comparison with the conventional vocabulary teaching methods, particularly using dictionaries. The study was conducted with two non-English major advanced groups of L2 learners at a public university. The experimental group studied 16 pre-selected formal academic words and their strong and weak collocations with corpus (COCA, the corpus of contemporary American English), while the comparison group studied the same collocations using advanced learner's dictionaries. The instruments for collecting data included the Oxford placement test, pretest, posttest, and exercises devised for particular teaching points of collocations. Results of Repeated Measures ANOVA tests showed no significant difference between the two experimental groups. However, the corpus-based approach showed more impact on the reception of strong collocations acquired by the corpus group at a slightly better performance rate, as evidenced by the group's mean scores (Corpus=45.91, Dictionary= 44.06). Interestingly, the acquisition of weak collocations was better for the dictionary use group (Corpus=54.08, Dictionary= 57.18). The paper thus offers some implications for teaching and assessing collocation knowledge and makes suggestions that EFL practitioners should create variations in instructional methodologies through gaining awareness of the increasing availability of innovative technologies. Further research on collocations' assessment has also been suggested.

1. INTRODUCTION

Numerous studies and publications have emphasized the contribution of corpora to the language learning environment, and corpora are being used more frequently as a reference tool for language teachers and learners as a result of the growing availability of advanced technology. Initially, corpora were used mainly for the production of dictionaries and language textbooks (e.g., Barlow & Burdine, 2006; Gilquin et al., 2007; Sinclair, 2001; Thurstun & Candlin, 1997). The common use of corpora in material development is reported as a result of the effectiveness of the published materials in foreign and second language classrooms. The use of authentic language samples from corpora serves as a comprehensible input for language learning settings,

*CONTACT: Hatice ALTUN ✉ haticealtun@gmail.com 📍 Pamukkale University, The School of Foreign Languages, 20070, Denizli, Turkey

particularly in foreign language (EFL) classrooms where it is rather challenging to expose language learners to various uses and contexts of a word studied. Corpora were also used as a source of linguistic research on lexical studies, grammar, discourse analysis, pragmatics, and linguistics (e.g., Benesch, 2001; Biber et al., 1999; Cortes, 2002; Flowerdew & Peacock, 2001).

More recently, direct access to corpora by learners comprises the subject of a number of studies (e.g., Bernardini, 2002; Boulton & Cobb, 2017; Chambers, 2005; Chambers & O'Sullivan, 2004; Gaskell & Cobb, 2004; Gilmore, 2009; Yoon & Hirvela, 2004). In all these corpora studies, language learners interact with the text in the concordancer to observe, speculate, and explore language patterns, word forms, and collocations. Learners can make generalizations about grammatical features, syntax, agreement, and stylistics thanks to this inductive learning approach. This is particularly important in the EFL contexts where students usually receive most of their language education through another medium but English. Learner's direct access to corpora promotes lexical consciousness, through which students familiarize themselves with the various contexts of the lexical items. For instance, if the students create a list of vocabulary and prepositions used in context, the concordance lines help students to understand that the same lexical items can be used in multiple contexts. This process can promote students' guessing ability by demonstrating the various uses of language items studied (Johns, 1991).

A new path for corpus use is applying corpus linguistic methods and tools in the design and validation process of language teaching and assessments. Some recent studies particularly focus on the potential benefits of exploiting a learner corpus for testing and assessment of L2 proficiency in writing and also speaking (Callies, 2016; Callies & Götz, 2015). Although corpus studies with the testing focus are still at an early stage, they contribute a lot to the research on the assessment of L2 proficiency (Deshors et al., 2016; McCarthy, 2013). This study is also expected to offer some potential beneficial implementations for assessing L2 vocabulary proficiency, particularly in the context of the *Common European Framework of Reference for Languages* (CEFR).

Another flourishing interest area among language teachers and researchers is formulaic expressions and idiomatic language use (Biber et al., 2004; Wray, 2002, 2008). It is considered that mastery of formulaic expressions is essential to acquire lexical competence and an idiomatic control of language (Ellis, 2002, 2003). The phenomenon of collocations occupies a focal point in the scheme of formulaic language research (Firth, 1957; Lewis, 1993; Lewis, 1997, 2000; Liu, 2010; Nesselhauf, 2003). The study of collocations is of great interest in language teaching because language learners are considered to benefit from the naturally occurring word combinations to gain a more natural phraseology of L2. Thus, instead of memorizing long chunks of words, the learners would be able to produce some of the collocation combinations and would also develop some understanding of linguistic features and processes which affect the way collocations are formed (Walker, 2011). Recently emerging awareness on the importance of corpus consultation, especially corpus concordancing, in the study of collocations has led to the penning a number of studies devoted to this issue (Breyer, 2009; Chan & Liou, 2005; Cheng et al., 2003; Durrant & Schmitt, 2009; Lee & Swales, 2006; Liu, 2010). Nevertheless, despite a plethora of research articles and projects comparing the effectiveness of traditional methods and dictionaries to corpora (Basal, 2019; Çelik, 2011; Daskalovska, 2015; Lai & Chen, 2015), corpus-based language teaching focusing on learners' corpus consultation about different collocation types (i.e.; strong vs. weak collocations) is still on all fours, and more effort is needed to draw up-and-coming implications for EFL contexts.

This paper attempts to contribute to the above-stated niche as a way to teach collocations. More specifically, it aims to see if concordancing exercises, which rely on collocation competence, can enhance the nature of vocabulary learning. This experimental study, therefore, aims to

explore the potential benefits of hands-on concordancing over dictionary use in-class activities for teaching strong and weak collocations over five weeks.

2. LITERATURE REVIEW

2.1. Corpora and Language Learning

A corpus is the accumulation of vast spoken or written electronic text archives (Anderson & Corbert, 2017). The texts are machine-readable and can easily be manipulated by software that can analyze the linguistic constructs in question. A careful analysis can provide insights into how language is used typically and commonly. The size of a corpus can change from millions to billions of words, and it may contain several genres which learners found useful to explore. A concordancing program enables researchers to view all of the occurrences of a particular word in its immediate environment in a corpus. The immediate environment contains several words before and after the search word itself. The full concordance lines indicate the larger text in which examples occur (*ibid.*). Concordancing allows the researchers to perform basic qualitative and quantitative analysis to show all aspects of the nature of the word as well as its frequency in a specific context (Flowerdew, 1996).

Corpora may provide learners with valuable tools such as basic lexical, grammatical, and organizational details for the genre (Tribble, 2001). With a corpus and a concordancer, learners not only see the authentic examples provided but also have the opportunity to study language patterns (Biber et al., 1999). Corpora display word collocations via the concordancing program. Learners can see preceding and subsequent data for the term they are searching for by looking at collocational frequencies. Another advantage of a corpus is the context it brings in examples (Biber et al., 2004). Learners can appreciate the sense in which terms should be used by looking at the examples. By making inferences, students can be able to figure out what a word means. Corpora may also foster an atmosphere conducive to inductive learning (Flowerdew, 2009). This gives students power over their language learning. In this sense, foreign language students take on the position of linguistic researchers, analyzing data and coming up with their own rules and conclusions.

Some scholars and language teachers (Johns, 1991; Tribble, 2001) have strongly supported the use of corpora instead of dictionaries and traditional activities to develop competencies in various skills on account of the fact that concordances are argued to promote learners' analytical thinking skills and autonomy. Adherents of corpora have also argued that traditional learning tools, including dictionaries, are tedious and tiring and also nonproductive tools, particularly for vocabulary learning. The inauthentic examples and the vague language use in dictionaries prevent learners from realizing various authentic contexts of words (Tribble & Johns, 1990). However, according to Cobb (2003), in spite of all the burdensome and time-consuming effects of dictionaries, many language learners still depend upon the dictionaries to learn vocabulary. Additionally, in their meta-analysis, Lee et al. (2018) argue that corpora can increase vocabulary gains considerably, particularly in in-depth vocabulary knowledge of collocations. Also, some collocations which are even difficult to be recognized by the native speakers can easily be taught through concordances.

As opposed to the importance credited to the exploitation of corpora in language teaching, however, total reliance on it may be problematic in that corpora may pose some challenges and obstacles for some learners. First of all, all learners may not have positive attitudes towards inductive discovery learning (Flowerdew, 2009). According to Flowerdew (2009), corpus use is typically correlated with an inductive approach, which may not be suitable for all students due to their differing cognitive styles. This style of learning can benefit field-dependent students who enjoy discussions based on the application of rules from examples (*ibid.*). Field-independent learners, on the other hand, who prefer simple rule instruction will not find it

useful. Cobb (1998) also raises another practical question about corpora exploitation. Lexical information is massive and maybe potentially confusing to the learners. While words occur in a wide range of contexts, many of the words in the concordance lines are unfamiliar, and the contexts are short, incomplete, and do not indicate a coherent and unified context (Cobb 1998). As a result, the teacher's function as a facilitator is essential to overcome the challenge caused by the context (Flowerdew, 2009).

2.2. Collocations and Corpora

Collocations are words that appear together in a text more often than their individual frequencies or than would be predicted by chance (Halliday, 1966). Collocating words predict each other, i.e., when one part of a collocating pair is detected, the odds of discovering the other part improve (Hoey, 1991; Jones & Sinclair, 1974). However, there is no set definition of what word combinations are considered as collocations among language educators. The controversy often stems from the disagreement over how structurally fixed and meaningfully transparent a word combination should be to be considered a collocation. Yet most educators agree that collocations are word forms with restricted structural variations and vary from free words, and to alleviate the problem of this arbitrariness, some scholars offered a scale with subcategories, such as 'strong,' 'medium strength,' (Crowther et al., 2002) or 'strong,' 'weak' and 'fixed' (O'Dell & McCarthy, 2008). For this study, the researcher exploited this scale of collocations and focused particularly on strong and weak collocations to be studied by advanced L2 learners.

Language users need to develop collocational links for an efficient lexical network. However, Nesselhauf (2003) and Altenberg and Granger (2001) argue that even advanced English learners have issues with the correct use of collocations. In the EFL settings, developing collocational competence is rather challenging due to the arbitrary nature of collocations. Collocational mistakes are usually the most dominant ones in EFL learners' outputs (Gui & H., 2002; Hsu & Chiu, 2008). Koç (2006) also discovered that one of the main problems with Turkish EFL learners is the lack of collocational competence. Learners tend to learn vocabulary as isolated units rather than as formulaic sequences of words in combination with each other. Furthermore, Prodromou (2003) contends that collocations, either fixed or more flexible, are formed after many years of habitual use by the native speakers of a language. Collocations offer 'chunks' of English that are part of formulaic language ready to be used; therefore, the automation of collocations enables 'native speakers' to express themselves fluently. Second language learners, however, lack this automation and, thus, are more prone to using unnatural phraseologies. In order to achieve automaticity in collocational use, second language learners should be aware that they need to develop an ability to comprehend and produce collocations as unanalyzed chunks (Prodromou, 2003).

Since mastering collocations is rather challenging (Wray, 2000), a large body of study has concentrated on learner mistakes and the primary challenges second language learners encounter while studying collocation norms (Howarth, 1998; Liu, 2010; Nesselhauf, 2003, 2005). It is also well known that second language learners tend to rely on weak collocations, which are non-restricted word combinations (e.g., nice memories, a good meal, bad friends) (Hasselgren, 1994; Nesselhauf, 2005). Considering that word frequency is one of the imperative determiners in making lexical choices (Foster & Chamber, 1973), it is not surprising that high-frequency weak collocations are processed quickly. So the element of familiarity plays a vital role to clutch for the words learners feel safe with, and even advanced learners systematically overgeneralize these 'lexical teddy bears' – "core words – learnt early, widely useable, and above all safe (because they do not show up as errors)" (Hasselgren, 1994, p. 250). On the other hand, strong collocations – low frequency, more clear-cut lexical combinations – take a longer time to learn and are less likely to be used by second language learners (Conzett, 2000). However, strong collocates are expected to facilitate the processing of the following noun

because they prime the subsequent noun and make it more restricted than the same word preceded by a weak collocate (e.g., auburn hair vs. brown hair, inclement weather vs. bad weather) (Hoey, 2005). This inherent paradoxical nature of strong and weak collocations poses an additional challenge to learning collocations. Therefore, the contradictory effects of strong and weak collocations – learners’ reliance on weak collocations but their being less predictable or strong collocations’ facilitating effect of the subsequent word but being difficult to be processed – on gaining collocational competence need to be studied more in EFL settings. There are, however, few studies focusing on learning above mentioned collocations through concordancing (Conzett, 2000).

Only a few studies have looked into the effects of using concordancers to teach EFL students. Some notable ones are as follows. In Sun and Wang’s (2003) study, the efficacy of inductive and deductive teaching approaches on EFL students was investigated. Participants used an online monolingual concordancer to research collocations of various difficulty levels. The inductive group benefited substantially more than the deductive group after the posttest. There was no significant difference between the learners’ performance affected by the teaching method, inductive or deductive, in terms of tricky collocations. However, the inductive approach was more effective in teaching easier collocations with the help of corpora.

Daskalovska (2015), in another notable study, explored the influence of concordance on 44 first-year English language and literature learners’ adverb-noun collocation knowledge. The experimental group outperformed the control group, who studied the collocations through traditional exercises and dictionaries. She underpinned the valuable contribution of concordance use on the collocational production of ELT learners. One last research study worth mentioning is Nesselhauf’s (2003) groundwork. She conducted an exploratory study on verb-object-noun collocations in a corpus of academic essays written by non-native speakers of English. He concluded that although rote learning and behaviorism are discredited, a number of collocations need to be taught and learned explicitly; in this case, the criteria for the selection of collocations to be taught can be determined based on the acceptability and frequency of collocations in any special register of interest to the learner.

Nesselhauf (2005) suggests three criteria to select collocations to be taught to advanced level students: frequency, difficulty, and degree of disruption. Frequency is the number of occurrences of a collocation set in a certain text that students need to study. Collocations with high-frequency and wide-range collocations are deemed worthy of teaching in some studies (Hill, 2000; Hill et al., 2000), and in others, collocations with medium or weak strength (Hill, 2000). Degree of difficulty, i.e., degree of susceptibility to deviation, is the second criterion in the model in which two types of difficulty are explained: absolute difficulty and relative difficulty. Both of them serve as a rating scale for the learnability of a collocation. Deviation in the model means using unnatural phraseology or ungrammatical word combinations. The third criterion is the degree of disruption, i.e., the extent to which a deviant expression confuses the reader or listener and obstructs the quality of meaning to be conveyed or even disrupts the communication. Nesselhauf (2005) admits that this criterion, the disruption criterion, is rather challenging to measure because it is hard to express the degree of disruption in numbers. Moreover, the fuzziness of the idea of disruption (e.g., according to whom and according to what situation) makes the criterion challenging to justify.

Collocations, as a necessary form of vocabulary awareness, have caused learning problems for EFL learners, according to the studies described above (Liu, 2010; Nesselhauf, 2003). The selection of collocations to be taught does not seem to be applicable to all proficiency levels. In the case of advanced levels, learners strive for high proficiency; thus, learners’ needs should be considered as a criterion as well as other dimensions related to collocations. Furthermore, depending on collocation instructions, various forms of collocations seem to behave differently.

Drawing on the criterion of frequency and degree of difficulty in Nesselhauf's (2005) model, the current study, therefore, aims at exploring advanced level Turkish EFL learners' learning processes of strong and weak collocations (Crowther et al., 2002; O'Dell & McCarthy, 2008) with the help of a corpus, i.e., the corpus of American English, COCA- (Davies, 2008). The study looks into the causes of individual treatment differences (with or without concordancers) and various collocation types in order to fill in the gaps identified in the previous research survey.

3. METHOD

3.1. Design

This study addresses the possible aftereffects of hands-on concordancing exercises on advanced level Turkish EFL learners' learning strong and weak collocations in comparison to traditional dictionary use. The study has employed a pretest and posttest design, with 44 participants in two groups, a control and an experimental. Both groups took part in the treatment sessions between pretest and posttest. The control group studied the selected collocations through dictionaries and the experimental group via a corpus. The dependent variable of the study is learners' achievement on a collocation test developed by the researcher. The independent variables are two groups who study using concordancing activities and an online dictionary, and the type of collocations taught: strong and weak. Instruction was delivered to both groups through explicit classroom teaching based on the activities prepared by the researcher. The participant groups showed differences as to whether they used dictionaries or concordances during the treatment. The experimental group explored concordance lines of COCA to make meaningful deductions about the collocations to be learned. The control group studied the same vocabulary using the traditional advanced learners' dictionary.

3.2. Research Questions

The following research questions have been addressed:

1. Do concordancing exercises have any impact on L2 learners' collocation competence in comparison with traditional dictionary use?
2. Does the reception level of strong and weak collocations reveal a significant difference in advanced L2 learners?

3.3. The Hypotheses

It was hypothesized that:

H₀: Statistically no significant difference will be observed in students' posttest scores across the two groups after a period of explicit vocabulary teaching.

H₀: Statistically no significant difference will be observed in students' performances in posttest scores with regard to strong and weak collocations across the two groups after a period of explicit vocabulary teaching.

The first research question was investigated by assessing the performance of the experimental group against the control group. The second research question was explored by comparing the potential development of the two groups with regard to the strong and weak collocations.

3.4. Participants

All the participants were EFL learners enrolled in an academic writing class at a public university, and they were all native speakers of Turkish. These 51 students took the course of academic writing, during which the collocation treatment was administered for five weeks. Eight participants did not take the posttest; therefore, they were excluded from the data. In total, 44 students took part in all phases of the study. All participants had an upper-intermediate or advanced level of English language proficiency. Their proficiency level was checked using an

Oxford placement test as part of the study. The mean score was 43.86 (SD, 3.968), which was classified as B2 level-upper intermediate by CEFR. According to their own assessment, their computer skills ranged from basic to intermediate and more advanced. None of the participants had any previous knowledge of corpus linguistics.

3.5. Data Collection Instrument and Target Structures

A multi-faceted protocol was adopted during the vocabulary collection and test creation phases. Drawing on the criterion of frequency and degree of difficulty in Nesselhauf's model (2005, see literature review 2.2 for the detailed account of the model), the researcher has identified several strong and weak collocations from the teaching materials used in classes and exams in order to meet the advanced EFL learners' needs. The lexical items with a medium degree of difficulty but the relatively low frequency, or vice versa, received a fair amount of attention while preparing the list of collocations to be taught, and they were tested later on in the study. The relative degree of difficulty is measured by comparing the number of deviant expressions of collocation with its overall number in a particular text.

The collocations were selected from among those identified as important because they were considered to be of help to the advanced learners of English in their written and spoken English outputs. Additionally, the researcher focused on collocations that are not immediately obvious (e.g., *adhere to standards*, *auburn hair*, and *broad accent*), considering that those collocations would be helpful for their language exams given in the school and also for the standard exams such as TOEFL, IELTS, and GRE that they might need to take according to their future aspirations.

All collocations tested were adjective-noun bigrams. According to corpus studies, the most common grammatical element in academic texts is nouns (300,000 nouns per million words) (Biber & Conrad, 1999; Biber & Gray, 2016; Biber et al., 1999). The other two most common grammatical functions are adjectives and prepositions (Biber & Gray, 2011). Due to their frequency in the teaching and testing materials, only adj+noun collocations were included as the items to be used in the treatment. Additionally, it is considered that students would encounter adj+noun collocations in most of the high stake tests as well, so these combinations seemed like the most appropriate choice from among the other collocation types.

The selected collocations were divided into two categories. The first category is defined as strong collocations, in which the words are very closely associated, e.g., *mitigating circumstances* or *factors* (see the literature review for the detailed information about strong and weak collocation types). The second one is that of weak collocations in which words collocate with a range of other words. For example, *broad* collocates with a broad range of different nouns, e.g., *broad avenue*, *accent*, *view*. It is also considered that, in terms of their fixedness and idiomaticity, the weak and strong collocations form a continuum, with stronger ones at one end and weaker ones at the other (Conzett, 2000). Most collocations lie somewhere between the two.

The strength of collocations was operationalized through Mutual Information (MI) scores calculated for selected adjective-noun bigrams in the COCA. From among the other association measurements (AM) like T-scores and Log Dice, only MI scores were used as a reference to calculate the probability of co-occurrence of the collocations for some reasons. First, T-scores are considered to be the best indicator for lexical PP-verb collocations among all association measures (Hoffmann et al., 2008) so it was not the best alternative to measure adj+noun combinations' strengths. Although another AM, Log Dice, has been introduced as an alternative to MI scores, it has not been explored enough in language learning research yet (Gablasova et al., 2017). Therefore, MI scores seemed to be the most relevant measure to give information about the bond of probability between the adjectives and nouns used in the current study.

Additionally, MI scores are one of the most frequent and reliable measurement tools recommended in the literature to calculate the strength of collocations (Hunston, 2002; Hunston & Laviosa, 2000; Walter, 2012).

MI scores calculate the extent to which specific words co-occur compared to the number of times they appear separately, and they strongly rely on frequencies. Therefore, in order to make sure about the strength of collocations, MI scores were checked using COCA and BNC (British National Corpus). In total, 16 collocations were identified: 8 strong and eight weak ones. Weak collocations were defined as adjective-noun bigrams with an MI score lower than 3, and strong collocations were defined as adjective-noun bigrams with an MI score higher than 8. These cutting edges were recommended by Hunston (2002, p.71). It is generally accepted that MI scores lower than 3 suggest an insignificant likelihood of co-occurrence between the node and its collocate. Therefore, the MI score over 8 would show a highly significant relation of the probability between the searched items.

The classroom exercises were designed to explore the collocates of the pre-determined 16 words. Due to the semantic unrelatedness of these 16-word collocations, exercises focused on discrete items in a rather structured way in a multiple-choice test. In tandem with Nesselhauf (2003), the researcher adopted an explicit teaching method while studying the collocations with learners. The five-week teaching material comprised matching, gap filling, paraphrasing, error correction, and production type of exercises, which allowed learners to explore the selected words and their collocations. The materials also sought to assess learners' ability to adapt their vocabulary information to new contexts. These exercises were studied as part of the academic writing course for almost half an hour every week.

The collocational knowledge test utilized in the study was also designed and developed by the researcher and was used to evaluate students' collocation competence. The multiple-choice test format was chosen for the receptive collocational test, given the objectivity of scoring it allows. The test instructed participants to determine the correct collocate of the highlighted 16 words.

All the distractors were chosen from among the pseudo-collocates, weakly collocated or unrelated items in the lists of COCA and BNC in relation to the search item. For each item, the strong collocates were defined after a thorough search on both corpora. Those collocates that has the highest frequency rate were chosen as the correct answer. Then all the distractors' frequency and strengths were checked in order to make sure that the correct answer is the best option. The piloting of the collocational knowledge test was conducted with 20 ELT students at a different public university and with three English teachers. All the necessary items and distractors' developments were done based on the results obtained from piloting. Cronbach alpha was .815 for the collocation test, which indicated a high internal consistency. One week before and after the five-week experiment, pretests and posttests, which were basically the same test, were administered.

3.6. Treatment

The two participant groups in the experiment were assigned according to lists provided for the academic writing course; that is to say, section one was the first group, and section two the second. The first group (G1), called the corpus group, studied the words and their collocations with concordance and corpus-based activities, but the other group (G2), the control (dictionary) group, used traditional dictionaries while studying the same words. There were five sections in each course, and at each session, four collocation combinations were studied. The activities were completed in half an hour under the guidance of the course instructor. After the administration of the pretest, an introductory lesson in which the collocations and their particular uses were taught by the course instructor was conducted in the first part of the experiment. The corpus group received additional information on the utilization and searching

with a concordancer. The dictionary group dwelled upon the exploitation of dictionaries while studying collocations during the introduction week.

The main part of the experiment, the five-week teaching treatment, was unique to the groups. Corpus group (G1) delved into the corpus queries with COCA, one of the largest corpora in the world with one billion words from eight different genres. Although COCA’s web page offers several linguistic search opportunities, the learners were only asked to use the frequency counts for the collocation search. All the learners in the corpus group performed the classroom task, which required searching through COCA using their own computers during the class period. The control group (G2) used several advanced learners’ dictionaries to do the same collocation searches. They completed the same tasks with the corpus group, and they were not introduced concordances. After the five-week treatment, the posttest was administered to evaluate participants’ performance in collocation learning. Participants’ test scores in each group were accumulated to conduct the necessary analyses.

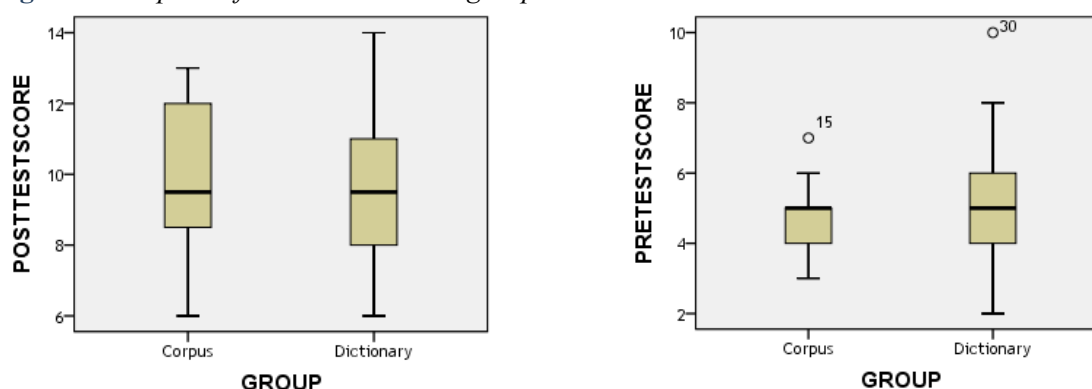
3.7. Data Analysis

A one-way ANOVA with repeated measures was used to explore the effect of the two treatments on advanced L2 learners’ collocational competence in two stages. The test type and the collocation types were taken as within-measures of the study. To assess the assumption of a one-way ANOVA, the researcher first checked the normality condition of the data set before making a decision on which statistical method should be used, using skewness and kurtosis indexes along with the Shapiro-Wilk test. Table 1 shows that all skewness and kurtosis values of the data were between -1.96 and +1.96, a threshold recommended by Ghasemi and Zahediasl (2012). This suggests the normal distribution of the data sets of the study. The result of the Shapiro-Wilk test also showed that all four data sets satisfied the normality condition, $p > .05$; therefore, the null hypothesis fails to be rejected. As is shown in the boxplots (see Figure 1), there seems to be one outlier in each group, so they were excluded from the data to conduct the analysis.

Table 1. Descriptive statistics of two groups in pre and posttest.

Groups	N		Mean	SD	Skewness	Kurtosis	Shapiro-Wilk
Corpus (G1)	20	Pretest	4.70	1.081	.117	-.212	.919
Dictionary (G2)	24		5.13	1.849	.521	.918	.953
Corpus (G1)	20	Posttest	9.70	2.130	-.072	-.749	.936
Dictionary (G2)	24		9.33	2.297	.046	-.580	.942

Figure 1. Boxplots of test scores across groups.



Since the within-subject test time variable has only two levels, the test for sphericity could not be applied. After the assumptions were met to conduct a repeated-measures ANOVA design, the test was run to pursue the analysis.

4. RESULTS

Both groups performed at a similar rate, according to the means of the pretest results ($G1_M = 5.56 / G2_M = 4.97$). These findings showed that there were no major variations in pre-learning histories between the groups prior to the pretest. The means of the posttest findings, on the other hand, showed a positive variance in favor of G1, which explored lexical items using corpora and concordance-based exercises ($G1 = 67.24 / G2 = 64.81$). [Table 2](#) below presents the summary of descriptive statistics of the pre and posttest results of the collocation test.

Table 2. Descriptive statistics from RM ANOVA for pre and posttest of collocation test, M (means) - SD (Standard Deviation).

Group	M/SD	Pretest	Posttest
G1 (20)	M	32.74%	67.24%
	SD	7.269	7.269
G2 (24)	M	35.17%	64.81%
	SD	8.923	8.924

A repeated measures one-way ANOVA was used to see whether the observed difference in the means of the posttest findings for the two groups was statistically meaningful. The findings, as presented in [Table 3](#), showed that the variance in posttest results favoring the corpus group (G1) was not significant ($F(1, 42) = .955, p = .334$). In response to the first research question about whether corpus activities create a significant difference between the two groups' collocational performance, it can be concluded that the null hypothesis cannot be rejected based on this result. In other words, the groups performed in a parallel manner on both collocation tests. Although the increase rate (of means) in both groups was quite large (nearly 30 points), there was not a significant difference between the groups. However, it can still be commented that regardless of the collocation type, the groups' learning performance during the practice period in the context of collocations was positive.

Table 3. Test of within-subjects effects from RM ANOVA for test results.

	Sum of Squares	df	F	Sig.	Partial Eta Squared	Observed Power
Pre/posttest	22313.204	1.000	165.243	.000	.797	1.000
WithinGroups	128.989	1.000	.955	.334	.022	.15
Error	5671.370	42				

[Table 4](#) shows descriptive statistics for the pre and posttest results in terms of collocation form. When the pretest outcomes of both strong and weak collocations were compared, the means of both classes were found to be reasonably similar ($G1$ strong collocation (sc) = 47.68/ weak collocation (wc) = 47.68), ($G2$ sc = 49.62 / wc = 50.41). $G1$ showed better performance in weak collocation items, but $G2$ revealed equally better performance in strong collocation items in the pretest. The means of posttest results of both groups as compared to those of pretest results indicated a decline in terms of strong collocation items. Despite this stated decline, the study revealed that the decline in the experimental group was less than the control group ($G1$ sc = 47.68 / 45.91, $G2$ sc = 49.62 / 44.06). Both groups, on the other hand, revealed better

performance in the weak collocation items in the posttest in comparison to the pretest results, but the results showed positive variance in terms of G2 this time (G1 wc = 54.08, G2 wc = 57.18).

Table 4. Descriptive statistics from RM ANOVA for pre and posttest with regard to strong and weak collocations, *M* (means) - *SD* (Standard Deviation).

Group	M-SD	Pretest Strong C	Pretest Weak C	Posttest Strong C	Posttest Weak C
G1 (20)	M	47.68%	52.37%	45.91%	54.08%
	SD	20.82	20.86	14.96	14.96
G2 (24)	M	49.62%	50.41%	44.06%	57.18%
	SD	17.58	17.54	12.96	12.54

A repeated-measures ANOVA analysis was run to see if the variance in the means of the posttest on strong collocation items was statistically significant, and the results (see Table 5) obtained from the test revealed that the difference in the strong collocation items observed in favor of corpus G1 was not statistically significant ($F_{1,42} = .421, p = >.05$).

Table 5. Test of within-subjects effects from RM ANOVA for strong collocations.

	Sum of Squares	<i>df</i>	F	Sig.	Partial Eta Squared	Observed Power ^a
Collocation type	323.789	1.000	1.733	.195	.040	.251
WithinGroups	78.660	1.000	.421	.520	.010	.097
Error	7848.761	42				

The test results of the repeated measures analysis conducted for weak collocation items within the tests revealed that the difference in the weak collocation items observed in favor of dictionary G2 was not statistically significant ($F_{1,42} = .361, p = >.05$), either, (see Table 6).

Table 6. Test of within-subjects effects from RM ANOVA for weak collocations.

	Sum of Squares	<i>df</i>	F	Sig.	Partial Eta Squared	Observed Power ^a
Collocation type	440.116	1.000	2.701	.108	.060	.362
WithinGroup	139.426	1.000	.856	.360	.020	.148
Error	6843.968	42				

The second research question explores the acquisition level of strong and weak collocations in both groups. From this analysis, it can be concluded that we cannot reject the null hypothesis, which claims that there was no significant difference between groups in terms of their competence with regard to collocation types. The experimental group showed slightly better performance, as evidenced by the groups' mean scores (G1 sc = 45.91, G2 sc = 44.06). However, the acquisition of weak collocations was slightly better for the dictionary use group despite not being evidenced by the statistical result (G1 wc = 54.08, G2 wc = 57.18). That is to say; it was found from the study that the corpus-based approach might have created some impact by chance on the reception of strong collocations.

5. DISCUSSION

The current study aimed to determine the more effective way of teaching strong and weak adjective-noun collocations using either concordancing tools or traditional learning tools of dictionaries. The results of the study did not support the hypothesis that corpus-based treatment would be better in teaching collocations, unlike some other studies which provided some profound effects in favor of corpus use in the literature (Chan & Liou, 2005; Daskalovska, 2015; Tsai, 2019). However, in terms of the collocation learning after the treatment, it can be argued that if enough time and effort spent on the study of collocations, L2 learners improve their lexical competence through guided teaching (Flowerdew, 2009). Although there is no significant difference between the experimental group and the control group in terms of learning collocations, the overall performance of both groups improved considerably. Particularly, in terms of the experimental group's experience, it might be weakly assumed that minimal training about how to use concordancing tools enabled learners to use concordance software well enough to conduct independent searches. In that regard, it can be argued that the study might offer some insight into the contemporarily debated research topic of whether teacher-prepared concordance lines or students' use of concordances on their own should be a more efficient way of teaching. It can be inferred from the study that learners' independent and direct use of corpus and concordancing tools have the potential to help learners to have control over their learning and thus boost their self-autonomy (Gaskell & Cobb, 2004; Sun & Wang, 2003).

Statistically not significant, but the relative success of the experimental group can be associated with the novelty effect of corpora, i.e., Hawthorne effect (Levitt & List, 2011). The students had no prior knowledge and experience of using corpus and concordancing activities. They were aware that they were studying a new and engaging tool to study collocations and expected to perform better. Therefore, this novelty effect might have contributed to their relative success in the posttest. Additionally, the rich input provided by concordance lines allowed students to engage actively in target collocations and to expose themselves repeatedly to the collocations. Lee, Warschauer, and Lee's (2018) meta-analysis demonstrates that corpus use improves in-depth vocabulary knowledge more than definitional knowledge or productive useability. In that sense, with regard to the relative success of the experimental group, it can cautiously be argued that corpus tools provide students with easy and ample access to explore the several aspects of a lexical item. Students' active involvement and spending time on the environment of a word increases the thought process, which may lead to more successful vocabulary gains. On the other hand, for the control group, limited access to the example uses and what is involved in better exploring a word did not require deep processing of the input about a word combination. Therefore, they might have scored slightly less in the posttest.

When the results are explored closely, it can be observed that there are interesting points with regard to the developments in different collocation types. Although the results are not statistically significant, it may be assumed that the experimental group's performance on strong collocation type could be associated with the instruction provided for this group through the corpus considering the previous literature about inductive learning (Sun & Wang, 2003). Strong collocations by nature are less frequent but more fixed collocations in comparison to weak ones. As was hypothesized in the literature (Hoey, 2005), strong collocations make the preceding nouns more marked; thus, it takes lesser time to process them than when a noun is preceded by a weaker collocate. Corpus, in that regard, might have allowed the experimental group to observe and explore ample and authentic use of target strong collocations. It seems quite likely that collocations observed in corpus provide cues on which learners can draw easily. Yet of course further research should be conducted to make strong arguments about it. Students' spending time on the collocates increases the thought process which may facilitate learning challenging strong collocations. Students might have found online concordancing motivational

and engaging while focusing on strong collocations. The control group, on the other hand, continued to rely on lexical teddy bears, i.e., weak collocations in our case, as indicated in the literature (Hasselgren, 1994; Siyanova & Schmitt, 2008). Even though they were advanced learners, once again, learners' dependence on the familiar was revealed through their overgeneralized use of the weak collocations.

When we examine the results from a pedagogical perspective, we can offer a combined methodology of corpora and dictionaries to teach collocations. Although today's language learners are '*digital natives*' (Prensky, 2001), who have sophisticated skills to use digital technologies and also developed new cognitive capacities adaptable to these new technologies, it is evident that some paper-based traditional teaching methodologies still apply to some learners' cognitive styles. As Flowerdew (2009) cautioned us, some field-independent students may not enjoy the inductive learning approach that corpus use adopts. Some students in the experimental group might, in this sense, not have had a positive attitude towards inductive discovery learning on account of their cognitive tendency.

There is consensus in the literature that teaching instruction should guarantee learners to develop an extensive repertoire of formulaic sequences – in our case, particularly collocations (Wray, 2002). The findings presented here seem to support this proposition with regards to collocation learning. The current study was conducted by comparing two instructional methodologies while teaching two different types of collocations and the results of the pre and posttests demonstrated that language teachers should combine concordancing activities with dictionary tasks in order to address various learning needs and styles. Web-based activities can also offer new possibilities to supplement the existing teaching materials.

5.1. Implications for Corpus Use in EFL Classes and Exploitation of Corpus for Testing

Two directions of pedagogical implications can be extrapolated from the present study. First, L2 learners are in need of hybrid teaching tools such as web-based tools and dictionaries to compensate for the limitations of each tool when they are used exclusively in an EFL setting. Dictionaries have been in good use for a long time in language classrooms. But a corpus is a relatively new tool for learners and teachers in particular EFL settings. Therefore, corpus tools should be introduced to both teachers and students in order to gain advantages of using corpus-based teaching/learning activities to address the needs of today's digital-native students. However, total reliance on corpus can pose several challenges on students, as warned by Flowerdew (2009). Since corpus use is based on inductive discovery learning, field-independent students might not benefit from corpus use as much as field-dependent learners. At this point, dictionary use with clear instructions would be more fruitful for the setting. The training sessions for the corpus group had three steps: 1) explicitly describing and teaching several corpora and the concordance, 2) demonstrating how the concordancers and collocation search is conducted, 3) having students hands-on practices in a flexible time frame. So if all the students were given a similar training, they would all make most of the use of corpus tools in vocabulary learning. The scope of this study is focused on vocabulary learning; however, corpus tools could be exploited in teaching many language skills such as writing and speaking. Thus, corpus tool, as a new type of learning aid, mediates language learning when appropriate training is provided for students.

The second direction of implications can focus on exploiting corpus as a testing aid. Teaching collocations is a rather challenging task due to the inherently complex nature of collocations. Choosing collocations to be taught is another task that poses difficulties for teachers. For this study, the researcher chose several adjective-noun collocations with various difficulty and frequency levels. An additional challenge is caused by the paradoxical nature of strong and weak collocations exploited. Participants' errors could provide some insight for teachers about what to focus on and how to improve the lacking information regarding the collocation type.

Every teacher might want to build their own learner corpus in order to custom their learners' needs and test their particular proficiency. So, this emerging research field, i.e., exploiting corpus for language testing and assessment, relies on learner corpora which comprise learners' outputs. Learners' errors provide valuable insight for teachers while preparing tests to assess proficiency levels in different constructs of language in the context of CEFR. Using learner corpus improves test content and also decreases the subjectivity of human raters whose holistic ratings are inevitably affected by their value judgment. Thus corpus-driven assessment also helps to validate human raters' claims (See Callies & Götz, 2015 for further research).

6. CONCLUSION

The current study is an attempt to delve into the area of learning collocations using different tools, i.e., corpora and dictionaries. Corpus and concordance programs are powerful tools in EFL settings. According to the results of the study, potential differences in learners' performance on collocation tests and their improvement in learning collocations cannot be attributed only to the corpus-based approach. Dictionaries still contribute to the language learning environment; therefore, a combined approach could be a better choice in studying collocations. Many researchers are strong proponents of corpus use in language teaching, yet some reservations about the benefits of corpus exploitations in language classrooms could still be valid in terms of learners' learning styles and needs (Cook, 1998; Widdowson, 2000). Therefore, traditional teaching materials like dictionaries should be supplemented with concordance programs to improve educational settings to respond to the various needs of language learners.

Although the results were not significant, relatively higher mean scores could still be considered to mean that corpus-based pedagogies may be more suitable for today's generations, who were grown up as computer and Internet literates and thus demand faster and cheaper technologies. In that regard, corpora can be a solution to some problems about vocabulary learning in language classrooms. Concordance-based activities provide learners with a chance to conduct research by allowing them to take on their own learning responsibilities (Johns, 1991) and expose them to authentic language (Biber, 2004). To a certain extent, the results of this study also support the literature postulating that corpus-based vocabulary learning exercises have a positive impact in improving lexical competence (Biber, 2004; Cobb, 1997, 2003). The findings of the study are compatible with the corresponding research in the research field (Cobb, 1997; Anđ, 2006).

However, it should be noted that the researcher is fully aware of the fact that a deeper and more detailed analysis would be necessary regarding the linguistics and psycholinguistics factors that affect the intrinsic difficulty of collocations. Therefore, the results should be regarded with caution. The short period of research time and lack of student training about corpus use were among the several limitations of the current study. Additionally, the sample size was not enough to draw generalizable results. The two groups were divided unevenly due to outliers. The number of participants in the corpus group was fewer in number, which could have impacted the results to be statistically significant. Unfortunately, the small sample size did not allow the researcher to draw reliable conclusions about whether the exploitation of corpus or dictionary could improve collocation learning. Mainly because of the sample size for the type of collocations (weak and strong), the researcher did not have enough statistical power to compute the within-effects between the variables. A follow-up qualitative research study could give us some detailed information about the learners' particular vocabulary choice concerning weak and strong collocations. It is necessary to conduct a more longitudinal study with a larger sample size in different settings to explore the effect of corpus use on the collocational competence of advanced students. For further research, a study based on learner corpora would give a more satisfying insight as to why participants made certain errors in collocational pairs

and how these lexical misselections can contribute to L2 vocabulary gain. Further development in computer technology will definitely spawn more efficient tools for incorporating corpus exploitation in L2 vocabulary learning, which will merit further empirical research.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics (Ethics Committee Approval: 178.233.42.148). The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

ORCID

Hatice Altun  <https://orcid.org/0000-0003-4096-4018>

7. REFERENCES

- Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of make in native and non-native student writing. *Applied Linguistics*, 22(2), 173-194.
- Anderson, W., & Corbert, J. (2017). *Exploring English with Online Corpora*. Palgrave.
- Barlow, M., & Burdine, S. (2006). *Phrasal verbs*. Athelstan Pubns.
- Basal, A. (2019). Learning collocations: Effects of online tools on teaching English adjective - noun collocations [Article]. *British Journal of Educational Technology*, 50(1), 342-356. <https://doi.org/10.1111/bjet.12562>
- Benesch, S. (2001). *Critical English for academic purposes: Theory, politics, and practice*. . Lawrence Erlbaum Associates.
- Bernardini, S. (2002). Exploring new directions for discovery learning. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 165–182). Rodopi.
- Biber, D., & Conrad, S. (1999). Lexical Bundles in Conversations and Academic Prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: studies in honour of Stig Johansson*. Rodopi.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25, 371-405.
- Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: the influence of written language use. *English Language and Linguistics*, 15(2), 223-250. <https://doi.org/10.1017/S1360674311000025>
- Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English : linguistic change in writing*. Cambridge University Press.
- Biber, D., Johansson, S., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.
- Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning*, 67(2), 348-393. <https://doi.org/10.1111/lang.12224>
- Breyer, Y. (2009). Learning and teaching with corpora: reflections by student teachers. *Computer Assisted Language Learning*, 22(2), 153-172.
- Callies, M. (2016). Towards a process-oriented approach to comparing EFL and ESL varieties [Article]. *International Journal of Learner Corpus Research (IJLCR)*, 2(2), 229-251. <https://doi.org/10.1075/ijlcr.2.2.05cal>
- Callies, M., & Götz, S. (2015). *Learner Corpora in Language Testing and Assessment* (Vol. 70). John Benjamins Publishing Company.
- Çelik, S. (2011). Developing Collocational Competence Through Web Based Concordance Activities. *Novitas Royal research on Youth and Language*, 5(2), 273-286.
- Chambers, A. (2005). Integrating corpus consultation in language studies *Language Learning and Technology*, 9 (2), 111–125.
- Chambers, A., & O’Sullivan, I. (2004). Corpus consultation and advanced learners’ writing skills in French. *ReCALL* 16(1), 158–172.

- Chan, T.-p., & Liou, H.-C. (2005). Effects of Web-based Concordancing Instruction on EFL Students' Learning of Verb-Noun Collocations. *Computer Assisted Language Learning*, 18(3), 231-251. <https://doi.org/10.1080/09588220500185769>
- Cheng, W., Warren, M., & Xu, X. (2003). The language learner as language researcher: putting corpus linguistics on the timetable. *System*, 31 (2), 173–186.
- Cobb, T. (2003). Analyzing Late Interlanguage with Learner Corpora: Québec Replications of Three European Studies. *Canadian Modern Language Review/ La Revue canadienne des langues vivantes*, 59(3), 393-424. <https://doi.org/10.3138/cmlr.59.3.393>
- Conzett, J. (2000). Integrating collocation into a reading and writing course. In M. Michael Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 70-87). Language Teaching Publications.
- Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal*, 52, 57-64.
- Cortes, V. (2002). Lexical bundles in Freshman composition. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131-145). Benjamins.
- Crowther, J., Dignen, S., & Lea, D. E. (2002). Oxford collocations dictionary for students of English. In *Oxford collocations dictionary for students of English*.
- Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, 28(2), 130-144. <https://doi.org/10.1080/09588221.2013.803982>
- Davies, M. (2008). *The Corpus of Contemporary American English: 520 million words, 1990-present*. <http://corpus.byu.edu/coca/>.
- Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL*, 47, 157-177.
- Ellis, N. C. (2002). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*, 24, 297-339.
- Ellis, N. C. (Ed.). (2003). *Constructions, chunking, and connectionism: the emergence of second language structure*. Blackwell.
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis* (pp. 1-32). Oxford.
- Flowerdew, J., & Peacock, M. (2001). Issues in EAP: A preliminary perspective. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 8-24). Cambridge University Press.
- Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics*, 14, 393-417. <http://dx.doi.org/10.1075/ijcl.14.3.05flo>
- Foster, K., & Chamber, S. M. (1973). Lexical Access and Naming Time. *Journal of Verbal Learning and Verbal Behavior*, 12, 627-635.
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence: Collocations in Corpus-Based Language Learning Research. *Language Learning*, 67(S1), 155-179. <https://doi.org/10.1111/lang.12225>
- Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors?. *System*, 32(3), 301–319.
- Gilmore, A. (2009). Using online corpora to develop students' writing skills. *ELT Journal*, 63(4), 363-372.
- Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4), 319-335.
- Gui, S., & H., Y. (2002). *Chinese learner English corpus (CLEC)*. Shanghai Foreign Language Education Press.

- Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robins (Eds.), *In Memory of J.R. Firth* (pp. 148-162). Longman.
- Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, 4(2), 237-258.
- Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching Collocation* (pp. 47-69).
- Hill, J., Lewis, M., & Lewis, M. (2000). Classroom strategies, activities and exercises. In M. Lewis (Ed.), *Teaching Collocation* (pp. 88-177).
- Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press.
- Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge.
- Hoffmann, S., Evert, S., Smith, N., Lee, D., & Prytz, Y. B. (2008). *Corpus linguistics with BNCweb: a practical guide*. Peter Lang.
- Howarth, P. (1998). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*. (pp. 161-168). Clarendon Press.
- Hsu, J. T., & Chiu, C. Y. (2008). Lexical collocations and their relation to speaking proficiency of college EFL learners in Taiwan. *The Asian EFL Journal*, 10(1), 181-204.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.
- Hunston, S., & Laviosa, S. (2000). *Corpus Linguistics*. School of English CELS.
- Johns, T. (Ed.). (1991). *Should you be persuaded: two samples of data-driven learning materials* (Vol. 4).
- Jones, S., & Sinclair, J. M. (1974). English lexical collocations. A study in computational linguistics. *Cahiers de lexicologie*, 24, 15-61.
- Koç, G. (2006). *Developing Collocational Awareness* [MA Thesis]. Bilkent University.
- Lai, S.-L., & Chen, H.-J. H. (2015). Dictionaries vs concordancers: actual practice of the two different tools in EFL writing. *Computer Assisted Language Learning*, 28(4), 341-363. <https://doi.org/10.1080/09588221.2013.839567>
- Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25, 56-75.
- Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, 40(5), 721–753. <https://doi.org/10.1093/applin/amy012>
- Levitt, S. D., & List, J. A. (2011). Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments". *American Economic Journal: Applied Economics*, 3, 224-238. <https://doi.org/doi:10.1257/app.3.1.224>
- Lewis, M. (1993). *The Lexical Approach: the state of ELT and a way forward*. Language Teaching Publications
- Lewis, M. (1997). *Implementing the Lexical Approach* Language Teaching Publications.
- Lewis, M. (Ed.). (2000). *Teaching collocations: further developments in the lexical approach*. Thomson.
- Liu, D. (2010). Going Beyond Patterns: Involving Cognitive Analysis in the Learning of Collocations. *TESOL quarterly*, 44(1), 4-30. <https://doi.org/10.5054/tq.2010.214046>
- Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223-242.
- Nesselhauf, N. (2005). *Collocations in a learner corpus*. J. Benjamins Pub. Co.
- O'Dell, F., & McCarthy, M. (2008). English collocations in use: Advanced. In Prensky, M. (2001). *Digital Natives, Digital Immigrants*.
- Prodromou, L. (2003). *Collocations*. Macmillan Publishers.

-
- Sinclair, J. (2001). Preface In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT* (pp. vii-xv). Benjamins.
- Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, 64(3), 429-458.
- Sun, Y.-C., & Wang, L.-Y. (2003). Concordancers in the EFL Classroom: Cognitive Approaches and Collocation Difficulty [doi: 10.1076/call.16.1.83.15528]. *Computer Assisted Language Learning*, 16(1), 83-94. <https://doi.org/10.1076/call.16.1.83.15528>
- Thurstun, J., & Candlin, C. N. (1997). *Exploring academic English: A workbook for student essay writing*. National Centre for English Language Teaching and Research.
- Tribble, C. (2001). *Corpora and language teaching: Adjusting the gaze*. Universite' Catholique de Louvain.
- Tribble, C., & Johns, G. (1990). *Concordances in the classroom*. Longman.
- Tsai, K.-J. (2019). Corpora and dictionaries as learning aids: inductive versus deductive approaches to constructing vocabulary knowledge. *Computer Assisted Language Learning*, 32(8), 805-826. <https://doi.org/10.1080/09588221.2018.1527366>
- Walker, C. P. (2011). A Corpus-Based Study of the Linguistic Features and Processes Which Influence the Way Collocations Are Formed: Some Implications for the Learning of Collocations. *TESOL quarterly*, 45(2), 291-312. <https://doi.org/10.5054/tq.2011.247710>
- Walter, E. (2012). Using a corpus to write dictionaries In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* Routledge.
- Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21, 3-25.
- Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, 21(4), 463-489.
- Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press.
- Wray, A. (2008). *Formulaic Language: pushing the boundaries*. Oxford University Press.
- Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257-283.

The Continuity of Students' Disengaged Responding in Low-stakes Assessments: Evidence from Response Times

Hatice Cigdem Bulut ^{1,*}

¹Cukurova University, Faculty of Education, Department of Educational Sciences, Adana, Turkey

ARTICLE HISTORY

Received: Sep. 01, 2020

Revised: May 20, 2021

Accepted: May 22, 2021

Keywords:

Response time,
Disengaged responding,
Insufficient effort
responding,
Validity,
Low-stakes assessments.

Abstract: Several studies have been published on disengaged test respondents, and others have analyzed disengaged survey respondents separately. For many large-scale assessments, students answer questionnaire and test items in succession. This study examines the percentage of students who continuously engage in disengaged responding behaviors across sections in a low-stakes assessment. The effects on calculated scores of filtering students, based on their responding behaviors, are also analyzed. Data of this study came from the 2015 administration of PISA. For data analysis, frequencies and percentages of engaged students in the sessions were initially calculated using students' response times. To investigate the impact of filtering disengaged respondents on parameter estimation, three groups were created, namely engaged in both measures, engaged only in the test, and engaged only in the questionnaire. Next, several validity checks were performed on each group to verify the accuracy of the classifications and the impact of filtering student groups based on their responding behavior. The results indicate that students who are disengaged in tests tend to continue this behavior when responding to the questionnaire items in PISA. Moreover, the rate of continuity of disengaged responding is non-negligible as can be seen from the effect sizes. On the other hand, removing disengaged students in both measures led to higher or nearly the same performance ratings compared to the other groups. Researchers analyzing the dataset including achievement tests and survey items are recommended to review disengaged responses and filter out students who are continuously showing disengaged responding before performing further statistical analysis.

1. INTRODUCTION

Low-stakes assessments are designed to determine the achievements of students and the factors related to students' achievements. Educational stakeholders shape their strategies and make educational decisions based on the results of many low-stakes assessments, which are conducted regularly in various grade bands. Although these low-stakes assessments provide valuable information for education stakeholders, generally they are not designed to benefit students directly. Thus, students sometimes neglect to perform at their best when answering test and survey/questionnaire items in low-stakes assessments. When students do not devote their full effort, this performance is often referred to as "disengaged responding."

*CONTACT: Hatice Cigdem Bulut ✉ hcyavuz@cu.edu.tr 📍 Cukurova University, Faculty of Education, Department of Educational Sciences, Adana, Turkey

e-ISSN: 2148-7456 /© IJATE 2021

Several studies have been published on disengaged test respondents, and others have analyzed disengaged survey/questionnaire respondents; but these are usually examined separately (Maniaci & Rogge, 2014; Wise, 2017). For many large-scale assessments, such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), students answer test and questionnaire items in succession. In such cases, any low-effort responses threaten the validity of scores obtained from both the test and the questionnaire.

Rates of disengaged responses can reach up to 28% in tests (Wise *et al.*, 2019) and 50% in surveys (Buchanan & Scofield, 2018). Due to these significant numbers, researchers have used various approaches to deal with the negative effects of this threat. Notably, there has been a greater interest in detecting disengaged respondents in achievement tests than in surveys (Soland *et al.*, 2019). Regardless of which instrument of a low-stakes assessment is investigated in terms of disengaged responding, the scores obtained from measures of the students are linked and evaluated accordingly. Students who are strongly motivated in the first session of an assessment may or may not continue to be so engaged in the following sessions and vice versa. Hence, disengaged responding may be considered as a single category in low-stakes assessments. For instance, some students may devote their effort to the first session of an assessment but fail to engage when answering items in the second session of the assessment. Depending on the percentage of students who become disengaged across sessions, data quality from large-scale assessments and resulting conclusions may be considerably affected. It is critical to decide whether we should include disengaged students' responses to make conclusions when creating student profiles. Before answering such a question, it is important to know the percentage of students who are disengaged during an overall session of a low-stakes assessment. However, to the best of the author's knowledge, measures of this factor have not been presented in the literature to date. This study was designed to show the percentage of students who are disengaged during full sessions of low-stakes assessments. Also analyzed is the impact of filtering student data based on their response behaviors concerning item parameter estimation. To begin, disengaged responding behavior is defined, and an overview of prior research on surveys and psychometrics is provided.

1.1. Literature Review

Due to the utilization of technology in assessments, the issue of disengaged responding has received considerable attention. In the survey and measurement research literature, this kind of responding is variously referred to as rapid guessing, low test-taking motivation, effortless responding, disengaged responding, insufficient responding, careless responding, and inattentive responding (Huang *et al.*, 2012; Niessen *et al.*, 2016; Wise & DeMars, 2005; Wise & Kingsbury, 2016; Wise, 2017). Regardless of which term is used, the common idea is that respondents do not devote their full effort or express their real emotions/thoughts when responding on measures. "Disengaged" responding is a construct-irrelevant factor, which threatens the validity of scores (Eklöf, 2006; Wise, 2005). A great deal of research on disengaged responding has demonstrated its negative consequences for scores (Huang *et al.*, 2012; Wise & DeMars, 2005; Wise & Kingsbury, 2016). In response to these findings, new methods have been developed both in survey research and psychometric studies to handle this threat.

1.1.1. Disengaged responding in achievement tests

Before the advent of technological advancements in assessment, early studies used self-reports to measure students' engagement after a test event (Sundre & Moore, 2002; Sundre & Wise, 2003; Wise & DeMars, 2005). Self-reports are vulnerable to potential biases such as social desirability and response biases (Wise & Gao, 2017). However, these studies show that the

validity of scores improves when data cleaning is performed. The availability of recording time on computers led to the emergence of alternative methods for measuring disengaged responding. The new methods allow much more direct detection (Wise & Kong, 2005). This is because, ideally, respondents who intend to devote their full effort are supposed to spend some time on each item to understand it before offering an answer for it. If respondents quickly pass from one item to another, then their responding behavior is rated as lacking in effort (Wise, 2006, 2017). Using computerized testing, it is possible to monitor respondents' behaviors during a test event; with this information, the tester might be able to avoid the effects of disengaged responding on the validity of the scores. Experimental studies (Wise *et al.*, 2006; Wise *et al.*, 2019) show that providing a warning or notification to respondents during testing (about their low engagement with the items) is effective for increasing their engagement.

Researchers have also attempted to suppress disengaged responding behavior in achievement tests using different methods. Some have filtered out disengaged respondents' data (DeMars, 2007; Guo *et al.*, 2016; Wise, 2006, 2019; Wise & Kong, 2005; Wise & Ma, 2012). The results of those studies reveal that filtering increases the validity of the scores. The main concern about filtering data is deciding the cut-off scores while classifying respondents. Methods include fixed measures or visual inspections of items (DeMars, 2007; Wise, 2006), and normative measures to identify responding behavior (Wise & Kong, 2005; Wise & Ma, 2012; Wise, 2019). Regardless of the method, these studies yielded more valid results when they employed filtering. Other studies used response times while estimating parameters (Guo *et al.*, 2016; Meyer, 2010; van der Linden, 2009; Wang & Xu, 2015; Wise & DeMars, 2005). This method of estimating parameters with time data jointly also helped to achieve more precise item and person parameters. Several studies have assessed the consequences of cleaning the data of disengaged respondents by means of different approaches; the overall conclusion is that this is an efficient way to improve model fit and decrease biased parameter estimation in calibration and scoring (Wise & DeMars 2005; Wise & Kong 2005).

These results can help researchers to handle validity concerns for low-stakes assessments. Recent evidence suggests that the rate of disengaged responding can extend to 28% in large scale assessments; the exact rate depends on many factors, such as item positions (Wise *et al.*, 2009), time of the test event (Wise *et al.*, 2013), test structure (Setzer *et al.*, 2013), and the ethnicity and gender of the respondents (Goldhammer *et al.*, 2016). Thus, there appears to be ample evidence that disengaged responding is a validity threat for low-stakes assessments and that it affects conclusions that are based on the scores.

1.1.2. *Disengaged responding in surveys*

Disengaged responding can cause a validity threat for surveys as well. Recent interest in this threat in survey research has been sparked by advancements in online survey platforms (Huang *et al.*, 2012; Zhang, & Conrad, 2014). Disengaged responding behavior in surveys harms the accuracy of conclusions drawn from the scores. However, unlike disengaged responding in achievement tests, disengaged responding in surveys occurs in two ways: when respondents answer items in the survey randomly (Karabatsos, 2003; Meade & Craig, 2012), or when they answer in a non-random way (Johnson, 2005; Meade & Craig, 2012).

Curran (2016) discusses the most efficient methods to detect random and non-random disengaged responding in surveys: (1) *response time*, (2) *long-string analysis*, (3) *Mahalanobis distance*, (4) *odd-even consistency*, (5) *resampled individual reliability*, (6) *semantic antonyms/synonym*, (7) *psychometric antonyms/synonyms*, (8) *inter-item standard deviation*, (9) *polytomous Guttman errors*, (10) *person total correlation*, (11) *bogus/infrequency items*, (12) *attention check item*, (13) *instructional manipulation checks*, and (14) *self-report scales*. Among these methods, response time analysis has been recently utilized by many researchers in this context (Curran, 2016; Huang *et al.*, 2012; Meade & Craig, 2012; Zhang & Conrad,

2014). Several studies show that the rate of disengaged responding varies from 10% to 50% in surveys (Huang *et al.*, 2012; Meade & Craig, 2012; Buchanan & Scofield, 2018; Soland *et al.*, 2019). As with achievement tests, many researchers are using different thresholds for response time data, such as a fixed two-second rule (Huang *et al.*, 2012), 300 milliseconds (Zhang & Conrad, 2014), and a normative method (Soland *et al.*, 2019). Mostly, these studies support the utilization of several methods, in addition to response times, to classify students' responding behavior (Buchanan & Scofield, 2018; Zhang & Conrad, 2014).

Similar to research on disengaged responding behavior in achievement tests, there is a large volume of published studies that discuss removing invalid data in surveys. These studies report that removing that data helps to reduce measurement errors, so that more valid results regarding means, variance, and the reliability of scales may be obtained (Huang *et al.*, 2015; Maniaci & Rogge, 2014; Woods, 2006).

1.1.3. Study objectives

The current state of research indicates that disengaged responding negatively affects both achievement tests and questionnaires in low-stakes assessments. Moreover, consideration of continuity in disengaged responding by students across sections of assessments is lacking in all the aforementioned studies. This indicates a need for investigation across all large-scale assessment events because some researchers use students' responses for all measures. To address this need, this study examines the percentage of students who continuously engage in disengaged responding behaviors across sections in a low-stakes assessment. The effects on calculated scores of filtering students, based on their responding behaviors, are also analyzed. The goal is to assess whether the degree of disengaged responding continuity is significant or negligible and to document the effects on scores obtained from achievement tests and questionnaires.

2. METHOD

The data of this study came from the 2015 administration of PISA (Organisation for Economic Cooperation and Development [OECD], 2017). PISA is a large-scale, international assessment that measures 15-year-old students' achievement in reading, mathematics, and science literacy. After completing these cognitive assessments, students also take a questionnaire that focuses on students' attitudes toward their homes, schools, and learning experiences. Although around 500,000 students took PISA 2015, only 69,426 of the students were included in the analysis based on some selection criteria. These criteria will be explained in the method section. [Table 1](#) shows the selected students' frequency and percentage across countries.

Table 1. Students' frequency and percent across countries.

Country	N	%		N	%
United Arab Emirates	2215	3.2	Lithuania	998	1.4
Australia	2261	3.3	Luxembourg	819	1.2
Austria	1109	1.6	Latvia	770	1.1
Belgium	1471	2.1	Macao	681	1.0
Bulgaria	942	1.4	Mexico	1158	1.7
Brazil	3582	5.2	Montenegro	884	1.3
Canada	3100	4.5	Malaysia	1399	2.0
Switzerland	682	1.0	Netherlands	796	1.1
Chile	1094	1.6	Norway	836	1.2
COL	1830	2.6	New Zealand	742	1.1
Colombia	956	1.4	Peru	1079	1.6
Czech Republic	1039	1.5	Poland	537	.8
Germany	1008	1.5	Portugal	1120	1.6

Table 1. *Continues.*

Denmark	1093	1.6	Qatar	1377	2.0
Dominican Republic	558	.8	B-S-J-G (China)	1524	2.2
Spain	1049	1.5	Spain (Regions)	5046	7.3
Estonia	863	1.2	Massachusetts	256	.4
Finland	910	1.3	North Carolina	270	.4
France	930	1.3	Russian Federation	936	1.3
United Kingdom	2223	3.2	Singapore	944	1.4
Greece	859	1.2	Slovak Republic	956	1.4
Hong Kong	842	1.2	Slovenia	967	1.4
Croatia	924	1.3	Sweden	857	1.2
Hungary	894	1.3	Chinese Taipei	1188	1.7
Ireland	681	1.0	Thailand	1280	1.8
Iceland	498	.7	Tunisia	845	1.2
Israel	1016	1.5	Turkey	928	1.3
Italy	1847	2.7	Uruguay	965	1.4
Japan	1043	1.5	United States	872	1.3
Korea	877	1.3	Total	69426	100.0

2.1. Measures

Science literacy tests: In PISA 2015, the major domain was science literacy, in which there were 67 forms (i.e., booklets), each containing seven science clusters and items related to other domains. Only 21 of these forms prioritized science clusters over the other domains. In this study, five forms (33, 44, 45, 91, 93) were randomly selected from the 21 forms to avoid position and other types of contextual effects among the forms. Data analysis was undertaken in each of the clusters in every five forms.

Student questionnaire: Students’ responses in the cognitive assessments were combined with their questionnaire responses. From the questionnaire, a science-related module including eight scales (see [Table 2](#)) with 51 items was selected.

Table 2. *The science-related module in PISA 2015.*

Scales	Number of Items	Description
ENVAWARE	7	Environmental awareness
ENVOPT	7	Environmental optimism
ENVOPT	5	Enjoyment of science
INTBRSCI	5	Interest in broad science topics
INSTSCIE	4	Instrumental motivation
SCIEEFF	8	Science self-efficacy
EPIST	6	Epistemological beliefs
SCIEACT	9	Science activities

2.2. Procedure

To classify the students as either disengaged or engaged respondents, their response times from the PISA 2015 database were used. Disengaged students were determined based on the normative threshold (NT10) method (Wise & Ma, 2012). NT10 method is one of the most effective methods for determining disengaged respondents in achievement tests (Wise, 2020). For each item, the time threshold is calculated “as a percentage of the elapsed time between when the item is displayed and the mean of the response time distribution for the item, up to a maximum threshold value of 10 seconds” (Wise & Ma, 2012; p. 9). Setzer *et al.* (2013) suggested that spending longer than 10 seconds on an item should not be defined as disengaged responding. By utilizing NT10 method, we classified the students’ engagement for each item

(i.e., 1 = engaged in answering the item; 0 = disengaged in answering the item). Then, we calculated students' total engagement scores (ESs) by summing all the binary classifications generated from the items. Finally, students were classified as "disengaged in the test" if they showed disengaged responding on more than 90% of the items (i.e., .90 threshold) in the test (Wise & Kong, 2005). For example, assume a respondent answered 20 items and this respondent showed disengaged behavior in 10 items based on the NT10 method. Then, this respondent's engagement score would be 10, meaning that the respondent would be classified as "engaged in the test" based on .90 threshold as the score was less than 18.

Disengaged students in the questionnaire were determined using the two-second method proposed by Huang *et al.* (2012). In PISA 2015, there were eight scales in the science-related module presented on a single page. As a result, students' response times included the time spent per scale, not the time per item. Hence, we followed Soland *et al.*'s (2019) approach by calculating the response time for each item as the time spent on the scale divided by the number of items in the scale. Then, we classified the students' engagement in the items separately by using the two-second threshold. Then, we calculated students' total engagement scores (ESs) for the questionnaire and used the .90 threshold again. In this way, students were classified as disengaged in the questionnaire if they showed disengaged responding behavior to more than 90% of the items in the questionnaire. To investigate the impact of filtering disengaged respondents on parameter estimation, three groups were created, namely engaged in both measures (the test and the questionnaire), engaged only in the test, and engaged only in the questionnaire. The group of engaged in both measures will be mentioned as 2, engaged in the test as 3, engaged in the questionnaire as 4, and full sample as 1 in the remainder of the paper.

For data analysis, frequencies, and percentages of engaged students in the sessions were initially calculated. Next, several validity checks were performed on each group to verify the accuracy of the classifications. The idea behind this step was to learn whether or not removing disengaged students made a difference in the parameter estimation, and which group had the highest quality data across the three engagement classification groups. First, all parameter estimations were conducted separately on all groups using the same item response theory modeling approach as PISA utilized, namely the two-parameter-logistic model (2PLM; Birnbaum, 1968) for dichotomously scored responses and Generalized Partial Credit Model (Muraki, 1992) for polytomously scored responses. Besides, classical test theory analysis was carried out with test items. Second, reliability coefficients, effect sizes, and correlations between scores were calculated. Third, fit indices related to factor structures of scales were compared. Note that, the second group was used for all comparisons while reporting results, however, the only third group was taken into consideration when comparisons were done for tests and the fourth group when comparisons were done for questionnaires. All analyses were conducted in R (R Core Team, 2019) using ShinyItemAnalysis (Martinkova *et al.*, 2017), ltm (Rizopoulos, 2006), and lavaan (Rosseel, 2011) packages.

3. FINDINGS

Only the results from the first cluster of Form 33 were reported as similar results were found for the other clusters. The results obtained from other clusters are available from the author upon request. The results showed that although the proportion of disengaged respondents changed across the clusters, a great number of disengaged students in the test also continued their disengaged responding behavior in the questionnaire session. Among the disengaged students who took the test, approximately 38-43% followed the same type of disengagement when responding to the questionnaire items. Specifically, when we look at the proportion in the first cluster (see Table 3), only 49% of students appear to be engaged in both measures. Most students (80%) were engaged in the test session while only 60% of students were engaged in the questionnaire session. This finding suggests that some disengaged students in the test

became engaged respondents in the questionnaire session. Furthermore, Appendix 1 shows student percentages based on responding behaviors across countries in PISA 2015. When we look at the countries in Appendix 1, especially the most successful East Asian countries, they have relatively smaller percentages of disengaged students in the test, but mostly higher percentages of disengaged students in the questionnaire.

Table 3. Descriptive Statistics of Ability Estimates and Engagement Scores (ESs).

Group	N	Ability estimates		ESs based on the test		ESs based on the questionnaire	
		\bar{X}	SD	\bar{X}	SD	\bar{X}	SD
1	2182	-.01	.96	.93	.11	.85	.25
2	1075	.27	.85	.98	.03	1	0
3	1745	.11	.93	.97	.03	.88	.22
4	1278	.19	.88	.94	.10	1	0

Note: 1 = Full sample; 2 = Engaged in both measures; 3= Engaged in the test; 4= Engaged in the questionnaire.

Table 3 shows that the difference in the mean ability estimates between the groups were not negligible, especially between the full sample (group 1) and the group of students engaged in both measures (group 2); Cohen’s d ranged from 0.12 to 0.29 [d1-2=-0.29, d1-3=-0.12, d2-3=.18]. The ability estimates were lowest in the group of students engaged in both measures. This is because easy items tended to get even easier after filtering out students based on their response behaviors. Furthermore, as seen in Figure 1, the test information appears to be much greater for the group of students engaged in both measures in the lower ability range. This is to be expected, given the removal of low-accuracy responses by disengaged students. The same results apply to the scores obtained from the questionnaires. For example, Figure 2 shows the test information functions of EPIST. Since the thresholds tended to be lower after filtering out the related students, the information appears to be generally less in the lower theta range. However, Figures 1 and 2 show that the test information appears to be much greater for the full sample between the -2 and 2 theta range. Therefore, it is possible that item parameters might be overestimated, and the measurement model inflated test information in this range due to the presence of disengaged responses.

Figure 1. Test information functions of the first cluster of the 33rd form

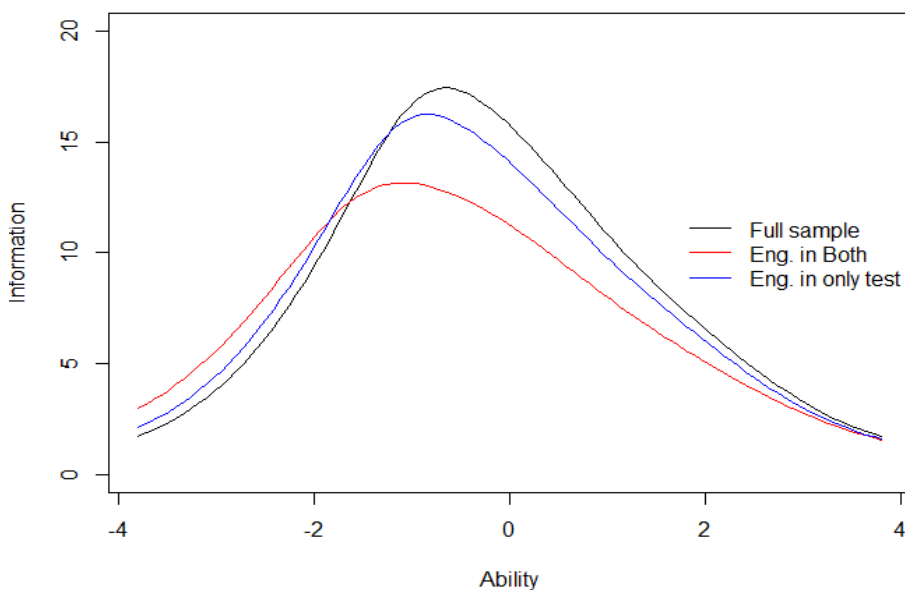
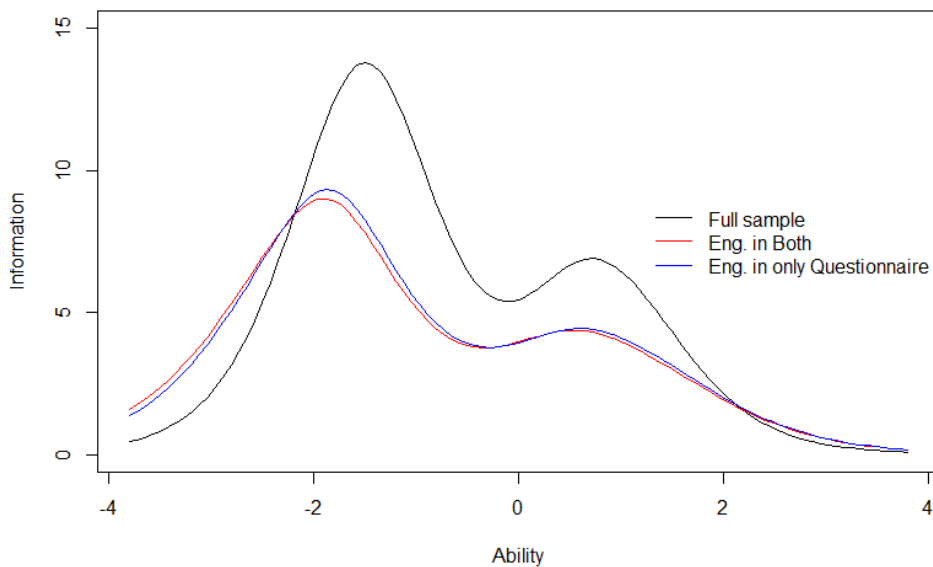


Figure 2. Test information functions of EPIST.

The alpha reliability coefficients for the test (.88, .86, .88 for groups 1, 2, and 3 respectively) and questionnaires slightly changed for the groups (see Table 4), but there were generally no big differences except that the reliability coefficient calculated from the group of students engaged in both measures was significantly lower in value than the others.

Table 4. The reliability coefficients of the first cluster of 33rd form.

	Reliabilities			Significance
	1	2	4	
ENVAWARE	.86	.84	.85	1-2, 1-4
ENVOPT	.87	.84	.85	1-2, 1-4
JOYSCIE	.94	.93	.93	1-2, 1-4
INTBRSCI	.81	.73	.74	1-2, 1-4
INSTSCIE	.92	.92	.92	-
SCIEEFF	.89	.86	.87	1-2, 1-4
EPIST	.88	.83	.83	1-2, 1-4
SCIEACT	.93	.90	.90	1-2, 1-4

Note: 1 = Full sample; 2 = Engaged in both measures; 4= Engaged in the questionnaire

Table 5 shows the correlations of domains and ES which can be interpreted regarding the validity evidence for calculated ESs. The correlations of ability estimates and ES are .24 ($p < .01$), .01 ($p > .05$), .08 ($p < .05$) in groups 1, 2, and 3 respectively. There was a significant low correlation between the ESs obtained from the tests and scales (.23, $p < .01$). In the full sample, calculated ES in the questionnaires was significantly correlated with the students' thetas estimated from the questionnaires. However, these correlations were not significant within each group. As expected, those correlations were significantly lower in the opposite direction in the group of students engaged in both measures. This suggests that both ESs were effective in removing disengaged students who caused a negative significant correlation between the overall ability estimates and thetas.

Table 5. *The correlation coefficients.*

Domains	ES in the questionnaire	Correlations with ability estimates			Cohen's q (1-2), (1-4), (2-4)
		1	2	4	
ENVAWARE	-.17**	-.05*	.01	.03	-.06, -.08, -.02
ENVOPT	-.25**	-.09*	.01	.04	-.09, -.13, -.04
JOYSCIE	-.22**	-.11**	-.02	-.05	-.09, -.06, -.02
INTBRSCI	-.23**	-.10*	.03	-.03	-.14, -.07, .07
INSTSCIE	-.14**	-.03	.05	.07*	-.07, -.10, -.03
SCIEEFF	-.33**	-.13**	.01	.06*	-.14, -.07, .07
EPIST	-.25**	-.14*	.06*	.01	-.08, -.15, -.06
SCIEACT	-.31**	-.16*	.02	.02	-.19, -.18, -.01

Note: 1 = Full sample; 2 = Engaged in both measures; 4= Engaged in the questionnaire

Table 6 shows the model fit indices obtained from the confirmatory factor analysis conducted separately for all the domains in the questionnaire. The confirmatory factor analysis of the fit indices of all the domains shows that a 1-factor model fits the data well. Although there were no big differences between the indices, the indices obtained from the fourth group are slightly better, suggesting that the method based on calculating ES provides good performance for the underlying construct.

Table 6. *The model fit indices of the first cluster of 33rd form.*

Domains	1			2			4		
	RMSEA	CFI	TLI	RMSEA	CFI	TLI	RMSEA	CFI	TLI
ENVAWARE	.11	.94	.91	.09	.96	.94	.10	.94	.91
ENVOPT	.10	.95	.93	.11	.93	.90	.10	.94	.91
JOYSCIE	.05	.99	.99	.03	.1	.1	.04	.1	.1
INTBRSCI	.17	.91	.83	.17	.88	.75	.17	.88	.77
INSTSCIE	.16	.98	.95	.14	.99	.96	.13	.99	.96
SCIEEFF	.07	.97	.96	.05	.98	.97	.05	.98	.97
EPIST	.16	.92	.87	.17	.86	.76	.17	.86	.77
SCIEACT	.17	.87	.83	.18	.82	.76	.18	.83	.77

Note: 1 = Full sample; 2 = Engaged in both measures; 4= Engaged in the questionnaire

Overall, the second group performed better or nearly the same as the third and fourth groups in terms of obtained results. This suggests that even if conservative methods are selected for identifying disengaged respondents, as in this study, some students still may not be assigned to the correct group. That is why the third and fourth groups did not appear to perform much better than the second group. The decision not to filter disengaged students may significantly affect the estimation of the scores in both measures.

4. DISCUSSION and CONCLUSION

Because disengaged responding behavior in tests and questionnaires causes a validity threat, this study was designed to examine the percentage of students who continuously demonstrate disengaged responding behaviors across the sessions of a low-stakes assessment. This paper contributes to research in the field of both questionnaires and tests and applies to disengaged responding generally in large-scale assessments. Another question asked is whether the effects of continuously disengaged behavior are significant or negligible in scores obtained from achievement tests and scales.

The results indicate that students who are disengaged in tests tend to continue this behavior when responding to the questionnaire items in PISA. Moreover, the rate of continuity of disengaged responding is non-negligible as can be seen from the effect sizes. This makes it critical to use large-scale assessments' data for educational decisions and policies without first screening for disengaged responding. Recent studies that focused on data from achievement tests reveal that disengaged responding behaviors affect the country rank orderings of international assessments (Eklöf *et al.*, 2014; Zamarro *et al.*, 2019). Hence, when we consider both cognitive and non-cognitive data sets together, disengaged responding may cause validity issues.

The percentage of students who were engaged in the cognitive part of the assessment in the current study was higher than the percentage of students who were engaged in the non-cognitive part of the assessment. This can be explained using the expectancy-value theory (see Wigfield & Eccles, 2000). According to expectancy-value theory, students' engagement in measures depends on their perceived value for the measure or expectancy for the test. For example, some students might assign more importance or value to the cognitive session (e.g., achievement tests) in the large-scale assessment. Ultimately, this influences their engagement across the sessions. Wise *et al.* (2019) reported a similar situation concerning the initial and final parts of a test. Inconsistencies in engaged responding across sessions of PISA 2015 are more obvious for some countries. Furthermore, several studies support that respondents' cultural backgrounds affect the occurrence of disengaged responses in questionnaires (e.g., Palaniappan & Kum, 2019). Respondents coming from collectivistic cultures tend to show more disengaged responding in questionnaires.

The results of this study also show that the information obtained from both measures appeared to be generally less in the lower theta range within the full sample. Removing disengaged students in both measures led to higher or nearly the same performance ratings compared to the other groups. These results are similar to those of several studies in the literature (Maniaci & Rogge, 2014; Meade & Craig, 2012; Wise & DeMars, 2006; Wise & Kingsbury, 2016), which all suggest that the removal of disengaged respondents' data provides more valid results. Alternatively, methods such as sending warning notifications (see Wise *et al.*, 2006; Wise *et al.*, 2019) to disengaged respondents before upcoming sessions of the assessment can be adopted to promote engagement in those upcoming measures. Further results of this study suggest that removing disengaged students can change the negative significant correlation to a non-significant correlation between the overall ability estimates and thetas. These results highlight an important area of further research.

Although the current study has yielded important results, the examination was constrained by several limitations. The main limitation in this study involves the use of a limited number of (randomly selected) science achievement tests and only the science module in the student questionnaire in PISA 2015. Another limitation involves the methods used to classify the students into engagement groups. As reported by Curran (2016), incurring a Type I error when using conservative methods is inevitable. A further limitation of the study relates to the use of response times for each scale, rather than for each item, during the process of classifying the students in the questionnaire session. This limitation can cause several problems, as Soland *et al.* (2019) indicated, and might ultimately limit the generalizability of the results. Therefore, more research should be conducted, using different low-stakes assessment data that include response times for each item in the questionnaire, and different methods and measures for classifying the students.

In conclusion, the present study unveils that disengaged respondents become a validity threat not only for the inferences of achievement scores but also for the information gathered from student questionnaires. Therefore, researchers analyzing the PISA dataset are recommended to

review disengaged responding behaviors. More importantly, researchers intended to use students' both cognitive and non-cognitive data sets are strongly recommended to filter out students who are continuously showing disengaged responding before performing further statistical analysis.

Declaration of Conflicting Interests and Ethics

The author 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.

ORCID

Hatice Cigdem Bulut  <https://orcid.org/0000-0003-2585-3686>

5. REFERENCES

- Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord and M.R. Novick (eds.), *Statistical theories of mental test scores*. Addison-Wesley.
- Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low-quality data and its implication for psychological research. *Behavior Research Methods*, 2018, (50), 2586–2596. <https://doi.org/10.3758/s13428-018-1035-6>
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4-19. <https://doi.org/10.1016/j.jesp.2015.07.006>
- DeMars, C. E. (2007). Changes in rapid-guessing behavior series of assessments. *Educational Assessment*, 12(1), 23–45. <https://doi.org/10.1080/10627190709336946>
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual review of psychology*, 53(1), 109-132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>
- Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, 66, 643–656. <https://doi.org/10.1177/0013164405278574>
- Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, 27(1), 31–45. <https://doi.org/10.1080/08957347.2013.853070>
- Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers, No. 133). OECD Publishing.
- Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, 29, 173-183. <https://doi.org/10.1080/08957347.2016.1171766>
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114. <https://doi.org/10.1007/s10869-011-9231-8>
- Huang, J.L., Bowling, N.A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, 30, 299–311. <https://doi.org/10.1007/s10869-014-9357-6>
- Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., & DeShon, R.P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27, 99–114. <https://doi.org/10.1007/s10869-011-9231-8>
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39, 103-129. <https://doi.org/10.1016/j.jrp.2004.09.009>

- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16, 277-298. <https://doi.org/10.1207/S15324818AME1604>
- Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61-83. <https://doi.org/10.1016/j.jrp.2013.09.008>
- Martinkova, P., Drabinova, A., Leder, O., & Houdek, J. (2017). ShinyItemAnalysis: Test and item analysis via shiny [Computer software manual]. <https://CRAN.R-project.org/package=ShinyItemAnalysis>.
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17, 437-455. <https://doi.org/10.1037/a0028085>
- Meyer, P. J. (2010). A mixture Rasch model with response time components. *Applied Psychological Measurement*, 34, 521-538. <https://doi.org/10.1177/0146621609355451>
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176. <https://doi.org/10.1002/j.2333-8504.1992.tb01436.x>
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use?. *Journal of Research in Personality*, 63, 1-11. <https://doi.org/10.1016/j.jrp.2016.04.010>
- OECD. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematics, financial literacy and collaborative problem solving*. OECD Publishing. <https://doi.org/10.1787/9789264281820-en>
- Palaniappan, K., & Kum, I. Y. S. (2019). Underlying Causes behind Research Study Participants' Careless and Biased Responses in the Field of Sciences. *Current Psychology*, 38(6), 1737-1747. <https://doi.org/10.1007/s12144-017-9733-2>
- R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL, <https://www.R-project.org/>.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1-25.
- Rosseel, Y. (2011). *lavaan: An R package for structural equation modeling and more* (Version 0.4-10 beta).
- Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a low-stakes assessment. *Applied Measurement in Education*, 26(1), 34-49. <https://doi.org/10.1080/08957347.2013.739453>
- Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, 14(1), 8-9.
- Sundre, D. L., & Wise, S. L. (2003, April). 'Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests. *Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago*.
- van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46(3), 247-272. <https://doi.org/10.1111/j.1745-3984.2009.00080.x>
- Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 68(3), 456-477. <https://doi.org/10.1111/bmsp.12054>
- Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education*, 19, 95-114. https://doi.org/10.1207/s15324818ame1902_2

- Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretations, and implications. *Educational Measurement: Issues and Practice*, 36(4), 52–61. <https://doi.org/10.1111/emip.12165>
- Wise, S. L. (2019). An Information-Based Approach to Identifying Rapid-Guessing Thresholds. *Applied Measurement in Education*, 32(4), 325-336, <https://doi.org/10.1080/08957347.2019.1660350>
- Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10, 1-18. https://doi.org/10.1207/s15326977ea1001_1
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43, 19-38. <https://doi.org/10.1111/j.1745-3984.2006.00002.x>
- Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, 30(4), 343-354. <https://doi.org/10.1080/08957347.2017.1353992>
- Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, 53, 86–105. <https://doi.org/10.1111/jedm.2016.53.issue-1>
- Wise, S. L., & Ma, L. (2012, April). Setting response time thresholds for a CAT item pool: The normative threshold method. *In annual meeting of the National Council on Measurement in Education, Vancouver, Canada*.
- Wise, S. L., Soland, J., & Bo, Y. (2019). The (Non) Impact of Differential Test Taker Engagement on Aggregated Scores. *International Journal of Testing*, 1–21. <https://doi.org/10.1080/15305058.2019.1605999>
- Woods, C.M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28, 189–194. <https://doi.org/10.1007/s10862-005-9004-7>
- Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital*, 13(4), 519–552. <https://doi.org/10.1086/705799>
- Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *In Survey Research Methods*, 8, 127–135. <https://doi.org/10.18148/srm/2014.v8i2.5453>

6. APPENDIX

Appendix 1. Student percentages based on responding behaviors across countries in PISA 2015.

Country	Disengaged respondents' percent in test	Disengaged respondents' percent in the questionnaire	Disengaged respondents' percent in both measures	Performance means estimated in this study	Performance means estimated in PISA 2015
Singapore	14.09	56.91	8.47	0.60	556
Estonia	16.51	76.52	13.02	0.51	523
Hong Kong	9.97	29.91	2.60	0.51	534
Chinese Taipei	17.59	53.92	10.61	0.50	532
Japan	16.30	43.37	6.92	0.49	538
Massachusetts	11.33	41.11	5.53	0.45	-
B-S-J-G (China)	24.28	53.75	10.84	0.44	518
Finland	9.56	32.69	3.97	0.41	531
Macao	29.66	51.03	13.82	0.40	529
Germany	16.17	45.96	7.52	0.33	513
New Zealand	13.10	36.13	5.14	0.33	509
Canada	11.48	51.47	6.10	0.31	528
Netherlands	9.80	33.29	6.56	0.31	509
Belgium	13.60	31.63	4.04	0.28	502
Korea	13.11	80.83	12.06	0.28	516
Ireland	12.19	21.94	3.13	0.26	503
United Kingdom	9.36	38.59	4.92	0.24	509
Spain (Regions)	18.23	23.48	3.66	0.22	-
Switzerland	16.57	45.12	10.80	0.21	493
Spain	14.90	24.25	4.29	0.21	501
Norway	15.07	39.53	6.77	0.21	498
Poland	20.02	20.95	3.49	0.21	493
Austria	16.13	36.40	7.33	0.20	495
France	8.93	34.40	3.67	0.20	495
Czech Republic	10.61	31.66	2.96	0.18	481
Italy	10.49	30.71	3.50	0.18	493
Australia	11.63	44.65	4.92	0.17	510
Slovenia	11.07	33.86	4.81	0.16	513
Sweden	22.87	38.86	8.53	0.14	493
Russian Federation	20.51	32.56	6.64	0.12	487
Portugal	18.84	38.10	5.43	0.10	501
Denmark	16.74	41.38	6.64	0.09	496
United States	15.37	30.22	6.16	0.09	502
Hungary	9.62	35.21	2.94	0.08	477
Luxembourg	24.54	34.89	9.07	0.06	483
Iceland	13.33	49.07	4.83	0.05	-
Latvia	19.08	42.74	8.38	0.05	473
North Carolina	10.52	29.27	3.31	0.05	490
Israel	34.55	45.58	15.16	0.00	467
Croatia	10.28	34.53	2.35	-0.01	475
Lithuania	9.02	37.92	3.80	-0.09	475
Slovak Republic	13.08	43.19	5.57	-0.10	461
Greece	22.12	29.44	7.83	-0.11	455
Chile	23.40	28.71	7.03	-0.14	447
Malaysia	17.87	23.53	4.07	-0.17	-
Bulgaria	24.52	47.21	11.72	-0.27	446
Uruguay	28.91	40.86	10.61	-0.36	435
United Arab Emirates	21.81	45.35	12.60	-0.39	437

Appendix 1. Continues.

Country	Disengaged respondents' percent in test	Disengaged respondents' percent in the questionnaire	Disengaged respondents' percent in both measures	Performance means estimated in this study	Performance means estimated in PISA 2015
Thailand	17.66	53.21	7.46	-0.40	421
Turkey	14.66	53.10	7.94	-0.48	425
Montenegro	22.17	56.53	13.64	-0.52	411
COL	27.10	22.39	5.44	-0.53	416
Colombia	34.72	21.99	5.38	-0.53	416
Mexico	26.99	20.87	6.28	-0.53	420
Qatar	38.42	57.61	27.75	-0.64	418
Brazil	58.82	58.16	35.97	-0.68	401
Peru	51.99	13.72	6.24	-0.72	397
Tunisia	34.91	43.22	14.82	-0.85	386
Dominican Republic	54.12	32.31	18.00	-1.11	332

The Use of Exploratory Graph Analysis to Validate Trust in Relationships Scale

Akif Avcu ^{1,*}

¹Educational Sciences Department, Marmara University, Istanbul, Turkey

ARTICLE HISTORY

Received: Nov. 26, 2020

Revised: Apr. 05, 2021

Accepted: May 25, 2021

Keywords:

Trust in relationships,
Exploratory graph analysis,
Network psychometrics.

Abstract: Today, various methods have been developed with a purpose to determine the number of factors underlying a construct. However, there is no definitive agreement on which techniques to be preferred to extract the underlying dimensions. To this end, Exploratory Graphical Analysis (EGA), a recently proposed method, has been compared with traditional methods and the results have revealed that the EGA is less affected from conditions like sample size and inter-dimensional correlation. Besides, it provides more stable results across different conditions. Considering the attractive opportunities it offers, this method has taken its place in the literature as a remarkable alternative to traditional methods. The EGA provides unique outputs compared to other factor extraction techniques. Considering this, interpreting the results obtained within this new and promising framework is assumed to contribute to validation studies. Based on this reality, this study aims to apply the EGA method to Trust in Relationships Scale (TRS) and therefore to contribute to its validity. The investigation of TRS's reliability and validity has already been documented, presenting research opportunities to researchers in the field of positive psychology. The results revealed that, the EGA produces dimensionality structures identical to confirmatory factor analysis (CFA) and exploratory factor analysis (EFA). In addition, further psychometrical indicators within the framework of network analysis are provided. The findings of the study are believed to contribute to the validity of the already existing Trust in Relationships Scale.

1. INTRODUCTION

Uncovering the latent structure underlying human behavior and cognitive abilities is important in social science studies with a very old history. Deciding on the number of underlying dimensions of human behaviors or abilities was firstly made possible by the development of factor analysis (Spearman, 1904). Since its invention, factor analysis technique has become widely popular among researchers. Today, examining the structure of underlying latent traits or dimensions in multivariate data is an important issue in the process of designing and validating assessment tools in psychology (Timmerman & Lorenzo-Seva, 2011). Currently, factor analysis is an inevitably most widely used method as one of the first steps routinely applied in the process of studying construct validity (Osborne & Costello, 2009). Furthermore,

*CONTACT: Akif Avcu ✉ avcuakif@gmail.com 📍 Educational Sciences Department, Marmara University, Istanbul, Turkey

investigating the underlying dimensions of constructs is very important for a better understanding of characteristics of individuals and human behavior (Garcia-Garzon et al., 2019).

Today, various methods have been developed for factor extraction decisions. Traditionally, Kaiser's eigenvalue greater than 1 (K1) rule (Guttman, 1954; Kaiser, 1960) and scree plot test (Cattell, 1978) are the most common methods. This popularity is somewhat related to their old history and availability in most of the statistical software. Bandalos and Boehm-Kaufman (2009) state that most of the commercial software programs present K1 rule as default option factor extraction decisions. In addition, parallel analysis (PA) technique (Horn, 1965) and minimum average partial (MAP) technique (Velicer, 1976) are other commonly used methods.

Studies conducted in the past have shown that PA and MAP methods provide more robust and accurate results for factor extraction decisions (i.e. Ledesma & Valero-Mora, 2007; Osborne et al., 2008). However, there is no definitive agreement on which technique should be preferred to unveil the underlying dimensions. The studies carried out indicate that each of these techniques has their own limitations (Garrido et al., 2013; Keith et al., 2016; Velicer et al., 2000; Lubbe, 2019). This ambiguity reveals the necessity of developing new techniques in order to obtain more accurate estimates when deciding on the number of dimensions.

In response to this necessity, the efforts to develop new factor extraction techniques by researchers still continue today. The EGA is a recently proposed method and has already been compared with traditional techniques (Golino & Epskamp, 2017). Accordingly, the results of such studies revealed that the EGA provides comparable results to the traditional methods and outperforms them when the number of dimensions is higher when the number of items is less and the correlation between dimensions are higher. In addition, it has been reported that EGA's precision shows less fluctuation across different conditions like sample size and inter-dimensional correlation. All these results prove its robustness.

1.1. Overview of the EGA Approach

In a recently published study Golino and Epskamp (2017) introduced a new approach as an alternative to factor analysis. This method called as the EGA uses network psychometric to determine the number of dimensions in psychological data. Network psychometrics recently been adapted the network modeling approach to the quantitative field in psychometrics (Epskamp et al., 2017). In these network models, nodes represent random variables. These variables correspond to items in measurement instruments. Nodes are connected by edges or links and show the level of interaction between these variables. These models focus on the prediction of direct relationships between these variables rather than defining the observed variables as a function of a latent common cause. This approach extracts the dimensions by clustering the variables in the dataset.

The EGA uses undirected network models. In this method, the focus is on the estimation of the number of dimensions in the psychological datasets of undirected network models called Markov Random Fields (Lauritzen, 1996). EGA models are based on the Gaussian Graphical Model (GGM) and directly model the multivariate normal distribution network with a reverse covariance matrix. Each unit of the inverse covariance matrix corresponds to the edge. These edges can be standardized and visualized and the link between the two variables can be interpreted as associations between the nodes. (Lauritzen, 1996).

The use of partial correlations is the most common approach used for the estimation of network models; however, it poses an important problem in itself: Even if two variables are conditionally independent, the estimated coefficient is not possibly being estimated as zero due to sample variability (Epskamp & Fried, 2016). Even if there is no conditional association between the two nodes, the resulting estimated correlation value can be slightly different from zero. In this

case, partial correlation may reflect spurious correlations. This problem can be solved by using regularization techniques such as the least absolute shrinkage and selection operator (LASSO) algorithm as described by Tibshirani (1996). With the use of LASSO, the parameters corresponding to the low relationship between node pairs are estimated to be exactly zero and estimation of a model provides sparser networks. In this way, the interpretability of the network structure becomes easier and more meaningful. Because of these features, LASSO estimation has gained popularity as a preliminary analysis for the prediction of network models (van Borkulo et al., 2014). The level of correction, formally expressed as regularization, is determined by a tuning parameter to estimate GGM. Using this penalty approach, the researcher can avoid the risk of model overfitting, control the sparsity of the network and produce an optimum network model that diminishes the Extended Bayesian Information Criterion (EBIC) (Chen & Chen, 2008). The tuning parameter is set by the researcher before the analysis process starts.

In general terms, the EGA works as follows: firstly, the correlation values between the observed variables are calculated; then, using the LASSO estimation, a sparse inverse covariance matrix is obtained; and using the walktrap algorithm (Pons & Latapy, 2005), the number of dense subgraphs (factors, communities or clusters) is specified using the partial correlation matrix calculated in the previous step.

The walktrap algorithm provides a measure of the similarities between vertices based on random walks that can extract the community/cluster structure in the graph (Pons & Latapy, 2005). The number of clusters identified corresponds to the number of latent factors in the dataset. These sub-graphs are undirected weighted networks in clusters. As a result of this process, the number of factors underlying the latent trait of interest and the size of each item's associations with the rest of items are estimated and presented in a graph consisting of nodes and edges. Traditionally, nodes are represented with green or blue circles in the graph. In addition, thinness of edges gives information about the association between node pairs as the association gets stronger, the lines get thicker.

2.1. Aim of the Study

Considering the attractive opportunities it offers, this method has taken its place in the literature as a remarkable alternative to traditional factor extraction techniques. Based on this reality, the aim of this study is to apply the EGA method with a real data set and to contribute to the validity of the Trust in Relations Scale by interpreting the obtained findings within this new and promising framework.

2. METHOD

2.1. Participants

A total of 736 university students were included in the current study. They were being selected from a large state owned university in a Metropolis in Turkey. The data were collected from the participants via an online data collection platform. Even online data collection posits some challenges to the validity of results (Al-Salom & Miller, 2017), the 2020 pandemic outbreak led universities to continue their education via online classes which made it impossible to collect data by meeting face-to-face. The participants were informed about the voluntary nature of participation and security of the information they provided to minimize those possible threats against the validity of the study. After completion of the data collection of process, 28 questionnaires were decided not to be included in the final dataset because incomplete information was available in them. The final dataset was composed of 611 (83%) female and 125 males (17%). Their ages varied between 18 and 27 (Mean=20.25±1.85).

2.2. Instrument

Trust in Relations Scale (TRS) was developed by Demirci and Ekşi (2018). The scale has two dimensions: trust and reliability. Each dimension is composed of five Likert type items. The dimensionality of TRS was evaluated with EFA and CFA. According to the results of EFA, two factors were extracted, explaining 54% of the total variance. In addition, CFA results also confirmed two dimensional structure of TRS [χ^2 (34, N = 450) = 63,40, $p < .001$; CFI = .99; NFI = .98; SRMR = .033; RMSEA = .044]. The criterion related validity of TRS was tested with the PERMA well-being scale (Demirci, Ekşi, Dinçer & Kardaş, 2017) and results revealed significant correlations between trust in relationships and well-being. The reliability of TRS was investigated by estimating Cronbach alpha coefficient and test-retest reliability. The results suggest that both trust and reliability sub dimensions have good internal consistency and stability of test scores over time.

2.3. Analysis

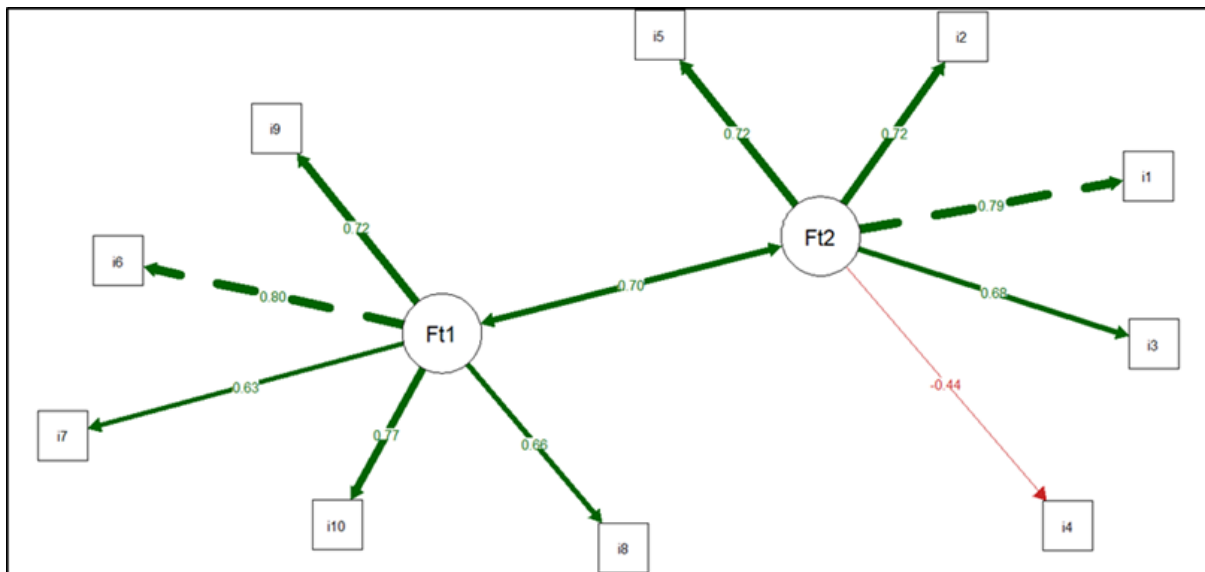
As stated previously, this study was carried out to investigate the underlying dimensionality of TRS using the EGA. The "EGAnet" package developed by Golino and Christensen (2020) was used. The package is available in the R environment (R Core Team, 2019). In addition, conventional EFA and CFA were also conducted for comparison. First of all, CFA was carried out using "lavaan" package (Rosseel, 2012) to analyze the dimensional structure of TRS. A further EFA was run with SPSS 21 to investigate the factor structure. Later, the network structure of TRS was examined based on the GLASSO algorithm using the "EGA" function. Because LASSO procedure includes the use of EBIC, a tuning parameter needs to be selected to control the sparsity of estimated network. For the current study the parameter was set as 0.5 which is used as default option in "EGAnet" package. By conducting this analysis, graphical model and edge weights were calculated. The weight matrix can be obtained with "EGA.estimate" function. After the model was estimated and two dimensional structure was obtained, "dimStability" function was used to examine the structural consistency of the predicted network model and the stability of the items in the extracted dimensions. After this inspection, "bootEGA" function was used to obtain the estimated network structure based on the bootstrap method. After obtaining bootstrapped model, factor loadings, termed as standardized node strengths, were calculated by using "net.loads" function followed by obtaining item stability statistics which indicate reliability of the scale. As a last step, EGA based standardized and unstandardized factor scores were calculated and compared with the conventional raw scores.

3. RESULTS / FINDINGS

3.1. Examining the dimensionality of TRS with CFA

Before estimating the network structure with the EGA, CFA was performed to provide evidence for the two-dimensional structure of TRS. The results showed that data fit well to the model [χ^2 148.328, df =34.000, χ^2/df =4.36, CFI=0.960, TLI=0.947, NFI=0.949, NNFI=0.947, RMSEA=0.068, SRMR, 0.041]. If "lavaan" (Rosseel, 2012) package is already installed and called with "library()" function in R program, "EGAnet" package provides a function to run CFA with the function "lavTestLRT" without running additional codes with another package. The graph for the CFA analysis was presented in Figure 1. In Figure 1, Ft1 represents trust dimension and Ft2 represents reliability dimension. Negative relationship was obtained only for item 4 (as inferred from redline between Ft2 cluster and item 4) in Trust dimension. This result was expected because the 4th item is negatively worded.

Figure 1. Dimensions Estimated via CFA.



Moreover, a further EFA was conducted to examine the underlying dimensionality of TRS. Results supported a two dimensional structure while these two dimensions explained 59.7% of the total variance. As in the CFA, EFA results yielded similar results: The first five items were retained in the first dimension and the second five items were in the second dimension.

3.2. Estimating Edge Weights Matrix

The EGA was estimated by using the GLASSO algorithm which estimated the model based on partial correlations and using penalty approach to obtain sparser networks. The EGA process primarily begins with the calculation of the weight matrices of the edges between the nodes. The estimated values are given in Table 1. The highest edge weight values are between item6-item10 and item1-item2 pairs. Higher values imply that these item pairs showed relatively higher associations. Table 1 also includes many zero values. These values result from the absence of links between the corresponding item pairs and occur due to applying LASSO algorithm. For example, there has been no connection of item 8 with items 2, 3, 4 and 5.

Table 1. Symmetric network edge weights estimated using GLASSO.

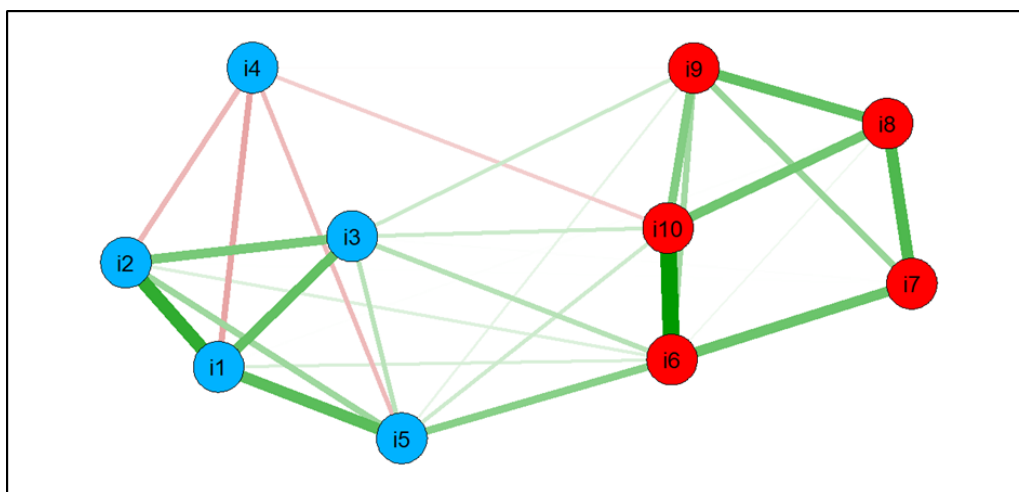
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Item 1	-	0.32	0.24	-0.14	0.26	0.06	0.00	0.00	0.00	0.00
Item 2		-	0.21	-0.11	0.15	0.05	0.00	0.00	0.00	0.00
Item 3			-	0.00	0.11	0.11	0.01	0.00	0.08	0.08
Item 4				-	-0.11	0.00	0.00	0.00	-0.00	-0.08
Item 5					-	0.18	0.00	0.00	0.04	0.08
Item 6						-	0.23	0.02	0.14	0.40
Item 7							-	0.27	0.16	0.00
Item 8								-	0.24	0.22
Item 9									-	0.19
Item 10										-

After obtaining weight matrix, the EGA model was graphed based on the estimated partial correlations. For this process, Walktrap algorithms, were used. This graphical presentation of estimated model is given in Figure 2. The resulting dimensions coincide with the original dimensional structure of the TRS scale. Accordingly, the first 5 items and the last 5 items of the

TRS scale constitute two different clusters. The partial correlations between items in reliability dimension (as represented with red lines) are relatively higher (as inferred from the thickness of lines). At the same time, the red lines in the network graph show that the relationships between the 4th item of trust dimension and the other items are negative this item is negatively worded. Also, this item is negatively related with the 10th item which belongs to reliability dimension.

The thickness of the edges between the items located in the same cluster is an indication of the homogeneity of the clusters. Although the relationships between the items in different dimensions are relatively thinner, the 5th and 6th items are connected with a relatively thicker edge. This implies that even those two items are not in the same dimensions, their associations are relatively higher. In addition, it was found that even located in the same clusters item pairs 1-2 and 7-8 are connected with relatively thinner lines.

Figure 2. The dimensions estimated using exploratory graph analysis.



3.3. Estimating Standardized Node Strengths

It was stated that node strengths are equivalent to factor loadings (Christensen, Golino & Silvia, 2019). Accordingly, they are regarded as the association of each node to the cluster to which it belongs. For the current study, these values were obtained by using the “*net.loads*” function. The obtained standardized node strengths of TRS items are given in Table 2. Accordingly, node strength values vary between 0.38 and 0.31 for the reliability dimension, while these values range between 0.48 and -0.18 for the trust dimension. On the other hand, as expected, each items’ association with the dimension it does not belong to is relatively weaker.

Table 2. Standardized Node Strength for TRS Items.

Item #	Reliability	Trust
Item 6	0.37	0.20
Item 7	0.31	0.00
Item 8	0.36	0.00
Item 9	0.34	0.06
Item 10	0.38	0.12
Item 1	0.03	0.48
Item 2	0.03	0.39
Item 3	0.13	0.28
Item 4	-0.04	-0.18
Item 5	0.14	0.31

3.4. Structural Consistency of TRS

Another concept related to standardized node strengths is structural consistency values. Structural consistency values are calculated for each dimension by evaluating the rate of times that items staying in the same dimension are indicative of the internal consistency of the clusters. In this process, the bootstrap technique was used by taking subsamples from empirical correlation matrix. Structural consistency values indicate the proportion of the times that items are located in the correct dimensions across iterations. It is possible to interpret these values as Cronbach Alpha values.

Prior to calculating the structural consistency, the first step is to apply bootstrap analysis. It can be performed using the “bootEGA” function. The analysis was performed with 500 replications as recommended by Golino and Christensen (2020). The predicted structural consistency value for the first dimension was 0.994 and the estimated structural consistency value for the second dimension was 0.998. These values show the dimensions to be consistently extracted as same rate.

By this analysis, it is also possible to examine the item-level consistencies to identify items that prevent the dimensions from being perfectly structurally consistent. Those statistics are similar to conventional reliability values if an item is deleted. Item level stability statistics values are given in Table 3. As can be seen in Table 3, items 6 and 10 for reliability dimensions were extracted in the unidentified 3rd dimension for 0.6% of replications. Likewise, item 3 in the trust dimension was located similarly in a different 3rd dimension for 0.2% of these replications. Those items could be regarded as distorting the stability of dimensions. Overall, those results suggested that almost all of the replications that located the items in their original dimensions except for the only ignorable rate of replications provided results that produce different structures.

Table 3. *Item Stability Across Dimensions of TRS.*

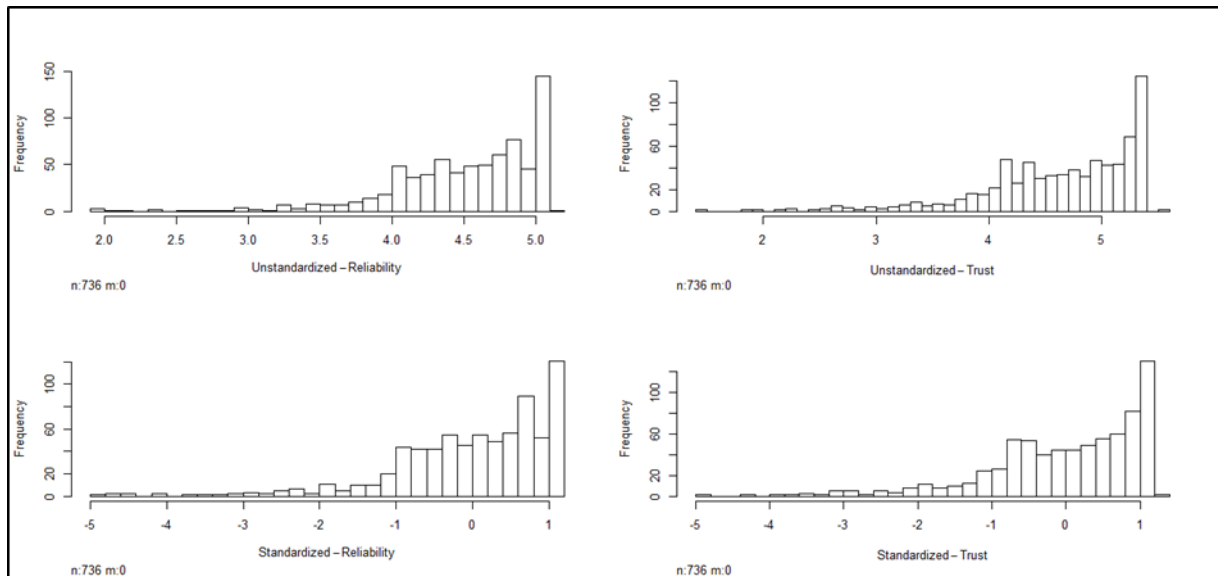
Item #	Reliability	Trust
Item 10	0.99	
Item 6	0.99	
Item 9	1	
Item 8	1	
Item 7	1	
Item 3		0.99
Item 5		1
Item 4		1
Item 2		1
Item 1		1

3.5. Obtaining Scores Based on the EGA Framework

It is also possible to obtain standardized and non-standardized network scale scores of individuals using the “*net.scores*” function available in the package. Network scores are calculated based on the node strength values within each factor. In the CFA approach, scores are generally calculated using a simple structure (items loaded on only one factor) and some regression-based techniques. As to the EFA approach, factor scores are calculated using saturated model approach where each item is allowed to be loaded on more than one factor. On the other hand, scores computed in network models are calculated using a complex structure and can be considered as a weighted composite rather than a latent factor (Christensen & Golino, 2021).

The distribution of the scores obtained for each sub-dimension of the TRS scale was provided in Figure 3. In addition, Pearson correlation coefficients between standardized scores and conventional raw scores were calculated to examine the relationship between the network scores and the observed raw scores. For the Trust sub-dimension, this value was estimated as 0.89 and, for the reliability sub-dimension, this value was estimated as 0.86. This finding suggested that estimated network model of TRS with the EGA approach provided similar ability scores with conventional observed raw scores.

Figure 3. *Standardized and Unstandardized Network Scores of TRS scale for Each Dimension.*



4. DISCUSSION and CONCLUSION

In this study, the factor structure TRS, which had been already reported as having two dimensional structure by Demirci and Ekşi (2018), was re-analyzed with a recently proposed EGA approach. In this study, the factor structure of TRS was firstly examined with the EFA and CFA. The results of these analyzes were found to be in line with the original two dimensional solution. After this preliminary checking, EGA model based on the network psychometric approach was estimated using GLASSO and the original two factor solution was therefore supported. Afterwards, the bootstrap method was used to see the stability of this predicted model and the results revealed that the stability of the model was 99%. These results further provided additional evidence and contribution to construct validity of TRS. In addition, these analyzes implied item-level stability for TRS. Finally, in this study, the raw scores obtained with the classical approach were compared with the standardized scores obtained with the EGA method, and they were found to be highly correlated and comparable. Regarding these findings, TRS can be inferred to be a valid scale within the network modeling perspective.

The results obtained in this study are consistent with the findings that traditional EFA and CFA yielded. This consistency supported EGA to be an alternative technique that can be preferred during validation studies. Further, considering the richness and the novelty of the output that the EGA provides, it can be said that researchers gain more insights into psychometrical properties of the construct they aim to validate by adopting EGA in their studies. To put it more clearly, EGA provides network graph for visual representations of the interconnectedness of items and help researchers about item level stability statistics. These outputs are easy to interpret and provide unique outputs for researchers to draw important implications for the factor structure of psychological constructs.

This study has shown that EGA is a considerable alternative to the methods traditionally used in the investigation of the underlying dimensionality of psychological latent traits. In this regard, the findings obtained by this study are similar to those of relevant literature (Golino & Epskamp, 2017; Golino & Demetriou, 2017) in terms of the similarity of the number of dimensions both EGA and traditional approaches extracted.

Existing literature on EGA has generally focused on modeling dichotomous items. On the other hand, in this study polytomous items were used. Considering that most of the assessment tools in Psychology use polytomous items, it can be inferred that this study contributes to our understanding for using EGA with polytomous items in the field of psychology. The psychological network approach offers a different way of understanding the psychological structures than the traditional methods. When the findings obtained in this study are considered together with the findings obtained in previous studies, it is clear that EGA can be used in many different subareas of psychology (Borsboom & Cramer, 2013; Kossakowski et al., 2015). Moreover, the EGA method can offer significant advantages compared to traditional factor extraction techniques as it offers an advantage of visual representation of observed relationship patterns between variables.

In this study, dimensional structure of TRS was analyzed. The scale is composed of five point Likert items, which are widely preferred in psychological instruments. On the other hand, since the effect of the number response category on EGA estimates is unknown, as also suggested in a previous study by Golino and Demetriou (2017), the effect of the number of response categories needs to be investigated. Such a study, could enlarge the applicability of EGA to different testing conditions. For this reason, in the future studies, the effectiveness of EGA in investigating the underlying factor structure can be examined by using alternative measurement tools with differing response categories (3 or 7 may be preferred). Overall, it would be useful to compare the similarities of the estimated factor loading of conventional methods with the node strengths and also simulation studies under which conditions this similarity may be affected should be conducted.

Acknowledgments

No acknowledgments of people, grants, and funds to declare.

Declaration of Conflicting Interests and Ethics

The author 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 author. **Ethics Committee Number:** Marmara University/Institute of Educational Sciences, 2000310195.

ORCID

Akif AVCU  <https://orcid.org/0000-0003-1977-7592>

5. REFERENCES

- Al-Salom, P., & Miller, C. J. (2017). The problem with online data collection: predicting invalid responding in undergraduate samples. *Modern Psychological Studies*, 22(2), 2.
- Bandalos, D. L., & Boehm-Kaufman, M. R. (2009). Four common misconceptions in exploratory factor analysis. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (p. 61–87). Routledge/Taylor & Francis Group.
- Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences* (1st ed). Plenum Press.
- Chen, J., Chen, Z., (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95, 759-771. <https://doi.org/10.1093/biomet/asn034>

- Christensen, A. P., Golino, H. F., & Silvia, P. (2019). A Psychometric Network Perspective on the Validity and Validation of Personality Trait Questionnaires. *PsyArXiv*. <https://doi.org/10.1002/per.2265>
- Christensen, A. P., & Golino, H. (2021). On the equivalency of factor and network loadings. *Behavior research methods*, Advance online publication. <https://orcid.org/10.3758/s13428-020-01500-6>
- Demirci, İ. & Ekşi. H. (2018). Keep calm and be happy: A mixed method study from character strengths to well-being. *Educational Sciences: Theory & Practice*, 18(29) 303–354.
- Demirci, İ., Ekşi, H., Dinçer, D. ve Kardaş, S. (2017). Beş boyutlu iyi oluş modeli: PERMA Ölçeği'nin Türkçe formunun geçerlik ve güvenirliği. *The Journal of Happiness & Well-Being*, 5(1), 60-77.
- Epskamp, S., & Fried, E.I. (2016). A tutorial on estimating regularized partial correlation networks. *PsyArXiv*, 1607.01367.
- Epskamp, S., Rhemtulla, M., Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, 82, 904-927. <https://doi.org/10.1007/s11336-017-9557-x>
- Garcia-Garzon, E., Abad, F. J., & Garrido, L. E. (2019). Searching for g: A new evaluation of spm-ls dimensionality. *Journal of Intelligence*, 7(3), 14. <https://doi.org/10.3390/jintelligence7030014>
- Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at horn's parallel analysis with ordinal variables. *Psychological Methods*, 18(4), 454-74. <https://doi.org/10.1037/a0030005>
- Golino, H., & Christensen, A. P. (2020). *EGAnet: Exploratory Graph Analysis -- A framework for estimating the number of dimensions in multivariate data using network psychometrics*. R package version 0.9.4.
- Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One*, 12(6), e0174035.
- Golino, H., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., Thiagarajan, J. A., & Martinez-Molina, A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*, 25(3), 292-320. <https://doi.org/10.1037/met0000255>
- Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika*, 19, 149-161. <https://doi.org/10.1007/BF02289162>
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185. <https://doi.org/10.1007/BF02289447>
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141-151. <https://doi.org/10.1177/001316446002000116>
- Keith, T. Z., Caemmerer, J. M. & Reynolds, M. R. (2016). Comparison of methods for factor extraction for cognitive test-like data: Which overfactor, which underfactor? *Intelligence*, 54, 37-54.
- Lauritzen, S. L. (1996b). *Graphical Models*. *Oxford Statistical Science Series*. volume 17. Oxford University Press.
- Ledesma, 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), 1-11. <https://doi.org/10.7275/wjnc-nm63>

-
- Lubbe, D. (2019). Parallel analysis with categorical variables: Impact of category probability proportions on dimensionality assessment accuracy. *Psychological Methods*, 24(3), 339–351. <https://doi.org/10.1037/met0000171>
- Osborne J.W. & Costello (2009). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review*, 12(2), 131-146.
- Osborne, J., Costello, A. & Kellow, J. (2008). Best practices in exploratory factor analysis. In Osborne, J. (Ed.), *Best practices in quantitative methods* (pp. 86-99). SAGE Publications, Inc.
- Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10, 191-218. https://doi.org/10.1007/11569596_31
- R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 9(1), 23–29. <https://doi.org/10.1027/1614-2241/a000051>
- 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>
- Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15, 201–293. <https://doi.org/10.2307/1412107>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 267-288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- 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>
- van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, 4, 5918. <https://doi.org/10.1038/srep05918>
- Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3), 321–327. <https://doi.org/10.1007/BF02293557>
- Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (p. 41–71). Kluwer Academic/Plenum Publishers.

Determination of cyber accessibility of teacher made tests/exams

Gulden Ozdemir ¹, Atilla Ozdemir ^{2,*}, Selahattin Gelbal ³

¹Ministry of National Education, Ankara, Turkey

²Süleyman Demirel University, Faculty of Education, Department of Mathematics and Science Education, Isparta, Turkey

³Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: Aug. 19, 2020

Revised: Apr. 03, 2021

Accepted: May 26, 2021

Keywords:

Teacher-made exams,
Cyber accessibility,
Exam reliability,
Teacher made tests/exams,
Assessment.

Abstract: This study aims at determining the cyber accessibility of teacher-made exams by analyzing the teachers', students', and parents' views on this subject. To fulfill this purpose, 60 exam papers in 4 different courses, including Turkish, Mathematics, Science, Social Studies/Atatürk's Principles and Revolutions, were examined through the technique of document analysis. Also, nine teachers, nine students, and nine parents were interviewed through semi-structured interview forms. According to the results, the level of cyber accessibility of secondary school teacher-made exams were; 87% for Turkish, 51% for mathematics, 99% for Science, and 88% for Social Studies/Atatürk's Principles and Revolutions courses. The highest level of cyber accessibility was found in multiple-choice and open-ended questions for Turkish courses, true-false questions for mathematics courses, open-ended, true-false, matching, and gap-filling questions for the Science courses and open-ended questions for Social Studies/Atatürk's Principles and Revolutions courses. Students and parents stated that they make use of cyber accessible websites while preparing for teacher-made exams. On the other hand, teachers indicated that they perceived themselves incompetent in assessment and evaluation.

1. INTRODUCTION

Education is a broad system where the input, process, output, and feedback mechanisms interact with each other. In the educational system, one of the most effective ways to provide feedback to students, parents, and school; in short, to all the stakeholders of educational process is the exams (Baykul, 2010). The knowledge and skills of individuals about a particular area are tested through exams (Büyüköztürk, 2016).

In this respect, according to the expertise of the author, there are two types of exams: teacher-made exams and standard exams. Teacher-made exams consist of tests used by teachers at school in the classroom. On the other hand, standard exams are exams prepared by experts or a measurement institution by following test development processes (Kilmen, 2014).

The results of teacher-made exams are used to provide feedback on missing or incorrect learning and the creation of students' End-of-the-Year Achievement Score (EYAS) and School

*CONTACT: Atilla ÖZDEMİR ✉ atillaozdemir@sdu.edu.tr 📧 Süleyman Demirel University, Faculty of Education, Department of Mathematics and Science Education, Isparta, Turkey

Achievement Score (SAS). Besides, according to The Guidelines for Transition to Secondary Education (MEB, 2019, p. 7-8) prepared by the Ministry of National Education (MEB) for the students who take the ‘Central Examination for Secondary Education Institutions that will Receive Students by Exam’, which is an exam applied for the transition from secondary school to high school;

“1.5.1. Central Placement

b) In case the central examination score is equal in the schools that accept students by examination; placement is made by looking at the School Achievement Score (SAS) superiority, the End-of-the-Year Achievement Score (EYAS) superiority (...) in the 8th, 7th and 6th grades, respectively.

1.5.2. Local Placement

a) Local placement (...) is done according to the students' residence addresses and the superiority of school achievement score (...) criteria, respectively. In case of equality in the evaluation, the placement will be made by looking at the success score of the 8th, 7th and 6th grades respectively.”

Accordingly, teacher-made exam results also play an effective role in students' transition to higher education. In the German education system, there is no central examination for the transition from primary to secondary education, but placement is made based on the German and mathematics grade averages and observations conducted by the teacher (Aytaç, 1999; Faust, 2011; as cited in Göçkan, 2019, s.64). In the Singapore education system, students' placement into eligible secondary education programs is fulfilled by the ‘Primary School Leaving Examination (PSLE)’. This test includes English, mother tongue, mathematics, and science questions. In the Finnish education system, academic achievements; and for vocational schools, work experiences are taken into consideration for transferring students to secondary education (Bal & Başar, 2014). It is possible to see that the results of teacher-made exams are determinant in the examples of transitions between educational stages in different countries as well. For this reason, teacher-made exams, which play an essential role in making decisions about the educational career of students, must be valid, reliable, and fair.

Teacher-made exams include different types of questions such as multiple-choice, open-ended, true-false, matching, and gap-filling. Teachers are expected to prepare these questions based on the objectives and achievements stated in the curriculum. However, it is assumed that the cyber accessibility of the questions in teacher-made exams will affect the reliability of the exam results; and thus, may lead to assessment and evaluation problems. Although the term cyber can be described as “something that exists in the form of software, not physical” as a computer term, it is used synonymously with the term virtual in daily life and is generally defined as an adjective used to mean belonging to the internet (Özdemir & Şahin; 2005). The increase in the use of information communication technologies has also increased access to information and communication between people. Computer use and internet access rates have increased due to these developments (Shariff & Gouin, 2005; Wang et al., 2009). Thus, the earth transformed into a kind of noosphere, a planet-wide consciousness was arisen, woven from one side of the globe to the other with fibers and networks. This change caused by the development of communication and internet technologies in our lives has carried a large part of our real-life to the cyber world and educational activities have been significantly affected by this change (Ateş, 2016). There are many educational platforms that share content for students and teachers in cyberspace. On these platforms; annual and daily course plan examples, written exam questions, tests, lecture videos, and many other similar contents can be accessed. In many cyber-accessible platforms with these contents, it is seen that teacher-made exam questions are shared for different grade levels and for different courses, including information such as the name of the course, period of education, name or title of the exam. In this study, the concept of cyber accessibility was defined by the researchers as accessing teacher-made exam questions through digital media such as the internet. If teacher-made exams are not prepared in a unique way and consist of questions that exist on any website before, this will cause advantages and

disadvantages for students who could and could not access these questions beforehand. As a result, the grades of the students who have been able to reach the exam questions will not reflect the truth. This situation will negatively affect the reliability of the exam results as well as the student achievement rankings in central exams where school achievement scores become important. Therefore, it will also pose a threat to the fairness of the decisions to be made about the transition of individuals to higher education. Fairness is a fundamental issue of validity and requires attention in all stages of test development and use (AERA et al., 2014).

When the relevant literature is reviewed, it is generally seen that the researchers have examined teacher-made exams taking Bloom's taxonomy into consideration (Ardahanlı, 2018; Çalışkan & Uymaz, 2019; Turan, 2017), and evaluated them regarding the target behaviors (Demircioğlu & Demircioğlu, 2009; Zorbaz, 2005). It is also possible to see other studies that compare teacher-made exams with Central Exam questions, PISA questions, and Central Common Exam scores (Akar, 2019; Bakırcı, 2019; Önder, 2016; Sınacı, 2019). However, no research that determines the cyber accessibility of teacher-made exams and examines the opinions of teachers, students, and parents has been observed in the reviewed literature.

As a consequence, it is necessary to define the cyber accessibility of teacher-made exams that directly affect students' lives and indirectly their social layers. In a similar vein, this research also aims to obtain the teachers', students', and parents' views regarding cyber accessibility of teacher-made exams. This study is significant since it will make a unique contribution to the literature determining the existing situation of the cyber accessibility of teacher-made exams. Additionally, different perspectives will be presented by taking the opinions of teachers, students, and parents on this subject. It is expected that the research results will offer significant benefits to the stakeholders of education and further studies as a source. Therefore, in this study, it is aimed at determining the cyber accessibility of teacher-made exams and examining the teachers', students', and parents' opinions on the subject. For this purpose, answers to the following questions have been sought;

1. What is the cyber accessibility level of the secondary School Turkish, mathematics, science, and social studies/Atatürk's principles and revolutions courses' teacher-made exams?
2. Regarding the cyber accessibility of secondary school teacher-made exams;
 - a) what are the teachers' opinions?
 - b) what are the students' opinions?
 - c) what are the parents' opinions?

2. METHOD

In this section, information about research design, data source, study group, data collection tools and techniques, data collection, and data analysis have been presented. This study adopts phenomenology design, one of the qualitative research methods. Qualitative research, according to Yıldırım & Şimşek (2013, p.46) is defined as "the research in which qualitative data collection methods such as observation, interview, and document analysis are used, and a qualitative process for the realization of perceptions and events realistically and holistically in the natural environment". The research is a two-stage qualitative study consisting of document analysis and descriptive analysis stages for interviews. In the first stage, the document review of teacher-made exam questions belonging to the Turkish, mathematics, science, and social studies/Atatürk's principles and revolutions courses have been conducted. In the second stage, the opinions of teachers, students, and parents on cyber accessibility of teacher-made exam questions were examined in-depth via descriptive analysis. The research models employed within the study are presented in Table 1.

Table 1. *Stages of the research model.*

Stages	Research Model
Stage I	Document Review
	Investigation of cyber accessibility of teacher-made exams for Turkish, mathematics, science and social studies/Atatürk's principles and revolutions courses
Stage II	Descriptive Analysis
	Examining the opinions of teachers, students and parents regarding the cyber accessibility of teacher-made exam questions with descriptive analysis

2.1. Source of Data

The teacher-made exams of Turkish, mathematics, science, and social studies/Atatürk's principles and revolutions courses make up the first stage of the study. They were gathered from three different secondary schools in Ankara that have volunteered to participate in the study.

2.1.1. Study group

In the second stage, the study group of the research has been determined. Criterion sampling, one of the purposive sampling methods, has been used to determine the study group. In criterion sampling, a study group is formed regarding the person, event, object, or situation with the appropriate qualifications for the problem statement (Patton, 2005). Accordingly, a total of nine people in each group, thrice teachers, parents, and students from each of the three volunteer schools were determined as the study group. Furthermore, in order to ensure the diversity of the study group, it is essential to take into account the differences in genders, branches, educational status, and professional seniority of the teachers that are interviewed. As for parents, however, the differences in gender and education status have been considered. And lastly, for students, the gender and class level differences have been taken into consideration.

Within the framework of the research's ethics, instead of using the names of teachers, parents, and students who participated in the research, participants were coded. Teachers were coded as T1, T2, T3,... T9, parents were coded as P1, P2, P3,... P9 and students were coded as S1, S2, S3,... S9. The demographic characteristics of teachers, parents, and students in the study group are presented in Table 2, Table 3, and Table 4, respectively.

Table 2. *Demographic features of teachers.*

Teachers	Gender	Major	Professional Experience (Years)	Educational Status
T1	Female	Turkish	17	BA
T2	Male	Science	25	BA
T3	Female	Mathematics	13	BA
T4	Female	Mathematics	7	BA
T5	Male	Social Studies	5	MA
T6	Female	Science	9	PhD
T7	Male	Turkish	14	BA
T8	Female	Science	15	BA
T9	Female	Social Studies	21	BA

As has been shown in Table 2, six of the teachers participating in the research are female, and three are male. According to the branches, two Turkish, two mathematics, three science, and two social studies teachers have participated in the study. The professional experience level of three teachers is between 5-10 years, the professional experience level of four is between 11-20 years, and the professional experience level of two is between 21-25 years. According to

their educational status, it has been observed that seven teachers have BA degrees, one has an MA degree, and one has a PhD degree.

Table 3. *Demographic features of students.*

Students	Gender	Grade Level
S1	Male	8 th
S2	Male	7 th
S3	Female	6 th
S4	Female	7 th
S5	Male	8 th
S6	Female	6 th
S7	Male	6 th
S8	Female	7 th
S9	Female	8 th

As has been presented in Table 3, four male and five female students have participated in the study. Three of the students are at the sixth grade, three of them are at the seventh grade, and the other three are at the eighth grade.

Table 4. *Demographic features of parents.*

Parents	Gender	Educational Status
P1	Female	Junior College
P2	Male	Undergraduate Degree
P3	Female	High School
P4	Male	Primary School
P5	Female	High School
P6	Female	Primary School
P7	Male	Junior College
P8	Female	Undergraduate Degree
P9	Male	Undergraduate Degree

Table 4 illustrates that five of the parents who have participated in the study are female, and the other four are male. Two of the parents have graduated from primary school, another two have graduated from high school, another two have graduated from junior college, and the other three parents have graduated from faculties.

2.2. Data Collection Tools and Techniques

In this research, the technique of document analysis has been employed to evaluate the cyber accessibility of teacher-made exams collected from schools, and semi-structured interview forms prepared separately for teachers, students, and parents have been utilized to determine the opinions of teachers, parents, and students about cyber accessibility of teacher-made exams.

2.2.1. Credibility (reliability) and transferability (validity) of the study

During the document analysis stage of the research, all exam papers have been classified and numbered according to the branches to ensure reliability and validity. First of all, each paper has been examined independently by the researchers, and the question types specific to the branches have been determined. Then, the differences have been eliminated by comparing the tables obtained by the researchers. In the next process, each question has been examined separately by two researchers on search engines, and the cyber accessibility status of questions

has been recorded. After this stage has been completed, comparisons have been made based on the exam paper, including the sources from which each question has been reached. Thus, a consensus has been reached by proving the evidence for the cyber accessibility of each question by both researchers. At the last stage of the document analysis, one of the researchers has processed the findings obtained according to the branch and question types into the database, and the other researcher has cross-checked and verified the data entry. Thus, internal and external validity and reliability of the first stage of data analysis is assumed to have been ensured.

It should also be noted that interviews have been conducted to support the findings of the document analysis. The findings obtained from the interviews have been analyzed with the technique of content analysis. In the analysis process, the answers to the questions, the repeated concepts, and expressions in each participant form have been compared. In this way, it is thought that a more reliable and valid coding could be made. First of all, the papers belonging to teachers, students, and parents have been named in a representative manner (For example, T1, S1, P1). Following this, each paper has been examined and coded sometimes on the basis of sentences, and sometimes on the basis of paragraphs. As a result, the main ideas have been formed according to themes. Therefore, the codes with repetitive or similar expressions have been simplified and finalized (Bogdan & Biklen, 2007; Gökçe, 2006). The final decision has been made by determining the similarities and differences in the codes that have been constructed separately by the two researchers. During this process, in order to prevent misunderstandings and misinterpretations, opinions of the third researcher have been obtained about the statements that were ambiguous. The same process has been repeated in all three groups of participants, and frequencies of the codes have been determined after it has been decided that the codes created according to these final code lists completely meet the expressions in answers. The codes that are concluded to be representative of the data and included in the final code list have been classified under certain categories and the analysis process finished (Creswell, 2005; Guba, 1981). The reliability of the analysis has been ensured in two ways. Firstly, the data obtained from the interviews have been coded separately by the first and the second researchers, and then the cases considered to be inconsistent have been re-evaluated by making comparisons together. Later, two researchers have come together again and the codes have been compared for a second time, and this time, support has been taken from the third researcher regarding the statements intended to be expressed in some answers. Consequently, the conflicts have been resolved and the codes have been finalized. Thus, in data analysis, the average reliability between coders has been computed as 87% in the first meetin, and 100% after the second meeting. These values have been calculated using the formula $(((\text{Consensus}) / (\text{Agreement} + \text{Disagreement})) \times 100)$ (Miles & Huberman, 2015). As the second method related to reliability, the observer triangulation method has been used and after comparing two separate analyzes, the resulting categories and codes have been examined by an expert who has not participated in the study to verify the process (Denzin, 1970).

2.2.2. Document analysis

In order to determine the cyber accessibility of teacher-made exams for the secondary school Turkish, mathematics, science, social studies/Atatürk's principles and revolutions courses, the data have been collected via the method of document analysis. According to Yıldırım & Şimşek (2013, p. 217), document review involves the analysis of written materials about the phenomenon or cases to be investigated. In addition to being used as a stand-alone data collection method in qualitative research, document analysis can also be used with different data collection methods. In this research, interview forms have been used together with document review.

2.2.3. Teacher, parent and student interview forms

To determine the opinions of teachers, students, and parents regarding the cyber accessibility of secondary school teacher-made exams, a teacher interview form consisting of seven open-ended questions; a student interview form consisting of four open-ended questions; and a parent interview form consisting of three open-ended questions have been developed. In the preparation stage of open-ended questions, questions that would enable participants to provide detailed information have been preferred. After reviewing the relevant literature, draft versions of semi-structured interview forms have been developed with the contributions of researchers. As a next step, expert opinions (experts in qualitative research and assessment and evaluation fields) have been collected. Interview forms have been revised in line with the recommendations of the experts. In order to test the intelligibility and relevance of the questions in the revised semi-structured interview forms, the relevant participants (one Turkish, one math teacher, two parents whose children are currently studying at secondary school, and two secondary school students) have been interviewed for each form. As a result of the pilot testing of the interviews, some questions have had to be rearranged and the final versions of interview forms have been obtained.

2.3. Collection of Data

Courses that constitute the first stage of the research are Turkish, mathematics, science, social studies/Atatürk's principles and revolutions. Teacher-made exams have been collected from three different secondary schools in Ankara that volunteered to participate in the study. While selecting the schools to be included in the study, voluntariness of the schools has been regarded as essential. However, the researchers have also paid attention to diversity of the schools in terms of academic success, socioeconomic level and facilities. Since the schools only archived the previous year's exam questions, 15 exam papers have been randomly selected from the exams conducted in the 2018-2019 academic year and the first semester of 2019-2020 academic year on the basis of branches. Accordingly, a total of 60 different exam papers have been included in the document review.

In the second stage, teachers', students', and parents' opinions and experiences regarding the cyber accessibility of secondary school teacher-made exams have been obtained using a semi-structured interview method. Data have been obtained through face-to-face interviews with nine participants (teachers, students, parents) in each study group by the researcher using the teacher, student, and parent interview forms. Before starting semi-structured interviews, the purpose of the research has been explained to the participants (teacher, student, parent). The participants have been informed that their credentials will remain confidential. For this reason, while expressing the participants' statements, codes have been preferred instead of their real names. Each speech has been transcribed verbatim, in researcher-participant order, without any correction by the researcher.

Firstly, three different secondary school principals in Ankara province have been contacted in order to arrange the meetings with the participants. To be more precise, a separate interview schedule has been created for each participant. Then, on the arranged days and hours, the same schools have been visited by the same researcher, and interviews have been conducted with the participants who volunteered to participate in this research. Interviews have been conducted in the first semester of the 2019-2020 academic year in January. The interviews have been conducted in the deputy principal room of the relevant schools and only the researcher and the participant have been present in the room. The fact that the researcher conducting the interviews is also a teacher made it easier to carry out the interviews with all the participants.

2.4. Data Analysis

2.4.1. Document analysis

In the first stage of the research, teacher-made exams of Turkish, mathematics, science, social studies/Atatürk's principles and revolutions courses have been examined via the technique of document analysis. At this stage, firstly, the total number of questions of the exam papers collected for each course has been grouped. As a next step, each exam paper has been examined on a question basis. The cyber accessibility of the questions has been checked using the google search engine. Thus, all questions for each course have been scanned in the search engine and links to the questions have been obtained. The number of questions examined in terms of branches is 437 in Turkish, 348 in Mathematics, 393 in science, and 454 in social studies/Atatürk's principles and revolutions.

2.4.2. Interview data analysis

In the analysis of the data obtained in the second stage of the research, descriptive analysis has been used to reveal the opinions, experiences of teachers, students and parents related to the cyber accessibility of secondary school teacher-made exams. The raw data obtained from the interviews have been coded and the categories have been identified. The data have been classified under these categories with the aim of making them meaningful for the reader. The coding and categorization process has been repeated by one of the researchers. Therefore, unnecessary codes have been removed by adhering to the problem and purpose of the research, and new codes have been added in the sections deemed as necessary. The researchers have tagged categories in cooperation. Finally, the tables where the frequency values of the participants' views on the subject have been obtained and examples from the participants' opinions have been included.

3. RESULT / FINDINGS

The collected data at this stage have been analyzed for the study. Document analysis has been conducted to determine the cyber accessibility of teacher-made exams and content analysis has been employed to determine the opinions of teachers, students, and parents on this subject. The first question and the findings obtained in the scope of the research are as follows:

3.1. What is the cyber accessibility level of the secondary school teacher-made exams?

In order to answer the first question of the study; each course has been examined separately. The findings obtained from the courses are as shown in Table 5. When the Turkish course's cyber accessibility levels presented in Table 5 are analyzed, it is seen that 381 of the 437 questions are cyber accessible. When this ratio is considered based on the question type, it is concluded that multiple-choice and open-ended questions are cyber accessible. However, when 15 different exam papers are evaluated separately, it is seen that all of the questions in 8 exam papers have been taken from different cyber accessible websites and all of the questions in 1 exam paper have been taken from a single website only by changing the name of the school.

When the findings of the mathematics course are examined, 348 questions have been examined and 51% of these questions are cyber accessible. It has been found that the highest rate is true and false, with 78%. Besides, 15 different exam papers have been examined and it is observed that all of the questions in 4 of them are taken from different cyber accessible websites. The fact that the questions consist mainly of shapes, symbols, and mathematical expressions in mathematics make it difficult for their cyber accessibility; therefore, in each exam paper, firstly verbal phrases have been searched by the researchers. As a result, no exam has been found in which all questions are taken from a single cyber accessible source for the mathematics course. After this stage, the document analysis of the science course questions has started.

Table 5. Courses cyber accessibility level.

Types of Questions /Courses		Multiple Choice	Open-End Questions	True-False Questions	Matching Questions	Gap-Filling Questions	Total
Turkish	NEQ*	248	40	5	102	42	437
	NCAQ** (%)	232 (94)	27 (94)	5 (100)	91 (89)	26 (62)	381 (87)
Mathematics	NEQ	179	61	37	34	37	348
	NCAQ (%)	104 (58)	12 (20)	29 (78)	10 (29)	22 (59)	177 (51)
Science	NEQ	186	12	76	96	23	393
	NCAQ (%)	180 (97)	12 (100)	76 (100)	96 (100)	23 (100)	387 (99)
Social Studies / Atatürk's	NEQ	220	11	25	171	27	454
Principles and Revolutions	NCAQ (%)	210 (96)	11 (100)	15 (60)	141 (83)	22 (82)	399 (87)

*NEQ: Number of Examined Questions; **NCAQ: Number of Cyber Achievable Questions

A total of 393 questions related to science course have been examined and the cyber accessibility rate is found to be 99%. When the cyber accessibility rates of all question types are examined in Table 5, it is seen that all of them are very high. However, it has been observed that 5 of 15 different exam papers are taken from various cyber accessible websites and 9 of them are taken from a single website only by changing the name of the school. As a result, all of the questions in science course have been taken from cyber accessible sources. After this stage, Social studies/Atatürk's principles and revolutions course questions' document analysis has started.

When the cyber accessibility level of Social studies/Atatürk's principles and revolutions course is examined, it is seen in Table 5 that 399 of 454 questions are cyber accessible. When this ratio is examined on the basis of question type, it is concluded that most open-ended questions are cyber accessible. However, when 15 different exam papers have been analyzed separately, it is observed that all of the questions in 7 exam papers are taken from different cyber accessible websites and all of the questions in 5 exam papers are taken from one website only by changing the name of the school.

In general, 1632 different questions from a total of 4 courses have been examined. As a result, the cyber accessibility ratio of these questions is computed as 82%. The question types are grouped under 5 different categories. In these categories, 833 multiple-choice, 124 open-ended, 143 true-false, 403 matching and 129 gap-filling questions have been examined. According to the question types, cyber accessibility rates are calculated as 87% for multiple-choice; 50% for open-ended; 87% for true-false; 84% for matching; and 72% for gap-filling. In other words, cyber accessibility rates are highest in multiple-choice and true-false questions and lowest in open-ended questions. Although the addresses where the questions are taken differ according to each branch, when the frequencies are examined, it has been found that certain addresses are common in all branches.

3.2. Findings from Interviews with Teachers, Students, and Parents

In the tables, the results obtained from the participants are presented for each question posed.

3.2.1. Opinions of teachers

3.2.1.1. Question 1. “Have you received any training on question preparation techniques in measurement and evaluation? Was this training enough for you? Why?” The frequency values of the teachers’ answers to Question 1 are presented in Table 6.

Table 6. Teachers’ opinions on question 1.

Categories	Codes	f	%
Education status	I took it as an undergraduate course	7	26
	I attended the training of a private institution	2	7
Qualifications	Sufficient	0	0
	Insufficient	9	33
Causes	Education remains at a theoretical level and not practiced	8	30
	Change of educational system	1	4
Total		27	100

According to Table 6, seven of the teachers who have participated in the research state that they have received training on question preparation techniques within the scope of measurement and evaluation course at the undergraduate level. All teachers think that the training they have received is insufficient. According to eight of the teachers, education remains at a theoretical level, and practice is not included. Some teachers expressed their opinions as follows:

The course I took at the university remained very academic; the practice was not included (T3). The training that I attended in a private institution was a general review regarding the concepts of measurement and evaluation since the time was limited; no question writing was performed (T7).

3.2.1.2. Question 2. “Do you experience any difficulties in preparing the exam questions in the school? Can you give an example?” The frequency values of the teachers’ answers to Question 2 are presented in Table 7.

Table 7. Teachers’ opinions on question 2.

Categories	Codes	f	%
Difficulty encountered	There is	7	44
	There isn’t	2	13
Example difficulty	Inability to get down to students’ level	5	31
	Not considering him-/herself sufficient in preparing a new generation question	1	6
	Preparing questions for every course outcome	1	6
Total		16	100

According to Table 7, seven of the teachers who have participated in the research state that they encounter difficulty preparing the exam questions. Five of the teachers refer to this difficulty as being unable to get down to students’ level. Some teachers expressed their opinions as follows:

Students have individual differences, and I find it challenging to get down to their level (T6). It is difficult to get down to the level of children. To me, it is easy to prepare difficult questions; it is challenging to prepare easy questions. I cannot predict what children can do because I am new to the school (T4).

The following words of a teacher on the subject are remarkable:

Students are not ready for the new generation questions. I do not consider myself sufficient in preparing new generation questions. We should get training on this subject (T1).

3.2.1.3. Question 3. “What resources do you use while preparing the exam questions for your course? Which of these sources do you use most?” The frequency values of the teachers’ answers to Question 3 are presented in Table 8.

Table 8. Teachers’ opinions on question 3.

Categories	Codes	f	%
Utilized resources	Educational websites	9	29
	Supplementary resources (Question bank, resource book etc.)	5	16
	Education Information Network (EIN)	3	10
	Textbook	3	10
	Notes printed on students’ notebook	2	6
Most used source	Educational websites	8	26
	Notes printed on students’ notebook	1	3
Total		31	100

According to Table 8, it is seen that teachers use multiple sources while preparing their exam questions for their courses. All of the teachers who have participated in the research state that they make use of educational websites while preparing the exam questions. On the other hand, eight teachers note that the most commonly utilized resource is educational websites when preparing exam questions. The following statements by two teachers about the subject are remarkable:

While preparing the exam questions, I benefit mostly from different educational websites because there is an example of a printed and prepared exam. The figure is ready, and I just change the questions (T3).

Since it is more up to date, I examine and organize the exam questions I have compiled from different educational websites in terms of subject, attainment, and class level (T5).

3.2.2. Opinions of Students

3.2.2.1. Question 1. “Have you ever encountered identical exam questions prepared by your teachers at school in a different place? Where have you seen these questions before?” The frequency values of the students’ answers to Question 1 are presented in Table 9.

Table 9. Students’ opinions on question 1.

Categories	Codes	f	%
Encounter	Yes	9	34
	No	0	0
Location encountered	Educational websites	8	30
	Textbook	3	12
	Supplementary resource (Question bank)	3	12
	Exercise book	2	8
	Educational videos	1	4
Total		26	100

According to Table 9, all of the students’ state that they have encountered identical exam questions prepared by their teachers at school in a different place before. Eight of the students who have participated in the study state that they have encountered identical exam questions on educational websites although the sources are different. Some students have expressed their views as follows:

In the Turkish exam we took last week, there were questions I saw on exam websites on the internet. One of them was the question in the textbook (S3).

In our science exam, there were questions that our teacher solved in the notebook and the questions in the question bank that I solved while preparing for the exam (S7).

3.2.2.2. Question 2. "Which sources do you utilize to prepare for the exams at school?" The frequency values of the students' answers to Question 2 are presented in Table 10.

Table 10. Students' opinions on question 2.

Category	Codes	f	%
Utilized resources	Educational websites	8	28
	Textbook	7	24
	Exercise book	5	17
	Education videos	5	17
	Supplementary resource (Question bank)	4	14
Total		29	100

According to Table 10, it is seen that students use more than one source while preparing for exams and they mostly benefit from educational websites. The following statements by three students on the subject are remarkable:

While studying for the exam, I usually solve the questions on the internet and read the textbook (S2).

Our teacher watches video in the course and asks the questions at the end of the video. I'm watching those videos too. Also, since I am preparing for LGS (National High School Entrance Exam), I solve tests from the test book (S5).

I usually read the notes our teachers have us write in our notebooks. Most of the time, I solve the questions on educational websites on the internet (S9).

3.2.3. Opinions of Parents

3.2.3.1. Question 1. "What resources does/do your child/children going to secondary school currently use when preparing for the exams held by teachers at school? Which of these sources do you think he/she/they uses/use the most? Frequency values regarding the answers given by parents to Question 1 are presented in Table 11.

Table 11. Parents' opinions on question 1.

Categories	Codes	f	%
Utilized resources	Educational websites	5	20
	Supplementary resource (Question bank)	4	16
	Textbook	3	12
	Educational videos	2	8
	Exercise book	1	4
	Republic of Turkey Ministry of National Education achievement tests	1	4
Most used source	Educational websites	3	12
	Supplementary resource (Question bank)	3	12
	Textbook	1	4
	Educational videos	1	4
	Exercise book	1	4
Total		25	100

According to Table 11, parents whose children attend secondary schools state that their children use more than one source while they are preparing for the exams held by teachers at school. Five of the parents who have participated in the study state that they use educational websites while preparing them for the exams at school. Three parents note that the most commonly used resource their children use to prepare for the exams held in school is the educational websites. Moreover, three parents refer to supplementary resources as the most commonly utilized sources. The following view of a parent on the subject is remarkable:

While my eldest son was going to middle school, we were downloading questions from the websites that had written questions on the internet to prepare them for exams. He solved these questions because the same questions appeared in the test and he was getting high marks. But this was not his real score. He scored low in TEOG (The Evaluation of the Transition Model from Elementary to Higher Education) and accepted by a high school he did not want. Therefore, I am not making the same mistake with my younger son. We do not solve any questions from that website. We use the MEB (Ministry of National Education) acquisition tests and continue from the sourcebooks (P3).

Another parent has expressed his/her opinion as follows:

When my daughter has an upcoming exam, she reads the topics from her textbook and from her supplementary sources in her room. Then, she solves exam questions from the websites she finds on the Internet and solves the questions and checks her mistakes from the supplementary resources she made us bought for her. As far as I can see, she spends most of her time on the computer and solves questions from internet sites (P6).

4. DISCUSSION and CONCLUSION

This study aims at obtaining the cyber accessibility levels of the secondary school teacher-made exams and the opinions of teachers, students, and parents on this subject. According to the results, the level of cyber accessibility of secondary school teacher-made exams for each course is as follows: 87% for Turkish, 51% for mathematics, 99% for science and 88% for social studies/ Atatürk's principles and revolutions. When the question types are examined, the highest cyber accessibility level is observed in multiple-choice and true-false questions. It is clear that the most common question type in the question papers is multiple choice and these results are consistent with those of Kılıç (2016) and Uymaz (2016).

Another remarkable finding at the end of the study is that 82% of all the questions are cyber accessible. This result is important in terms of questioning the reliability of teacher-made exams. In a similar vein, Sinacı examined teacher-made exams before and after TEOG in his study in 2019 and found that in some schools, students' scores in pre-TEOG teacher-made exams and in TEOG are inconsistent. Furthermore, post-TEOG teacher-made exams show higher correlations than pre-TEOG teacher-made exams with TEOG. It was stated that the placement scores obtained from the TEOG system were not calculated fairly because the results of the teacher-made exams were tried to be equated to TEOG results. Similarly, the study by Antenesh and Silesh (2018) found a positive but low correlation between teacher-made exams and regional exam results ($r=0.476$, $p<0.01$). The study conducted by Özdemir & Gelbal (2016) with the aim of investigating the predictive power of the same students' end-of-the-year achievement score from 7th to 12th grade on the raw scores of YGS sub-exams concludes that the best predictor of mathematics, Turkish, science, and social sciences sub-tests is the end-of-the-year achievement score of different courses at different grade levels and these variables explain 50% to 71% of success in the relevant sub-tests. These results indicate that teacher-made exams are not sufficient in terms of measurement and evaluation criteria.

When the teacher's opinions are examined, a vast majority of teachers state that they have problems in the measurement and evaluation process. Similarly, according to the research

results of Zorbaz (2005), most of the Turkish language teachers considered determining whether the subjects are comprehended or not and grading with the purpose of measurement and evaluation. As a matter of fact, it has been identified that the teachers who have participated in the study generally received their training on question preparation techniques within the scope of the ‘measurement and evaluation’ course given during their undergraduate years. Furthermore, it has been understood that this education remained at the theoretical level and the education received is insufficient because it did not include any practice dimension. The teachers state that they mostly encounter difficulties while preparing the exam questions due to their inability to reach the student level. Teachers also complain that they cannot write questions covering all subjects and do not perceive themselves competent in preparing new generation questions, leading to difficulties in preparing exam questions. Similarly, it was found that teacher-made exam questions examined by Uymaz (2016) and Zorbaz (2005) were mostly at the level of knowledge and comprehension and they did not include many questions at the metacognitive level.

In a similar fashion, teachers have remarked that they mostly benefit from educational websites while preparing their exam questions. In the document analysis process of the research, although the exam questions have been scanned separately as root and options, it has been observed that both the root and the options of the questions have been taken from cyber accessible educational sites without any modifications. The most important reason for this may be that teachers do not have sufficient knowledge about measurement and evaluation techniques. In addition, it has also been observed that there are several common addresses for each course. To be more precise, while all of the questions on certain exam papers could be accessed from various websites, it has also been noticed that on some exam papers, all questions were taken from a single cyber accessible source. When the student opinions on the cyber accessibility of teacher-made exams are examined, it is monitored that these opinions support the results obtained from the document examination. Accordingly, all of the students who have participated in the study state that they have encountered some of the exam questions prepared by their school teachers on educational websites. For this reason, it has been admitted that educational websites are the most beneficial sources for students while taking advantage of multiple resources during school exams.

As to the parents’ views on the cyber accessibility of teacher-made exams, the parents who have participated in the study believe that while their children prepare for the exams conducted by the teachers at school, they benefit from more than one source; however, they mostly use educational sites and supplementary resources (such as sourcebooks and question banks). Most significantly, it has been observed that both parents and students are aware of the cyber accessibility of teacher-made exams. Therefore, they tend to guide their children to the questions on the cyber accessible websites to ensure high scores in teacher-made exams at the cost of low scores in the central exam. This result supports the findings of studies conducted by Antenesh & Silesh (2018) and Sinacı (2019). To sum up, the cyber accessibility of teacher-made exam questions prepared for four basic courses at the secondary school level; namely, Turkish, mathematics, science and social studies/Atatürk’s principles and revolutions, is high and since this is known by both the students and the parents, it would be justified to argue that the reliability level of these exams is low. This study is limited to teacher-made exams collected for 4 basic courses and the participants in three secondary schools in Ankara. This situation requires the repetition of the study in different levels, schools and courses in order to support the results obtained from the research theoretically. In order to increase the quality of teacher-made exams, it is recommended to organize trainings for teachers on question preparation techniques.

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:** Suleyman Demirel University, 14.09.2020- 96/13.

Authorship Contribution Statement

Gulden Ozdemir: Investigation, Resources, Visualization, Software, Formal Analysis. **Atilla Ozdemir:** Investigation, Resources, Visualization, Software, Formal Analysis, and Writing - original draft. **Selahattin Gelbal:** Methodology, Supervision, and Validation.

ORCID

Gulden Ozdemir  <https://orcid.org/0000-0002-3150-9438>

Atilla Ozdemir  <https://orcid.org/0000-0003-4775-4435>

Selahattin Gelbal  <https://orcid.org/0000-0001-5181-7262>

5. REFERENCES

- AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Akar, Z. (2019). *Sekizinci sınıf Türkçe dersi yazılı sorularının merkezî sınav Türkçe soruları ile karşılaştırılması* [Comparison of eighth grade Turkish course written questions with central exam Turkish questions] [Unpublished master's thesis]. Ataturk University.
- Anteneh, M. M., & Silesh, B. D. (2018). Assessment practices and factors for the disparity between students' academic scores at teacher-made and regional exams: The case of Bench Maji zone grade 8 students. *Educational Research and Reviews*, 14(1), 1-24. <https://doi.org/10.5897/ERR2018.3581>
- Ardahanlı, Ö. (2018). *TEOG sınavı matematik soruları ile 8. sınıf matematik yazılı sınav sorularının yenilenmiş bloom taksonomisi'ne göre incelenmesi* [Analysis of questions in the TEOG examination and questions in the mathematics written exam of 8th grade mathematics courses according to the revised bloom's taxonomy] [Unpublished master's thesis]. Eskisehir Osmangazi University.
- Ateş, D. (2016). Possibilities of democracy transformation in cyberworld: the case of legislation assemblies. *Cyberpolitik Journal*, 1(1), 170-177.
- Aytaç, K. (1999). *Federal Almanya Cumhuriyeti'nde okul sistemi* [School system in the Federak Republic of Germany]. Engin Yayınevi.
- Bakırcı, G. (2019). *Ortaokul matematik öğretmenlerinin veri öğrenme alanına dair yazılı sınav soruları ile pisa sorularının karşılaştırmalı incelemesi* [Comparative investigation of the secondary school mathematics teachers written exam questions and PISA questions in the field of data handling] [Unpublished master's thesis]. Gaziantep University.
- Bal, B., & Başar, E. (2014, Ocak). Finlandiya, Almanya, Singapur ve Türkiye'nin eğitim sistemleri açısından kademeler arası geçiş sistemlerinin karşılaştırılması [The comparison the systems of passing among educational levels of Finland, Germany, Singapore and Turkey]. *Çukurova Üniversitesi Türkoloji Makale Bilgi Sistemi*, 18776, 1-24.
- Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması* [Measurement in education and psychology: Classical test theory and practice]. (2. Baskı). Pegem Akademi.
- Bogdan, R.C., & Biklen, S.K. (1992). *Qualitative research for education: A introduction to theory and methods*. Allyn and Bacon.
- Büyükköztürk, Ş. (2016). Sınavlar üzerine düşünceler – Türkiye'de ki ölçme değerlendirme sistemi üzerine güncel durum hakkında uzman değerlendirmesi [Thoughts on exams -

- expert assessment on the current situation on the assessment system in Turkey]. *Kalem Eğitim ve İnsan Bilimleri Dergisi*, 6(2), 345-356.
- Creswell, J. W. (2005). *Educational research: Planning, conducting and evaluating quantitative and qualitative research*. Upper Saddle River, Pearson Education, Inc.
- Demircioğlu, G., & Demircioğlu, H. (2009). Kimya öğretmenlerinin sınavlarda sordukları soruların hedef davranışlar açısından değerlendirilmesi [An evaluation of the questions chemistry teachers asked in exams in terms of the target behaviors]. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, 3(1), 80-98.
- Denzin, N. K. (1978). *The research act: A theoretical orientation to sociological methods*. McGraw-Hill.
- Göçkan, A. (2019). *Türk eğitim sistemi ile Alman eğitim sisteminde kademeler arası geçişlerin karşılaştırılması* [Comparison between the Turkish education system and the levels in the German education system] [Unpublished master's thesis]. Canakkale Onsekiz Mart University.
- Gökçe, O. (2006). *İçerik analizi: kuramsal ve pratik bilgiler* [Content analysis: theoretical and practical information]. Siyasal Kitabevi.
- Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology: A Journal of Theory, Research, and Development*, 29(2), 75- 91.
- Kılıç, A. F. (2015). *Temel eğitimden ortaöğretime geçiş ortak ve mazeret sınavındaki Türkçe ve matematik alt testlerinin psikometrik özelliklerinin karşılaştırılması* [The comparison of psychometric properties of standardised and make up maths and Turkish subtest questions in the exam of transition from basic to secondary education] [Unpublished master's thesis]. Hacettepe University.
- Kilmen, S. (2014). Ölçme ve değerlendirmede temel kavramlar, R. N. Demirtaşlı (Ed.), *Eğitimde ölçme ve değerlendirme içinde* [Basic concepts in measurement and evaluation, R. N. Demirtaşlı (Ed.), in *Measurement and evaluation in education*] (s. 30-64). Edge Akademi.
- MEB (2019). Ortaöğretime geçiş tercih ve yerleştirme kılavuzu [The Guidelines for Transition to Secondary Education]. http://www.meb.gov.tr/meb_iys_dosyalar/2019_06/25104443_evrak8071216911865128306.pdf
- Miles, M. B., & Huberman, A. M. (2015). *Nitel veri analizi* [Qualitative data analysis] (2. baskıdan çeviri) (S. Akbaba Altun, & A. Ersoy Çev. Eds). Pegem Akademi.
- Önder, R. (2016). 2014-2015 TEOG sınavına ilişkin paydaş görüşleri ile öğretmen yapımı testlerle olan ilişkisi [Relationship between basis of stakeholders opinion about TEOG with teachers making test in 2014-2015] [Unpublished master's thesis]. Akdeniz University.
- Özdemir, A., & Gelbal, S. (2016). Predictive power of primary and secondary school success criterion on transition to higher education examination scores. *Journal of Measurement and Evaluation in Education and Psychology*, 7(2), 309-334.
- Özdemir, S., & Şahin, Ş. (2005). Siberuzay nerede? [Where is the cyberspace?]. *Pivolka*, 4(19), 1-14.
- Patton, M. Q. (2005). *Qualitative research*. John Wiley & Sons, Ltd.
- Shariff, S., & Gouin, R. (2005, November). *Cyber-dilemmas: Gendered hierarchies, free expression and cyber-safety in schools*. In Oxford Internet Institute conference at Oxford University.
- Sınacı, B. (2019). *Temel eğitimden ortaöğretime geçiş (teog) sisteminde uygulanan sınavların puanları ile diğer puanların karşılaştırılması* [The comparison of scores in transition from basic education to secondary education (TEOG) and other scores] [Unpublished master's thesis]. Hacettepe University.

- Turan, S. D. (2017). *Ortaokul 5, 6, 7 ve 8. sınıfların bilim dersi yazılı sınav sorularının ölçme ve değerlendirmeye uygunluğu açısından incelenmesi* [An investigation on the written examinations of the science courses of 5, 6, 7 and 8 th grade with regard to the appropriateness of measurement and evaluation] [Unpublished master's thesis]. Mustafa Kemal University.
- Uymaz, M., & Çalışkan, H. (2019). Öğretmen yapımı sosyal bilgiler dersi sınav sorularının yenilenmiş bloom taksonomisine göre incelenmesi [An investigation on the teacher-made social studies course exam questions in terms of revised Bloom's taxonomy]. *Kastamonu Education Journal*, 27(1), 331-346. <https://doi.org/10.24106/kefdergi.2637>
- Wang J, Iannotti RJ, & Nansel TR. (2009). School bullying among adolescents in the United States: physical, verbal, relational, and cyber. *Journal of Adolescent Health*, 45(4), 368-75.
- Yıldırım, A., & Şimşek, H. (2013). *Sosyal bilimlerde nitel araştırma yöntemleri* [Qualitative research methods in the social sciences]. Seçkin Yayınevi.
- Zorbaz, K. Z. (2005). *İlköğretim okulları ikinci kademe türkçe öğretmenlerinin ölçme ve değerlendirmeye ilişkin görüşleri ve yazılı sınavlarda sordukları sorular üzerine bir değerlendirme* [An evaluation on the views of primary school Turkish teachers on measurement and evaluation and the questions they ask in written examinations] [Unpublished master's thesis]. Mustafa Kemal University.

Investigation of Measurement Invariance of State Test Anxiety Scale

Huseyin Selvi ^{1,*}

¹Faculty of Medicine, Department of Medical Education, Mersin University, Mersin, Turkey

ARTICLE HISTORY

Received: Nov. 17, 2020

Revised: May 20, 2021

Accepted: May, 31, 2021

Keywords:

Test Anxiety,
Measurement Invariance,
Multi Group Confirmatory
Factor Analysis,
Test Validity

Abstract: In this study, it was aimed to examine the measurement invariance of State Test Anxiety Scale and its sub-dimensions developed by Şahin (2019) in terms of different variables. For this purpose, data were collected from a total of 956 university students studying in different faculties. The measurement invariance of the scale was examined by multi-group confirmatory factor analysis in terms of gender, faculty and socioeconomic level variables. In the study, the measurement model was established for 22 items and three components of the state anxiety test scale (cognitive, psychosocial and physiological) and tested for configural, metric, scalar and strict equivalence by considering the hierarchical principle in terms of gender, faculty and socioeconomic level variables. The findings showed that Configural equivalence was provided for all dimensions except the cognitive and physiological subscales for the socioeconomic status variable. On the other hand, metric equivalence was achieved in cognitive, psychosocial and physiological dimensions for the gender variable. Metric equivalence was achieved in Cognitive dimension for faculty variable. And for the socioeconomic status variable, it was provided only for the scale as a whole. Scalar and strict equivalence conditions were not met by any of the variables examined in the study.

1. INTRODUCTION

Test anxiety is one of the important variables affecting the academic success of individuals. For this reason, it is one of the subject area that psychology has emphasized since the 1950s. Test anxiety is a special form of anxiety and it can affect individuals of all ages in the society (Sieber, 1980). The exams, which are carried out for different purposes such as selection, placement, diagnosis and guidance, especially the exams with wide participation, affect the lives of individuals significantly today. Considering the meaning attributed to these exams and the potential of these exams to affect the lives of individuals, the dimension of anxiety experienced by individuals and their families can be better understood.

In the literature anxiety is defined as; fear of anticipation of something bad will happen, restlessness and feeling of loss of control (Sapir & Aranson, 1990), as fear and tension felt under threat (Büyüköztürk, 1997), as sadness and distress caused by stressful situations (Özgüven, 2007). As can be understood from the definitions, the concept of "anxiety"; It includes feelings of sadness, distress, fear, failure, helplessness and loss of control (Cüceloğlu,

*CONTACT: Huseyin Selvi ✉ hsyn_selvi@yahoo.com.tr 📧 Faculty of Medicine, Department of Medical Education, Mersin University, Mersin, Turkey

1998). Increasing the level of anxiety reveals the preventive role of anxiety. High test anxiety; it is an important problem that negatively affects the learning process and academic achievement of individuals (Ergene, 1994). As a matter of fact, individuals with high test anxiety may encounter situations like; easily distract, worry about performance, tension, restlessness, sadness, distress, fear, helplessness, loss of control, incompetence, silence, loss of control, less speech, withdrawal, palm sweating, hands trembling, increased heart rate and panic attack (Geen, 1985; Öner, 1990; Zeidner, 2007).

On the other hand, anxiety is generally perceived as a negative emotional state. However, it does not always affect the person negatively. This situation, which is described as the facilitating effect of anxiety, is referred to as "facilitating anxiety" in the literature. Facilitating anxiety; It emerges as a result of the person developing more motivation to cope with this situation and making more effort due to the increase in perception and awareness of the anxiety situation (Albert & Haber, 1960). Studies conducted in the literature on test anxiety show that very low and very high-level test anxiety affects learning negatively and a medium-level test anxiety affects learning positively (Hill & Wingfield, 1984; Bados, 2005; Gençdoğan, 2006). For this reason, keeping test anxiety under control is important for individuals' academic success, self-confidence and motivation.

Another important requirement of keeping test anxiety under control is that the quality of measurement tools used to measure success is affected by this situation. As a matter of fact, the most important requirement of a qualified measurement process is that the tests used in the exams can measure the variable to be measured without mixing it with other variables (Turgut & Baykul, 2012; Alici, 2013). A measurement process that can be performed in a qualified manner independent from the negativity of variables such as test anxiety, test technique and motivation will increase the accuracy of the decisions to be taken (Hill & Wigfield, 1984).

The review of the literature on this issue shows that many studies have been conducted to measure test anxiety and to reveal the reasons by examining it in terms of different variables. Most of these studies consist of scale development / scale adaptation studies for measuring test anxiety and studies that attempt to reveal the causes of anxiety (McDonald, 2001; Driscoll, 2007; Totan & Yavuz, 2009; Akın *et al.*, 2012; Başol, 2017; Aydın & Bulgan, 2017; Bozkurt *et al.*, 2017; Yao-Ting Sung & Tzu-Yang Chao, 2015; Şahin, 2019).

When the scales developed and adapted to measure test anxiety in Turkey are examined, it is seen that the evidence for the reliability and validity of almost all of these scales is collected. Using these scales, many studies have been conducted in the literature to reveal the reasons for test anxiety and to reveal the differences between different subgroups such as gender, education level, socioeconomic status, school type (Totan & Yavuz, 2009; Akın *et al.*, 2012; Başol, 2017; Aydın & Bulgan, 2017; Şahin, 2019). However, it is not correct to explain the differentiation of values obtained by using these scales between groups by only linking the characteristics of individuals. As a matter of fact, the differences between groups can be caused by the measurement tool rather than the individuals. Although there are studies on linguistic validity in scale adaptation studies, the adapted measurement tool may not measure the same structure in the language / culture it was adapted to (Cheug & Rensvold; 2002). The same situation also applies to the implementation to different groups of the scale was developed in Turkey. The developed scale may not measure the same structure for females and males in terms of gender variable and high school and university graduates in terms of education level variable. This is explained by the fact that psychological structures do not exactly overlap in different groups and cultures, in other words, the behaviors related to the structure can be different in different groups and cultures (van de Vijver & Poortinga, 2005).

In order for the scores obtained from the scale to be compared between groups, it must first be shown that the scale measures the same structure in different subgroups, in other words,

measurement invariance is achieved (van de Vijver & Poortinga, 2005). In measurement invariance studies, the items in the measurement tool in different groups; It is examined whether the factor loadings, correlation patterns, error variances are the same. Measurement invariance is a prerequisite for comparison studies between groups (Erkuş & Selvi, 2019; Cheung & Rensvold, 2000). When measurement invariance is not achieved, it means that the measurement tool does not measure the same thing in different subgroups. This causes the comparison studies to lose their meaning. In summary, it is not possible to make a comparison between groups without measurement invariance (van de Vijver & Poortinga, 2005).

Measurement invariance studies are conducted to reveal whether the factor structure obtained for the scale is the same in the sub-groups by using multi-group confirmatory factor analysis (Başusta, 2010). Multi-group confirmatory factor analysis is a method frequently used in structural equating modeling analysis, and it is performed to examine whether the model created by the researcher is the same in more than one group based on the data obtained from the same measurement tool (Tabachnick & Fidel, 2001).

Test anxiety can be compared in terms of variables such as gender, class level, socioeconomic status, age group, school, university, department, education level. But first of all, the measurement tool used must provide measurement invariance. Findings regarding differences between individuals and groups cannot be interpreted without providing measurement invariance (Horn & Mc Ardle, 1992). Measurement invariance studies consist of 4 phased steps, each of which is the prerequisite of the next, such as configural, metric, scalar and strict equivalence (Meredith, 1993; van de Vijver, 1998). Configural equivalence is the most basic level of measurement invariance, and it is the step where it is tested that the factor structure revealed for the measured psychological variable is the same in all groups, in other words, that the free and fixed factor patterns are similar between the groups. As a matter of fact, in order for the groups to be compared, it must be demonstrated that the relevant measurement tool measures the same thing in all groups. If the measured structure is different in the groups to be compared, in other words, if the related measurement tool measures different things in different groups, it is not meaningful to make comparisons between groups. Depending on this condition, it is clear that metric, scalar and strict equivalencies cannot be examined.

After the Configural equivalence is achieved, metric equivalence is examined. Although the measurement tool measures the same structure in different groups, it may not be able to measure individuals in different groups with the same latent structure at the same size. For this reason, the invariance of the units of the measuring tool in metric equivalence between groups examine. In other words, the equality of the units is tested. In order to test the equality of the units, it examines whether the factor loadings obtained for the scale items change or not between groups. When metric equivalence cannot be achieved, a situation arises where scale items are biased and cannot be summed. Therefore, it is not possible to examine scalar and strict invariance in cases where metric invariance cannot be achieved.

For scalar invariance, the equality of the origins of the measuring instrument between groups is tested. In other words, it is tested whether 5 points in one group equal 5 points in the other group. If 5 points in one group equals 7 points in the other group, then the origins are unequal and the measuring tool contains possible bias. For scalar invariance, it is sufficient to demonstrate that group means and factor loadings are equal. In strict equivalence, in addition to configural, metric and scalar invariance, it is tested whether the error and factor variances obtained for the scale items are equal between the groups. If configural, metric, scalar and strict invariance is provided for different groups, it can be interpreted that the scale measures the same structure, the same size and the same precision in these groups. And this makes it possible to compare the scores obtained from these different groups with the same measurement tool.

Regarding test anxiety, the scale used in order to make comparisons between groups should meet the measurement invariance conditions. In this context, the scales developed and adapted to measure test anxiety were examined and summarized in [Table 1](#).

Table 1. Summary of scales developed and adapted in Turkey.

Scale Name	Developed / Adapted	Measurement Invariance
State Test Anxiety Scale	Developed by Şahin (2019)	Unreported
Revised Test Anxiety Scale	Developed by Benson & El-Zahhar (1994), Adapted by Akin <i>et al</i> (2012)	Unreported
IDA Test Anxiety Scale,	Developed by Başol (2017)	Unreported
Children’s Test Anxiety Scale	Developed by Wren & Benson (2004), Adapted by Aydın & Bulgan (2017)	Unreported
Cognitive Test Anxiety Scale-Revised Form	Developed by Cassady & Johnson (2002), Adapted by Bozkurt <i>et al</i> (2017)	Unreported
Friedben Test Anxiety Scale	Developed by Bados & Sanz (2005), Adapted by Akin <i>et al</i> (2013)	Unreported
Westside Test Anxiety Scale	Developed by Driscoll (2007), Adapted by Totan & Yavuz (2009)	Unreported

When [Table 1](#) is examined, no findings related to measurement invariance have been reported in any of the scales. Evidence regarding measurement invariance was not provided in any of the scales developed or adapted in Turkey to measure test anxiety.

Test anxiety is a variable that negatively affects individuals' learning process, academic achievement, and the quality of the measurement tool used, and it can affect individuals of almost all ages in the society. Therefore, studies conducted to reveal the reasons for test anxiety and to keep it under control are important for the literature.

For this reason in this study, it is aimed to provide a scale to the literature that provides the evidence for measurement invariance. In this study the measurement invariance of the State Test Anxiety Scale developed by Şahin (2019) in terms of variables of gender, socioeconomic level and faculty attended was examined.

2. METHOD

In this study, it was aimed to examine the measurement invariance of the state test anxiety scale. Ethics committee approval was obtained for the study.

2.1. Study Sample

Kline (2005) states that a sample of 200 people is generally sufficient in factor analysis studies. In addition, in different sources, it is recommended to reach 5-20 times the number of items for factor analysis (Alpar, 2016).

In order to reach a sample that can represent the range of the measured latent trait, the sample of the study were determined by purposeful sampling method and the data of the research were obtained from 956 university students. 572 of the students are female (59.8%) and 376 of them are male (39.3%). 8 (0.8%) of the students was not specify their gender. 8 (0.8%) of the students did not specify their gender. 400 of the students stated that they were medical school students (41.8%), 112 were dentistry students (11.7%), and 444 (46.4%) were health college students. 188 of the students (19.7%) are 1; 236 of them (24.7%) are 2; 428 of them (44.8%) are 3 and 96 of them (10%) are 4 grade. 8 (0.8%) of the students did not specify their class. 92 (9.6%) of the students stated that their socioeconomic level was low, 812 (84.9%) were medium, and 52 (5.4%) were high.

2.2. Data Collection Tool

The "State Test Anxiety Scale" was developed by Şahin (2019). The scale is scored in Likert type with 4 degrees and consists of 22 items in total. The scale consists of 3 components as 'cognitive', 'psychosocial' and 'physiological' and these components explain 59.21% of the total variance. This structure, which was revealed by the exploratory factor analysis, was also confirmed by the confirmatory factor analysis. Goodness of fit values were $\chi^2 / df = 1.72$, CFI = 0.96, NNFI = 0.96, IFI = 0.96, RMSEA = 0.05, SRMR = 0.05. The lowest score that can be obtained from the scale is "22" and the highest score is 88. The Cronbach alpha reliability of the scale was calculated as 0.94. Alpha reliabilities for sub-dimensions were calculated as 0.85 for physiological sub-dimension, 0.84 for psychosocial sub-dimension and 0.93 for cognitive sub-dimension. Similarly, the test-retest reliability of the scale, applied 4 weeks apart, was calculated as 0.74 for the physiological sub-dimension, 0.80 for the psychosocial sub-dimension, 0.78 for the cognitive sub-dimension, and 0.81 for the overall scale.

2.3. Data Analysis

The measurement invariance of the state test anxiety scale was analyzed in terms of gender, faculty and socioeconomic level variables using the multi-group confirmatory factor analysis method. In the multi-group confirmatory factor analysis, it is examined whether the model created by the researcher regarding the measurement tool is the same in different groups (Tabachnick & Fidell, 2001). For this purpose, equality limitations are imposed on the model established for subgroups in the multi-group confirmatory factor analysis, and the equivalence of intergroup parameters is examined by following a hierarchical order from the least limited to the most limited model (Başusta, 2010).

In the study, in order to examine the configural equivalence, it was tested whether the structure revealed by the state test anxiety scale measures the same formal structure in subgroups. For this purpose, the similarity of the number of factors and factor loadings of the measurements obtained from the subgroups was investigated. In the examination of metric equivalence, it was examined whether the units (distances between categories) of the conditionality test anxiety scale in different groups were equal. In the examination of scalar equivalence, it was examined whether the constants (origins) of linear regression equations between latent and observed variables change between groups. In the examination of strict equivalence, it was examined whether all the parameters estimated for the model presented were equal among the groups (Erkuş & Selvi, 2019).

In this study, ΔCFI values were used to examine the equivalence of parameters between groups and " $0.01 \geq \Delta CFI \geq -0.01$ " criterion was used for ΔCFI values. ΔCFI values provide information about the relationship between implicit scores and observed scores and are therefore recommended for evaluation of goodness of fit (Amery *et al.*, 2007; Brown, 2006; Vandenberg and Lance, 2000; cited in Uzun, 2010).

For model data fit, values of χ^2 / df , Root Mean Square of Approximate Errors (RMSEA), Comparative Fit Index (CFI), Adjusted Goodness of Fit Index (AGFI) and Root of Residual Means (RMR) were taken into consideration (Kline, 2005). As a criterion, the criteria given in Table 2 were taken into consideration (Çokluk *et al.*, 2010; Şimşek, 2007). Lisrel 8.7 program was used to analyze the data.

In the study, Little's MCAR test was used in missing data analysis. From the findings, it was concluded that the missing data were in random structure ($\chi^2 = 1083.123$, $df = 215$, $p = 0.001$). For this reason, expectation maximization method has been used to eliminate missing data. In addition, the data before the analysis were cleared of outliers and the skewness value calculated for the scale total score was calculated as 0.604 and the kurtosis value as -0.07. Şenocak (2014) states that the distribution can be considered normal if the skewness value is less than 1.00 and

the kurtosis value is less than 2.00. From here, it was accepted that the data of the study were normally distributed.

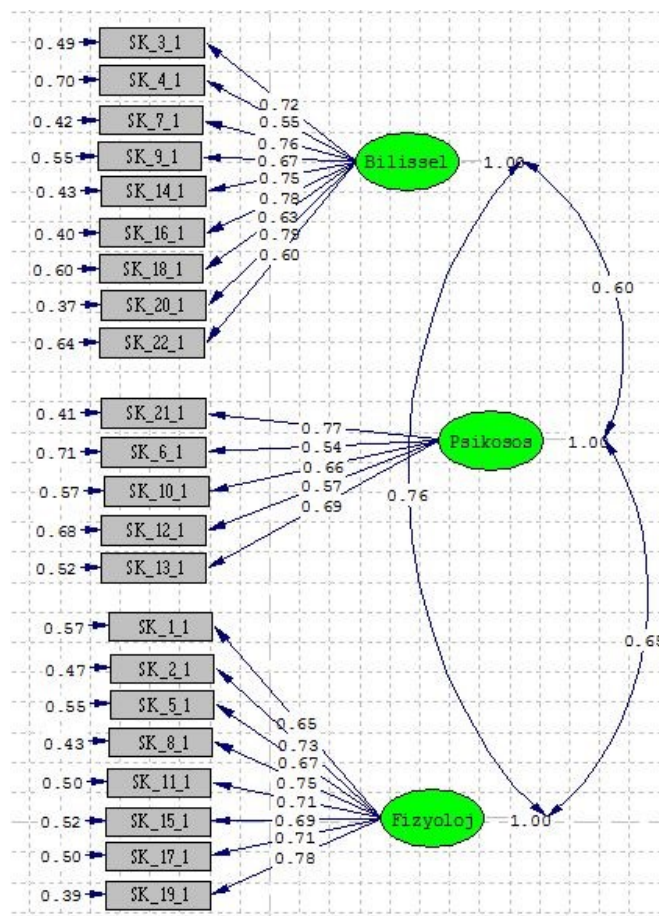
Table 2. Fit indices.

Fit Indices	Good Fit	Acceptable Fit
χ^2/df	$0 \leq \chi^2/df \leq 2$	$2 < \chi^2/df \leq 5$
RMSEA	$0 \leq RMSEA \leq .05$	$.05 < RMSEA \leq .08$
RMR	$0 \leq RMR \leq .05$	$.05 < RMR \leq .08$
SRMR	$0 \leq SRMR \leq .05$	$.05 < SRMR \leq .08$
CFI	$.95 \leq CFI \leq 1$	$.90 \leq CFI < .95$
NFI	$.95 \leq NFI \leq 1$	$.90 \leq NFI < .95$
NNFI	$.95 \leq NNFI \leq 1$	$.90 \leq NNFI < .95$
GFI	$.95 \leq GFI \leq 1$	$.90 \leq GFI < .95$
AGFI	$.90 \leq AGFI \leq 1$	$.85 \leq AGFI < .90$

3. RESULT / FINDINGS

The fit statistics of the state test anxiety scale were found as $\chi^2/df = 8.95$, RMSEA = 0.08, CFI = 0.94, GFI = 0.90, SRMR = 0.06 and NFI = 0.94. When the obtained values are compared with the criteria in Table 2, it is seen that the fit statistics for the model are within the 'acceptable fit' limits except for the χ^2/df value. It is known that the value of χ^2/df is affected by the sample size and therefore it should not be interpreted alone in measurement invariance studies (Cheung & Rensvold, 2002). For this reason, it was accepted that model data fit was achieved for the model created. The path diagram obtained for the measurement model of the state test anxiety scale is presented in Figure 1.

Figure 1. The path diagram of the measurement model of the state test anxiety scale.



In this study, the measurement invariance of the gender, faculty and socioeconomic level variables of the state test anxiety scale was examined by taking the principle of hierarchy (configural, metric, scalar and strict equivalence) into consideration, and the findings obtained in Table 3 for the gender variable, in Table 4 for faculty variable, in Table 5 for socioeconomic level variable and in Table 6 for summary information on whether the invariance is provided for all of the examined groups.

Table 3. Fit statistics obtained from the measurement invariance study on gender variable.

		Equivalence	χ^2/df	CFI	RMR	GFI	RMSEA	ΔCFI	Invariance
Gender	Cognitive	Configural	4.85	0.94	0.06	0.92	0.06	-	+
		Metric	8.52	0.94	0.07	0.91	0.07	0.00	+
		Scalar	10.54	0.90	0.08	0.85	0.14	0.03	-
	Psychoso- cial	Configural	8.16	0.96	0.03	0.99	0.07	-	+
		Metric	6.46	0.95	0.07	0.98	0.08	0.01	+
		Scalar	7.15	0.85	0.09	0.96	0.12	0.10	-
	Physiologi- cal	Configural	16.62	0.91	0.04	0.90	0.07	-	+
		Metric	15.6	0.90	0.07	0.90	0.08	0.01	+
		Scalar	13.8	0.85	0.06	0.85	0.16	0.05	-
Entire Scale	Configural	8.49	0.91	0.08	0.90	0.08	-	+	
	Metric	8.30	0.80	0.15	0.60	0.20	0.10	-	

When Table 4 is examined, it is seen that the configural equivalence is provided for the cognitive, psychosocial, physiological dimensions and the entire scale for the groups related to the faculty variable, and the metric invariance is provided for only the cognitive dimension. In addition, metric invariance could not be provided for the psychosocial, physiological dimensions and entire scale.

Scalar and strict invariance could not be provided for sub-dimensions and the entire scale.

Table 4. Fit statistics obtained from the measurement invariance study for faculty variable.

		Equivalence	χ^2/df	CFI	RMR	GFI	RMSEA	ΔCFI	Invariance
Faculty	Cognitive	Configural	8.07	0.93	0.06	0.91	0.07	-	+
		Metric	7.5	0.92	0.08	0.90	0.08	0.01	+
		Scalar	10.35	0.86	0.19	0.79	0.17	0.06	-
	Psychosocial	Configural	10.01	0.92	0.03	0.94	0.07	-	+
		Metric	9.9	0.87	0.09	0.90	0.16	0.05	-
	Physiologi- cal	Configural	13.4	0.92	0.05	0.90	0.08	-	+
		Metric	12.2	0.88	0.09	0.82	0.18	0.04	-
	Entire Scale	Configural	7.1	0.91	0.07	0.90	0.08	-	+
		Metric	6.9	0.83	0.09	0.72	0.13	0.08	-

When Table 4 is examined, it is seen that the configural equivalence is provided for the cognitive, psychosocial, physiological dimensions and the whole scale for the groups related to the faculty variable, and the metric equivalence is provided for only the cognitive dimension.

In addition, metric equivalence could not be provided for the entire scale, for psychosocial and physiological dimensions. Scalar and strict equivalence could not be provided for sub-dimensions and the entire scale.

Table 5. Fit statistics obtained from the measurement invariance study for socioeconomic level variable.

		Equivalence	χ^2/df	CFI	RMR	GFI	RMSEA	ΔCFI	Invariance
Socioeconomic Level	Cognitive	Configural	8.02	0.89	0.38	0.45	0.15	-	+
	Psychosocial	Configural	6.9	0.94	0.09	0.92	0.08	-	+
		Metric	7.2	0.90	0.23	0.62	0.14	0.04	-
	Physiological	Configural	12	0.78	0.12	-0.98	0.18	-	+
	Entire Scale	Configural	8.5	0.91	0.05	0.90	0.11	-	+
		Metric	8.1	0.91	0.05	0.89	0.11	0.00	+
Scalar		8.5	0.80	0.09	0.78	0.11	0.11	-	

When Table 5 is examined, it is seen that for groups related to the socioeconomic level variable, Configural equivalence is provided only for the psychosocial dimension and the entire scale, and metric equivalence is provided only for the entire scale.

In addition, metric equivalence was not provided for subdimensions. Scalar and strict equivalence could not be provided for sub-dimensions and the entire scale.

Table 6. The results of the measurement invariance study regarding the variables of gender, faculty and socioeconomic level.

Variable	Sub-Dimensions	Configural Equivalence	Metric Equivalence	Scalar Equivalence	Strict Equivalency
Gender	Cognitive	+	+	-	-
	Psychosocial	+	+	-	-
	Physiological	+	+	-	-
	Entire Scale	+	-	-	-
Faculty	Cognitive	+	+	-	-
	Psychosocial	+	-	-	-
	Physiological	+	-	-	-
	Entire Scale	+	-	-	-
Socioeconomic Level	Cognitive	-	-	-	-
	Psychosocial	+	-	-	-
	Physiological	-	-	-	-
	Entire Scale	+	+	-	-

When Table 6 is examined, configural equivalence was provided for the socioeconomic status variable, except for cognitive and physiological dimensions. Metric equivalence was provided for the cognitive, psychosocial and physiological dimensions for the gender variable, for the cognitive dimension for the faculty variable, and for the socioeconomic status variable only for the whole scale. Scalar and strict invariance, on the other hand, could not provided for any of the variables examined in the study.

4. DISCUSSION and CONCLUSION

Test anxiety can affect individuals of all ages and their families in society (Sieber, 1980). Test anxiety is a variable that disrupts the qualifications of measurement tools, especially their validity, and therefore, it is frequently studied in the literature to measure and determine its reasons.

In the literature, it is seen that almost all studies on test anxiety focus on scale development, scale adaptation, comparison and compilation studies (McDonald, 2001; Pekrun, 2004; Driscoll, 2007; Totan & Yavuz, 2009; Akin *et al.*, 2012; Başol, 2017; Aydın & Bulgan, 2017;

Bozkurt *et al.*, 2017; Bados & Sanz, 2005; Yao-Ting Sung & Tzu-Yang Chao, 2015; Şahin, 2019).

In comparison studies conducted in terms of different variables and groups, it is known that the obtained findings are accepted as the "real" difference between the groups in terms of the measured feature and comparison, interpretation and generalization are made in this direction (Mark & Wan, 2005). It should be kept in mind that these comparative studies and generalizations made without providing evidence regarding measurement invariance may produce erroneous results.

In this study, it was aimed to provide a test anxiety scale that provides evidence for measurement invariance to the literature. For this reason, the measurement invariance of the state test anxiety scale developed by Şahin (2019) was examined in terms of variables of gender, faculty and socioeconomic level.

The findings obtained showed that the state test anxiety scale provided Configural equivalence in groups related to the gender variable. In other words, the factor number and factor loadings pattern related to the measurements obtained from the groups of the gender variable with the state test anxiety scale are equivalent. This is an indication that the items of the measurement tool reveal the same formal structure in the groups examined (Sireci, Patsula & Hambleton, 2005). From this, it can be concluded that the state test anxiety scale measures the same formal structure in the groups of the gender variable.

Metric equivalence was provided within acceptable limits for the gender variable, but this is not valid for the whole scale. If metric equivalence is not achieved, the summability of the scale items is violated and it is emphasized that there may be bias in the items (Erkuş & Selvi, 2019). From here, it can be concluded that the items of the state test anxiety scale produce biased results in men and women. For this reason, it is recommended to examine biases on the scale and items by considering the gender variable and to edit the items found to be biased.

The hypotheses created in measurement invariance studies are examined by comparing the adaptation levels of the preceding model due to the principle of hierarchy (Başusta & Gelbal, 2015). For this reason, Scalar equivalence could not be examined because the metric invariance conditions of the scale were not met. In sub-dimensions of the scale provided metric equivalence, but was not provided scalar equivalence conditions. This shows that the constants used in linear regression equations between latent variables and observed variables are not equivalent between groups (Erkuş & Selvi, 2019). From here, it can be concluded that the scale and its sub-dimensions do not have the same origins for women and men, therefore, comparisons between groups on the basis of the gender variable cannot be made using the state test anxiety scale. On the other hand, strict invariance could not be studied since scalar invariance could not be achieved. Strict equivalence is based on the demonstration that all the predicted parameters are equivalent between groups, but it is known that this condition is often not met in social sciences (Erkuş & Selvi, 2019).

The findings obtained from the measurement invariance study conducted for the faculty variable show that the configural equivalence is provided for the cognitive, psychosocial, physiological sub-dimensions and the whole scale. This is an indicator that the state test anxiety scale reveals the same formal structure in the subgroups of the faculty variable. From here, it can be concluded that the state test anxiety scale measures the same formal structure in different subgroups of the faculty variable. Metric invariance was provided only for the cognitive sub-dimension. This shows that for the subgroups of the faculty variable, only the units of the cognitive sub-dimension of the scale are equivalent within acceptable limits. Units are not equivalent for other dimensions and for the whole scale. For this reason, it is recommended to conduct a bias study on the scale and items considering the faculty variable, and to edit the items found to be biased. Scalar and strict invariance could not be provided for the sub-

dimensions for the faculty variable and for the whole scale. From here, it can be concluded that the scale and its sub-dimensions do not have the same origins between faculties, so the comparison between groups on the basis of the faculty variable cannot be made using the state test anxiety scale. From here, it can be concluded that, apart from the cognitive sub-dimension, the conditionality test anxiety scale and its sub-dimensions produce biased results according to the variable of the faculty attended. This may be due to the different intensity and difficulties of the education programs of the faculties for which data were collected. As a matter of fact, in the faculties of medicine and dentistry, it is possible for students to repeat a year if they fail. But in the health college, students to repeat a course if they fail a course. This may be the reason why students' test anxiety cannot be explained in terms of equivalent origin and equivalent units. Findings on this subject also coincide with the findings of Erözkan (2004).

From the measurement invariance study conducted for the socioeconomic status variable, it is seen that the Configural equivalence was provided only for the psychosocial sub-dimension and the whole scale. In other words, in subgroups related to the socioeconomic status variable, the factor number and factor load of the state test anxiety scale are equivalent for the psychosocial sub-dimension and the whole scale. This is not the case in other sub-dimensions. From this, it can be concluded that the state test anxiety scale measures only the psychosocial sub-dimension in the subgroups of the socioeconomic status variable and the same formal structure in the whole scale. Metric invariance was not provided for sub-dimensions except for the entire scale. This situation shows that the units of the sub-dimensions are not equivalent in the subgroups related to the socioeconomic status variable. Considering the whole scale, the units for the relevant variable are equivalent. However, the scalar and strict invariance conditions for the socioeconomic level variable were not met. Therefore, it appears that the comparison between groups on the basis of the socioeconomic level variable cannot be made using the state test anxiety scale. This may be due to the students with low socioeconomic status not meeting their needs adequately. As a matter of fact, almost all of the situations such as suitable working environment, adequate nutrition, meeting the needs properly, future expectation depend on the socioeconomic level. This may be the reason why test anxieties of students with different socioeconomic levels cannot be explained with equivalent origin and equivalent units. Findings on this subject are in parallel with the findings of Softa, Ulaş Karahmetoğlu & Çabuk (2014).

From the findings obtained, it can be concluded that the state test anxiety scale and its sub-dimensions produce biased results in subgroups regarding the variables of gender, faculty and socioeconomic level, and considering these variables. So comparison between in this groups cannot be made and the findings cannot be generalized. Vanderberg & Lance (2000) state that testing the invariance of only a few parameters in measurement invariance studies may not produce sufficient results. For this reason, it may be suggested to carry out the "partial measurement invariance" study in terms of the variables examined within the scope of this study. On the other hand, this study was carried out on the students of Mersin University on the basis of gender, faculty and socio-economic status variables. It may be suggested to repeat the study on different variables, different universities and different faculties.

Acknowledgments

In this study, the measurement invariance of the state test anxiety scale developed by Şahin (2019) was investigated. I would like to thank Dr. Alper Şahin for allowing us to use this scale in the study.

Declaration of Conflicting Interests and Ethics

The author 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.

ORCID

Huseyin Selvi  <https://orcid.org/0000-0002-3513-0003>

5. REFERENCES

- Akın, A., Demirci, İ., & Arslan, S. (2012). Revize edilmiş sınav kaygısı ölçeği: geçerlik ve güvenilirlik çalışması [Revised Test Anxiety Scale: Validity and Reliability study]. *Eğitim Bilimleri ve Uygulama*, 11(21), 103-118.
- Akın, A., Akın, U., Sarıçam, H., Aşut, S., Arslan, S., Demirci, İ., Toprak, H., & Çardak, M. (2013). *Friedben sınav kaygısı ölçeği Türkçe formu'nun geçerlik ve güvenilirliği* [The validity and reliability of the Turkish form of the friedben test anxiety scale]. Paper presented at the international conference on innovation and challenges in education. Dumlupınar University.
- Alpert, R., & Haber, R. N. (1960). Anxiety in academic achievement situations. *The Journal of Abnormal and Social Psychology*, 61(2), 207-215. <https://doi.org/10.1037/h0045464>
- Alpar, R. (2016). *Spor, sağlık ve eğitim bilimlerinde örneklerle uygulamalı istatistik, geçerlik ve güvenilirlik* [Applied statistics validity and reliability with examples in sports, health and education sciences]. DetayYayıncılık.
- Alıcı, D. (2013). Okula yönelik tutum ölçeği'nin geliştirilmesi: güvenilirlik ve geçerlik çalışması [Development of attitude scale towards school: reliability and validity study]. *Eğitim ve Bilim*, 38(168).
- Aydın, U., & Bulgan, G. (2017). Çocuklarda sınav kaygısı ölçeği'nin Türkçe uyarlaması [Turkish adaptation of the test anxiety scale in children]. *İlköğretim Online*, 16(2), 860-899. <https://doi.org/10.17051/ilkonline.2017.304742>
- Bados, A. L., & Sanz, P. (2005). Validation of the revised test anxiety scale and the friedben test anxiety scale in a spanish sample. *Ansiedad y estrés*, 11(2), 163-174.
- Başol, G. (2017). Ayda sınav kaygısı ölçeği: geçerlik ve güvenilirlik çalışması [IDA test anxiety scale: validity and reliability study]. *Uluslararası Eğitim Bilimleri Dergisi*, 4(13), 173-193. <https://doi.org/10.16991/INESJOURNAL.1506>
- Başusta, B. (2010). Ölçme eşdeğerliği [Measurement invariance]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1(2), 58-64.
- Başusta, N.B., & Gelbal, S. (2015). Gruplararası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: Pisa öğrenci anketi örneği [Testing measurement invariance in comparisons between groups: Pisa student survey sample]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 30(4), 80-90.
- Benson, J., & El-Zahhar, N. (1994). Further refinement and validation of the revised test anxiety scale. *Structural Equation Modelling*, 1(3), 203-221. <https://doi.org/10.1080/10705519409539975>
- Bozkurt, S., Ekitli, G. B., Thomas, C. L., & Cassady, J. C. (2017). Validation of the Turkish version of the cognitive test anxiety scale-revised. *Sage Open*, 1(1), 1-9. <https://doi.org/10.1177/2158244016669549>
- Büyüköztürk, Ş. (1997) Araştırmaya yönelik kaygı ölçeğinin geliştirilmesi [Development of an anxiety scale for research]. *Eğitim Yönetimi*, 3(1), 453-464.
- Cassady, J.C., & Johnson, R. E. (2002). Cognitive test anxiety and academic procrastination. *Contemporary Educational Psychology*, 27(1), 270-295. <https://doi.org/10.1006/ceps.2001.1094>
- Cheung, G.W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, 31(2), 187-212. <https://doi.org/10.1177/0022022100031002003>

- Cheung, G., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance structural equation modeling. *A Multidisciplinary Journal*, 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5
- Cüceloğlu, D. (1998). *İnsan ve davranışı [Human and his behaviour]*. Remzi Kitabevi.
- Erkuş, A., & Selvi, H. (2019). *Psikolojide ölçme ve ölçek geliştirme III: ölçek uyarlama ve norm geliştirme [Measurement and scale development in psychology III: scale adaptation and norm development]*. Pegem Akademi.
- Ergene, T. (1994). Sınav kaygısı ile başa çıkma programının etkinliği [Effectiveness of coping program with test anxiety]. *Psikiyatri, Psikoloji ve Psikofarmakoloji (3P) Dergisi*, 2(1), 9-16.
- Erözkan, A. (2004). Üniversite öğrencilerinin sınav kaygısı ve başa çıkma davranışları [Test anxiety and coping behaviours of university students]. *Muğla Üniversitesi SBE Dergisi*, 12(1), 13-38.
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik (spss ve lisrel uygulamaları) [Multivariate statistics for social sciences]*. Pegem Akademi.
- Driscoll, R. (2007). Westside test anxiety scale validation. *Psychology*, 8(14), 1-6.
- Geen, R.G. (1985). Test anxiety and visual vigilance. *Journal of Personality and Social Psychology*, 49(4), 963-970. <https://doi.org/10.1037/0022-3514.49.4.963>
- Gençdoğan, B. (2006). Lise öğrencilerinin sınav kaygısı ile boyun eğicilik düzeyleri ve sosyal destek algısı arasındaki ilişkiler [Relationships between high school students' test anxiety and their level of submission and social support perception]. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 7(1), 153-164.
- Hill, K.T., & Wigfield, A. (1984). Test anxiety: a major educational problem and what can be done about it. *The Elementary School Journal*, 85(1), 105-126. <https://doi.org/10.1086/461395>
- Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(1), 117-144. <https://doi.org/10.1080/03610739208253916>
- Kline, R.B. (2005). *Principles and practice of structural equation modeling*. The Guilford Press.
- Mark, B. A. & Wan, T. T. (2005). Testing measurement equivalence in a patient satisfaction instrument. *Western Journal of Nursing Research*, 27(6), 772-787. <https://doi.org/10.1177/0193945905276336>
- McDonald, A. S. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology*, 21(1), 89-101. <https://doi.org/10.1080/01443410020019867>
- Sapir, S. & Aronson, A.E. (1990). The relationship between psychopathology and speech and language disorder in neurological patients. *Journal of Speech Hearing Disorder*, 55(1), 503-509. <https://doi.org/10.1044/jshd.5503.503>
- Sieber, J. E. (1980). *Defining test anxiety: problems and approaches*. Lawrence Erlbaum Associates.
- Sireci, S.G, Patsula, L., & Hambleton, R.K. (2005). *Statistical methods for identifying flaws in the test adaptation process*. Lawrence Erlbaum Associates.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543. <https://doi.org/10.1007/BF02294825>
- Pekrun, R. (2004). Test Anxiety and academic achievement, *Encyclopedia of the Social and Behavioral Sciences*. <https://doi.org/10.1016/B0-08-043076-7/02451-7>
- Softa Kaçan, H., Ulaş Karahmetoğlu G., & Çabuk, F. (2014). Lise son sınıf öğrencilerinin sınav kaygısı ve etkileyen faktörlerin incelenmesi [Examining of test anxiety of senior high

- scholl students and the factors affecting them]. *Kastamonu Eğitim Dergisi*, 23(4), 1481-1494.
- Sung, Y.T. & Chao, T.Y. (2015). Construction of the examination stress scale for adolescent students. *Measurement and Evaluation in Counseling and Development*, 48(1), 44-58. <https://doi.org/10.1177/0748175614538062>
- Şimşek, Ö.F. (2007). *Yapısal eşitlik modellemesine giriş (temel ilkeler ve lisrel uygulamaları) [Introduction to structural equating modeling (basic principles and applications of lisrel)]*. Ekinoks.
- Tabachnick, B. & Fidell, L. S. (2001). *Using multivariate statistics*. Allyn & Bacon.
- Şahin, A. (2019). Durumluk sınav kaygısı ölçeği (duskö): geçerlik ve güvenilirlik çalışması [State test anxiety scale: validity and reliability scale]. *Trakya Eğitim Dergisi*, 9(1), 78-90. <https://doi.org/10.24315/tred.450423>
- Şenocak, M. (2014). *Biyoistatistik ve araştırma yöntembilimi [Biostatistics and research methodology]*. İstanbul Tıp Yayınevi.
- Totan, T. & Yavuz, Y. (2009). Westside sınav kaygısı ölçeğinin Türkçe formunun geçerlik ve güvenilirlik çalışması [Validity and reliability study of the Turkish form of the Westside test anxiety scale]. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, 17(1), 95-109.
- Turgut, M. F. & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Pegem Yayıncılık.
- Uzun, B. & Öğretmen, T. (2010). Fen başarısı ile ilgili bazı değişkenlerin tıms-s-r örnekleminde cinsiyete göre ölçme değişmezliğinin değerlendirilmesi [Evaluation of measurement invariance of some variables related to science achievement by gender in TIMSS_R sample]. *Education and Science*, 35(1), 26-35.
- Öner, N. (1990). *Sınav kaygısı envanteri el kitabı [Test anxiety inventory handbook]*. Yüksek Öğretimde Rehberliği Tanıtma ve Rehber Yetistirme Vakfı Yayınları.
- Özgülven, İ. E. (2007). *Psikolojik testler [Psychological tests]*. PDREM Yayınları.
- Yao-Ting Sung & Tzu-Yang Chao (2015) Construction of the examination stress scale for adolescent students. *Measurement and Evaluation in Counseling and Development*, 48(1), 44-58. <https://doi.org/10.1177/0748175614538062>
- Vanderberg, R.J. & Lance, C. (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>
- Van der Vijver, F. J. R. & Poortinga, Y. H. (2005). *Conceptual and methodological issues in adapting tests*. Lawrence Erlbaum Associates.
- Wren, D. G. & Benson, J. (2004). Measuring test anxiety in children: Scale development and internal construct validation. *Anxiety, Stress and Coping*, 17(3), 227-240. <https://doi.org/10.1080/10615800412331292606>
- Zeidner, M. (2007). *Test anxiety in educational contexts: concepts, findings, and future directions*. Elsevier.

Comparison of G and Phi coefficients estimated in generalizability theory with real cases

Kaan Zulfikar Deniz ^{1,*}, Emel Ilican ²

¹Ankara University, Faculty of Educational Sciences, Ankara, Turkey

²Republic of Turkey Ministry of National Education, Ankara, Turkey

ARTICLE HISTORY

Received: Oct.03, 2020

Revised: Jan. 16, 2021

Accepted: May 31, 2021

Keywords:

Reliability,
Generalizability theory,
Decision study,
Item difficulty index.

Abstract: This study aims to compare the G and Phi coefficients as estimated by D studies for a measurement tool with the G and Phi coefficients obtained from real cases in which items of differing difficulty levels were added and also to determine the conditions under which the D studies estimated reliability coefficients closer to reality. The study group for this research consisted of 80 seventh-grade students from various public and private secondary schools in the provinces of Ankara, Istanbul, and Adana in Turkey. Four raters who served as Turkish teachers in various public secondary schools in Ankara were included in this study. A data collection tool consisting of 12 tasks was prepared to measure the participating seventh grade students' written expression skills in Turkish. The equation of the G and Phi coefficients estimated in the D study and obtained through the real cases was observed only when six tasks with item difficulty indexes close to the mean difficulty of the test were added in such a way that the mean difficulty of the test never changed. In other cases, where the mean difficulty of the test changed because of the addition of easy or difficult tasks, it was determined that the reliability coefficients estimated in the D study and obtained in real cases were similar, but they had different values.

1. INTRODUCTION

The most important psychometric properties sought in a measurement tool are grouped under the concepts of reliability, validity, and usability. Reliability is defined as the ability to repeat measurements of a feature performed on the same individuals with the same measurement tool under similar conditions or to give consistent results (Baykul, 2015; Crocker & Algina, 1986; Nitko, 2004). According to the Classical Test Theory (CTT), reliability coefficient is to be estimated regarding reliability. While making this estimation, the effect of variable situations such as the content, construct and application of items and tests on test scores is examined using various reliability estimation methods (Aiken, 2009; Anastasi & Urbina, 1997).

In some cases, when utilising reliability estimation methods that are based on CTT, any single application of the CTT model cannot clearly differentiate among multiple sources of error. To find a solution to the limitations of CTT, the Generalizability theory (G) was developed, which

*CONTACT: Kaan Zulfikar Deniz ✉ zlfkrdnz@yahoo.com 📍 Ankara University, Faculty of Educational Sciences, Ankara, Turkey

allows for the calculation of reliability coefficients based on differing sources of variation (i.e., error) that may occur within a single study. G theory liberalizes classical theory by employing ANOVA methods that allow an investigator to untangle multiple sources of error (Brennan, 2001).

As a result, with G theory studies, any facet (i.e., source) of error such as rater, time, forms, and/or item is evaluated simultaneously and as a group in order to estimate a comprehensive and single reliability coefficient. The basic idea of G theory is that error variance derives from different sources of variability as well as from the interactions that take place between them. In other words, the superiority of G theory over CTT is that different error sources can be simultaneously estimated through a single analysis. This process is completed with the help of variance analysis that allows for multiple variance sources to be analysed through a single analysis, while at the same time a determination can be made regarding the size of each variance source (Brennan, 2001; Shavelson & Webb, 1991).

Also, G theory allows for the calculation of two differing reliability coefficients regarding both relative decisions; namely, those decisions based on individual performance and the absolute decisions of these individual performances. As a result, these are the generalizability coefficients that make up the relative evaluations and the Phi (Φ) coefficient for the absolute evaluations. Importantly, generalizability (G) and decision (D) studies are carried out in order to determine the reliability coefficients utilising G theory. Through the G study process, the variance components of scores and the interactions between them are estimated simultaneously through ANOVA. These estimated variance components are then utilised in the subsequent step of the D study. In a D study, in order to create measurement situations with sufficient reliability, measurements are organized so that the measurement error can be minimised (Brennan, 2001; Shavelson & Webb, 1991).

To explain, a D study is an estimation, use, and interpretation of variance components in order to formulate decisions according to already well-defined measurement processes (Crocker & Algina, 1986). For example, in a case where more than one rater scores a group of students' ability to solve mathematical problems, a G study that utilises three raters and 20 items is followed by a D study; as a result, differing numbers of raters and differing numbers of items can be estimated and through this process the G and Phi coefficients can also be estimated. However, in the results of the D study, the G and Phi coefficients are provided when adding or subtracting items from the measurement tool, yet no information is given in regard to the difficulty of these items. For example, in a D study, the G and Phi coefficients are estimated after at least three items have been added to a measurement tool, but to what extent these coefficients are sensitive to the item difficulty index (p_j) of the added items remains unknown, and whether the items are easy or difficult also remains undefined.

In the literature, there are many studies in which items related to various measurement and evaluation practices have been considered a source of variability and reliability studies based on G Theory (Choi & Wilson, 2018; Çakıcı Eser & Gelbal, 2013; Deliceoğlu & Çıkrıkçı Demirtaşlı, 2012; Demir, 2016; Doğan & Anadol, 2017; Doğan & Bıkmaz Bilgen, 2017; Güler, 2011; Güler et al., 2014; Gülle et al., 2018; Hathcoat & Penn, 2012; Hill et al., 2012; Scherbaum et al., 2018; Solano-Flores & Li, 2013; Yılmaz Nalbantoğlu & Gelbal, 2011). Furthermore, in some of these studies (Doğan & Anadol, 2017; Scherbaum et al., 2018; Yılmaz Nalbantoğlu & Gelbal, 2011) comparisons were also made regarding the use of crossed and nested research designs within the scope of G theory. In other studies (Doğan & Bıkmaz Bilgen, 2017; Güler et al., 2014; Gülle et al., 2018; Hathcoat & Penn, 2012; Solano-Flores & Li, 2013) it was observed whether the reliability of performance-based measures could be examined through G theory. In addition, there are several studies (Çakıcı Eser & Gelbal, 2013; Deliceoğlu & Çıkrıkçı Demirtaşlı, 2012; Demir, 2016; Güler, 2011) in which the reliability of measurements was

examined through methods other than G theory. Apart from these studies, there are few studies in which the G and Phi coefficients estimated through a D study were compared with the reliability coefficients in real cases. Atılgan and Tezbaşaran (2005) compared the G and Phi coefficients acquired from D studies and real situations from a number of different raters by using data from two successive years of special skill selection exams conducted from a student selection program. In another study, the G and Phi coefficients estimated for two, three, and four raters from real cases in which it was not possible to randomly select raters from a population universe, were compared with the results from relevant D studies (Kamış & Doğan, 2017). However, there was no identified study that compared the predicted G and Phi coefficients in the D studies as well as the obtained G and Phi coefficients from real cases in which there were items of varying difficulty levels added and/or removed from the measuring tool.

While test items are considered as a source of variability and reliability in which studies based on G theory have been carried out, there can be a determination made to change the number of test items in order to obtain the reliability coefficients that have previously been predicted in the D study. At this stage, it is believed that knowing the difficulty level of items and under which conditions the D study accurately estimates the reliability coefficients in real cases will ultimately contribute to a more meaningful interpretation of D studies. In addition, this information is expected to facilitate the selection of items as a way of obtaining reliability coefficients as estimated in the D study as well as supporting the efficient completion of reliability studies.

As a result, the aim of this study was to compare the G and Phi coefficients as estimated by D studies as well as the G and Phi coefficients obtained in real cases in which the items of differing difficulty levels were added and to also determine the conditions under which the D studies estimated the reliability coefficients more in line with the real situation. In this respect, easy, moderate or difficult items were added to a measuring tool and these additional items were meant to reflect two conditions, both modifying and not-modifying the mean difficulty of the test. The sub-objectives determined for the general purpose of this study are as follows:

- a) To compare the G and Phi coefficients estimated by the D studies and the G and Phi coefficients obtained by increasing the total number of tasks to 18 that change the mean difficulty of the test: with six easy tasks; with six moderate tasks; and with six difficult tasks.
- b) To compare the G and Phi coefficients estimated by the D studies and the G and Phi coefficients obtained by increasing the total number of tasks to 18 that did not change the mean difficulty of the test: with two easy, two moderate, and two difficult tasks; and with six moderate tasks.
- c) To determine whether there were any significant differences between the G and Phi coefficients estimated by D studies and the G and Phi coefficients obtained in various real cases, where the total number of tasks was increased to 18.

2. METHOD

This section indicates the research design used in the study, the study group, the data collection, and the analysis of the data.

2.1. Research Design

This study followed a survey research model in which attempts were made to define a situation under a set of circumstances without changing and/or influencing that situation in any way. In addition, since this research was aimed at generating information, it was prepared and carried out in a basic manner (Büyüköztürk et al., 2012; Fraenkel et al., 2015; Karasar, 2016).

2.2. Study Group

The study group for this research consisted of 80 seventh grade students (ages 12-13) studying in various public and private secondary schools located in Ankara (n=30, 37.5%), Istanbul (n=25, 31.5%), and Adana (n=25, 31.5%), Turkey during the 2016-17 academic year. Of the students in the study group, 34 (42.5%) were male and 46 (57.5%) were female. Students for the study were selected from 26 schools, 20 of which were public and 6 were private. The study group of the research was selected from the sample of a study conducted by the Republic of Turkey Ministry of National Education that aimed to evaluate the Turkish written expression skills of students from various grade levels. The students who were applied one of the seventh grade test forms used in the study and raters assigned for item scoring were included in the study group of this research study. Four raters worked as Turkish language teachers in various public secondary schools. These teachers had previously received training on item scoring and were also informed about the use of rubrics prepared for this study. Importantly, the students and teachers included in this study group were selected from different schools.

2.3. Data Collection

The data of this research were obtained from the Ministry of National Education by official correspondence for research permission. In the data collection process of the study, the students and raters from the study group were briefly informed about the study process. A skill test consisting of 12 tasks was first prepared and then applied in order to measure the students' Turkish written expression skills. Then, four raters scored the skill test independently and the data were collected for analysis. Through the application of student tasks, each student answered the same 12 tasks and the four raters via a scoring rubric prepared for the test scored each student's responses. Thus, the research design for this study can be considered to follow a fully crossed (sxtxr) design.

2.3.1. Data collection tool

The test utilised in this study consisted of 12 tasks prepared to measure the Turkish written expression skills of seventh grade students. In completing the tasks included in the test, the participating students had to create sentences and paragraphs with a variety of characteristics. In the first task of sentence knowledge, the students were asked to select at least five words from a word pool provided and then form a sentence consisting of a minimum of eight words in total. In the second task, these students were asked to form a sentence consisting of a minimum of eight words in accordance with a visual prompt. In the third task, a dialogue was provided to the students and they were asked to complete the dialogue with an appropriate sentence consisting of at least five words. The subsequent four tasks of the test were related to a persuasion paragraph and then the remaining final five tasks involved writing a petition. Rubrics that can be scored from 0 to 4 were developed for each task of the test in this study. The experts in the study team formed by the Ministry of National Education developed these rubrics. As a result, the highest score a student could receive from task scoring was 48 and the lowest possible score was 0.

2.4. Data Generation and Analysis

In this study initially, variance sources were estimated from the G study of 12 tasks. Then, D study was conducted by using these variance sources and increasing the number of tasks to 18. The G and Phi coefficients were estimated for 18 tasks within the test through the D study. These coefficients were compared with the reliability coefficients estimated from the real situation of 18 tasks subsequent to adding tasks of various difficulty indexes, which ultimately changed the test's mean difficulty for some of the cases but not all. Since all of the 12 tasks initially included in the scale were rated at a moderate level of difficulty, there were randomly

selected tasks from the scale that were reused by adding moderate tasks to the test. The easy tasks added to the test were produced by increasing the points of the easiest tasks in the test by two points each except for those with full points. The difficult tasks added to the test were artificially created by dividing the scores of the most difficult tasks in the test into three and then decreasing the scores downward. Finally, the significance of the differences between the estimated G and Phi coefficients as well as the G and Phi coefficients obtained in various real cases was examined through a Fisher's z' test. Variance sources and the G and Phi coefficients were estimated in the analysis performed through crossed design (sxtxr) obtained by grading 80 students by four raters for 12 tasks. The EduG 6.1-e program was utilised in analysing the data obtained from this study.

3. RESULTS

In the results section, first, those results related to the estimated variance of the sources of variability from the fully crossed design are provided for different cases where the number of tasks was either 12 or 18. Second, in accordance with the sub-objectives of this study, the findings from the comparison of the G and Phi coefficients estimated through the D study as well as the G and Phi coefficients obtained in real cases were interpreted. In regards to the analysis findings, results related to the estimated variance components are provided in [Table 1](#).

Table 1. *Analysis of variance results and variance component estimates for students, tasks, raters, and their interactions.*

	Number of Tasks	Source of Variance	df	MS	Variance Component Estimates	Percentage of Total Variance Estimates
Actual status	12	s	79	97.63	1.66	22.50
		t	11	70.75	0.10	1.40
		r	3	278.77	0.25	3.40
		st	869	11.67	2.39	32.40
		sr	237	8.43	0.53	7.10
		tr	33	28.70	0.33	4.50
		str	2607	2.11	2.11	28.60
The mean difficulty of the test changes	18 (Six easy tasks added)	s	79	142.62	1.69	21.20
		t	17	502.85	1.48	18.60
		r	3	394.57	0.25	3.20
		st	1343	9.06	1.80	22.60
		sr	237	13.63	0.65	8.20
		tr	51	21.19	0.24	3.00
		str	4029	1.84	1.84	23.10
The mean difficulty of the test changes	18 (Six moderate tasks added)	s	79	187.59	2.28	26.30
		t	17	92.97	0.16	1.90
		r	3	306.76	0.19	2.20
		st	1343	14.84	3.21	37.10
		sr	237	10.85	0.49	5.70
		tr	51	27.98	0.32	3.80
		str	4029	2.00	2.00	23.10
The mean difficulty of the test changes	18 (Six difficult tasks added)	s	79	84.87	0.98	13.30
		t	17	747.30	2.24	30.50
		r	3	207.02	0.12	1.70
		st	1343	9.05	1.87	25.50

Table 1. *Continued.*

		sr	237	7.06	0.31	4.20
		tr	51	23.75	0.28	3.80
		str	4029	1.55	1.55	21.10
		s	79	149.38	1.77	23.80
		t	17	65.58	0.09	1.20
	18	r	3	405.34	0.26	3.40
	(Six moderate	st	1343	11.79	2.45	32.90
	tasks added)	sr	237	12.09	0.56	7.50
		tr	51	27.01	0.31	4.20
		str	4029	2.01	2.01	27.00
The mean difficulty of the test remains unchanged		s	79	134.02	1.59	19.50
	18	t	17	602.77	1.77	21.70
	(Six easy,	r	3	457.06	0.29	3.50
	moderate and	st	1343	9.00	1.80	22.10
	difficult tasks	sr	237	12.24	0.58	7.10
	added)	tr	51	29.69	0.35	4.30
		str	4029	1.78	1.78	21.80

Table 1 illustrates that in a majority of the cases studied; the ST interactive variance component value had the highest rate of total variance. Accordingly, it can be stated that the difficulty levels of the tasks differed from one student to another in the cases examined. In addition, when six difficult tasks were added to the test and the mean difficulty of the test decreased, it was determined that instead of ST, the T variance component value (2.24) had the highest rate (30.5%) in the total variance. Thus, after adding difficult tasks, it can be said that the tasks in the test become very different from each other in terms of their difficulty level. Among all the cases examined, it was observed that when six items with moderate difficulty were added to the test and the average difficulty of the test varied, the ST-interactive variance component value (3.21) was found to have the highest value. Here, it was the source of variability that explained the total variance with the highest rate (37.1%). The second highest rate in total variance generally belongs to residual component. Accordingly, it can be said that there is interaction between students, tasks, and raters and there are systematic or unsystematic sources of variability that cannot be measured in this study. In these cases, the variance component for students was generally high in total variance. This result demonstrated that the measured characteristics of students differed from each other; as a result, the measurement process proved successful in distinguishing students from one another according to their results. Finally, in all of the cases, it can be stated that the raters generally provided consistent scores because the overall rater ratio variance in total was negligible.

Table 2 illustrates the G and Phi coefficients obtained when the number of tasks in the test was actually 12 and then an estimate was produced for 18 tasks in the D study.

Table 2. *D study results.*

Number of Tasks	G	Phi
12	0.82	0.79
18	0.85	0.82

Table 2 displays that the G and Phi coefficients obtained from real cases where the number of tasks in the test was 12 were 0.82 and 0.79. Furthermore, according to the results of the decision study, in which the number of tasks was 18, the G and Phi coefficients were 0.85 and 0.82.

3.1. Results for the First Sub-Objective

The mean difficulty of the test and the G and Phi coefficients obtained in the G and D studies from the cases where the number of test tasks was increased to 18 and the test mean difficulty changed are provided in [Table 3](#).

Table 3. *G and phi coefficients obtained in cases where the test mean difficulty changed.*

Number of Tasks	Mean difficulty of the test	Actual Status		Decision Studies (estimated for 12 tasks)	
		G	Phi	G	Phi
12 ^a	0.51	0.82	0.79	-	-
18 ^b	0.60	0.85	0.79		
18 ^c	0.48	0.87	0.85	0.85	0.82
18 ^d	0.38	0.83	0.73		

^aOriginal scale

^bAdded six easy tasks

^cAdded six moderate tasks

^dAdded six difficult tasks

As can be seen in [Table 3](#), when six tasks of moderate difficulty ($p_j = 0.41-0.58$) were added and the mean difficulty of the test was least varied, the G and Phi coefficients were 0.87 and 0.85 for the first case. In addition, when the test had 12 tasks, the G and Phi coefficients estimated for the 18 tasks within the D study were 0.85 and 0.82. As a result, it can be stated that the G and Phi coefficients estimated for the 18 tasks from the D study were relatively smaller than those obtained in the real case where six moderate tasks had been added to the test.

The G and Phi coefficients were 0.85 and 0.79 for the second case where six easy tasks ($p_j = 0.76-0.80$) were added to the test and the test mean difficulty had changed more than the first case. Through the analysis results it was recognised that the G coefficient estimated in the D study for 18 tasks was equal to the G coefficient obtained in the real case where six easy tasks had been added to the test. Also, the Phi coefficient obtained after adding easy tasks to the test was less than the estimated Phi coefficient (0.82) from the D study with 18 tasks.

Finally, when six difficult tasks ($p_j = 0.12-0.13$) were added to the test, it was recognised that the mean difficulty of the test decreased/increased considerably compared to the first two cases. In this case, the G and Phi coefficients were acquired as 0.83 and 0.73 for the real situation in which difficult tasks had been added to the test, and as a result, the values were smaller than the G and Phi coefficients estimated in the D study for 18 tasks. In addition, these values ($G = 0.83$ and $\Phi = 0.73$) differed from the G (0.85) and Phi (0.82) coefficients estimated in the D study as compared to the other two cases where the mean difficulty of the test had changed less.

3.2. Results for the Second Sub-Objective

The mean difficulty of the test and G and Phi coefficients obtained from G and D studies where cases that had the number of test tasks increased to 18 and the test mean difficulty did not change are provided in [Table 4](#).

Table 4. *G and phi coefficients obtained in cases where the mean difficulty of the test did not change.*

Number of Tasks	Actual Status			Decision Studies (estimated for 12 items)	
	Mean difficulty of the test	G	Phi	G	Phi
12 ^a	0.51	0.82	0.79	-	-
18 ^b	0.51	0.85	0.78		
18 ^c	0.51	0.85	0.82	0.85	0.82

^aOriginal scale^bTwo of the six tasks added were easy, two were moderate and two were difficult.^cAdded six moderate tasks

As Table 4 presents the G and Phi coefficients were 0.85 and 0.78 in the first case when two easy ($p_j = 0.78$ and $p_j = 0.80$), two moderate ($p_j = 0.58$), and two difficult tasks ($p_j = 0.12$ and $p_j = 0.13$) were added to the test and the mean difficulty of the test ($p_j = 0.51$) remained unchanged. Very close to these values, next, in the second case the values remained close with the G and Phi coefficients obtained at 0.85 and 0.82 when six moderate tasks ($p_j = 0.41-0.58$) were added and the mean difficulty ($p_j = 0.51$) of the test again remained unchanged. As a result, the G coefficients acquired in both real cases were found to be equal to the G coefficient that had been estimated in the D study for the 18 tasks. In addition, the Phi coefficient (0.78) obtained in the first case was less than the Phi coefficient (0.82) estimated in the D study for the 18 tasks. Importantly, among all of the cases examined, only within the second case was the obtained G (0.85) and Phi (0.82) coefficient equal to the G (0.85) and Phi (0.82) coefficient estimated in the decision study for 18 tasks.

3.3. Results for the Third Sub-Objective

In order to determine whether the differences between the G and Phi coefficients estimated by the D studies in this research and those obtained through real cases were significant, all of the G and Phi coefficients were converted to z scores through the Fisher Z-transformation test. Accordingly, the G and Phi coefficients obtained and as well as the Fisher's z' scores calculated are provided in Table 5.

Table 5. *G and phi coefficients obtained in cases where the number of test tasks were 12 or 18 including the Fisher z' scores.*

	Number of tasks	Decision Studies (estimated for 12 tasks)		Actual Status	
		G	Phi	G (Fisher z')	Phi (Fisher z')
The mean difficulty of the test changes	18 ^a			0.85 (0)	0.79 (0.20)
	18 ^b	0.85	0.82	0.87 (-0.18)	0.85 (-0.23)
	18 ^c			0.83 (0.16)	0.73 (0.54)
The mean difficulty of the test remains unchanged	18 ^d			0.85 (0)	0.78 (0.26)
	18 ^e	0.85	0.82	0.85 (0)	0.82 (0)

^aAdded six easy tasks^bAdded six moderate tasks^cAdded six difficult tasks^dTwo of the six tasks added were easy, two were moderate and two were difficult^eAdded six moderate tasks

When the Fisher's z' test results provided in [Table 5](#) are examined, it can be recognised that all of the z' values calculated were between -1.96 and +1.96 (Kenny, 1987). As a result, this finding shows that there was not a significant difference between the reliability coefficients estimated in the D studies and those obtained in real cases.

4. DISCUSSION and CONCLUSION

As a result of the analyses conducted in this study, it was observed that the reliability coefficients predicted in the D studies and those obtained in real cases were different; however, in general, they remained quite close to each other. When the differentiated Phi coefficients were examined, it was also found that the values estimated in the D studies and obtained through real cases were different for four of the five cases examined. Next, the values obtained in real cases for the G coefficient were equal to the estimated G coefficient from the D study in three of the five cases studied. As a result, it can be stated that the reliability coefficients in the case where items were added to the measurement tool and estimated through the D studies, the Phi coefficient was more sensitive to the difficulty level of the added items in comparison to the G coefficient. This result is thought to be related to the fact that the item variance considered when calculating the Phi coefficient increased more than the G coefficient when the easy or difficult items were added to the measuring instrument (Brennan, 2001). In this study, it was observed that the task variance, which has the smallest value in the total variance in the real situation, increases when tasks with different difficulty levels are added to the test. Added easy or difficult tasks caused the Phi coefficient to decrease as the task variance and absolute error variance increased. As a result, although it was estimated that the Phi coefficient would increase if the number of tasks was increased from 12 to 18, it was instead recognised that the Phi coefficient did not change and/or decrease from the addition of either any easy and/or difficult tasks to the test. Furthermore, the relative error variance utilised in determining the G coefficient was acquired with the interactive variance components that included the students and was ultimately less affected by the change in variance that arose from the test tasks and was generally close in value to those predicted in the D studies (Güler et al., 2012). When the literature for this study was examined, it was determined that there were findings which increased the number of items that had a positive effect of ensuring the desired quality of reliability as well as that reliability would increase as the number of items increased (Ankenmann & Stone, 1992; Bıkmaz Bilgen & Doğan, 2017; Güler & Yetim, 2008; Hulin et al., 1982; Tavşancıl, 2005). In other previous studies, it was concluded that low reliability was in effect due to the low number of substances (i.e., items) included (Güler & Yetim, 2008; Kaya, 2005). This is important because in research where D studies were conducted based on G theory, it was concluded that reliability would increase if the number of items in the test increased (Deliceoğlu & Çıkrıkçı Demirtaşlı, 2012; Demir, 2016; Doğan & Bıkmaz Bilgen, 2017; Gülle et al., 2018; Hathcoat & Penn, 2012). However, as was determined in this study, an increase in the number of items may in effect not actually provide a higher reliability coefficient in all cases. Similarly, the research study by Giray and Şahin (2012) revealed that solely reducing the number of items did not in itself lead to a decrease in reliability.

In this present study, the equality of both the G and Phi coefficients obtained in the real situation as well as estimated in the D study could only be witnessed when six tasks of moderate difficulty ($p_j=0.41-0.58$) were added to the test but did not change the mean difficulty of the test ($p_j=0.51$). In addition, it was also determined that the difference between reliability coefficients, especially the Phi coefficient, which was obtained for the real cases and estimated in the D study, increased more when the mean difficulty of the test changed as a result of adding items. Accordingly, it can be stated that when the reliability coefficients estimated in the D study from the addition of items to the test were expected to be obtained in a real case, it would be beneficial in future research to select items that do not change the mean difficulty of the test or items with the

difficulty indexes closest to the mean difficulty of the existing test. On the other hand, it was also determined that there were no significant differences between the G and Phi coefficients obtained in various situations when the number of tasks in the test was actually 18 and the G and Phi coefficients estimated as a result of the D studies were made with 12 tasks. However, it is recommended that this situation be re-examined by utilising different measurement tools when added items can be changed in the mean difficulty of the test. In addition, it can be stated that these examinations may be useful for a test where the percentage of item variance in the total variance is greater. This is recommended because the G studies conducted in this study generally showed that item variance made up a small percentage of the total variance. While, in studies by Demirel and Epçaçan (2012) and Katrancı and Yangın (2003), very easy and very difficult items were removed from the test for a similar purpose, and as a result, sufficient KR-20 reliability coefficients were obtained. Similarly, for other studies (Çakır & Aldemir, 2011; Kaplan & Duran, 2016), some test items were excluded in order to obtain a higher reliability coefficient, but unfortunately no information was provided regarding the item difficulty index of the extracted items.

Also, previous research studies have pointed out that in decision studies with G theory the G and Phi coefficients will increase if the number of items and raters are increased (Güler et al., 2012). However, in this current study, it was determined that the Phi coefficient could remain the same or even decrease if easy or difficult tasks were added to a moderate scale. As a result, it was recognised that the Phi coefficients obtained from the real case where the number of items was increased might be smaller than the estimated Phi coefficients for the number of items in the D study. In addition, Kemiş and Doğan (2017) revealed that even though the number of raters increased in their study, the reliability coefficients could possibly decrease and could even be lower values than the predicted values from the related D studies. Furthermore, Atılgan and Tezbaşaran (2005) determined that the reliability coefficients obtained in real cases where the number of raters was increased were smaller than the reliability coefficients predicted within the D studies. In this current study, it is discovered that if the number of tasks increased, the G and Phi coefficients obtained for real situations may be larger, smaller, or equal to the G and Phi coefficients estimated in the D studies. As a result of this finding, it is believed that the difference between the findings of the two previous studies may be a result of whether or not the items/raters have been randomly selected from the population universe or the ratio of the item/rater variance in regards to the total variance within the study.

Finally, the significant findings of this study may show that since the reliability of real situations cannot be estimated completely and/or systematically through the utilisation of D studies in G theory, then it is recommended that these factors be taken into consideration when interpreting the results of future D studies. Since the scores on easy and/or difficult tasks were artificially produced in this study, future researchers are recommended that they perform similar studies utilising a real pool of items, in which the easy or difficult items can be added to a test at any point and with no concern of its effect on the outcome and/or validity of the test.

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 author(s).

Authorship Contribution Statement

Kaan Zülfikar Deniz: Investigation, research design, literature review, supervision and writing the manuscript. **Emel Ilican:** Research design, literature review, methodology, data collection, data analysis, and writing the manuscript.

ORCID

Kaan Zülfiyar Deniz  <https://orcid.org/0000-0003-0920-538X>

Emel Ilcan  <https://orcid.org/0000-0003-4244-6441>

5. REFERENCES

- Aiken, L., R. (2009). *Psychological testing and assessment* (Twelfth ed.). Pearson.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Pearson.
- Ankenmann, R. D., & Stone, C. A. (1992, April). *A monte carlo study of marginal maximum likelihood parameter estimates for the graded model*. Paper presented at the Annual Meeting of the Council on Measurement in Education, San Francisco, CA.
- Atılgan, H., & Tezbaşaran, A. A. (2005). Genellenebilirlik kuramı alternatif karar çalışmaları ile senaryolar ve gerçek durumlar için elde edilen G ve Phi katsayılarının tutarlılığının incelenmesi [An investigation on consistency of G and Phi coefficients obtained by generalizability theory alternative decisions study for scenarios and actual cases]. *Eurasian Journal of Educational Research*, 18, 236-252.
- Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması*. Pegem.
- Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Çok kategorili parametrik ve parametrik olmayan madde tepki kuramı modellerinin karşılaştırılması [Comparison of polytomous parametric and nonparametric item response theory models]. *Journal of Measurement and Evaluation in Education and Psychology*, 8(4), 354-372.
- Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.
- Büyüköztürk, Ş., Çakmak, E.K., Akgün, Ö.E., Karadeniz, Ş. ve Demirel, F. (2012). *Bilimsel Araştırma Yöntemleri*. Pegem.
- Choi, J., & Wilson, M. R. (2018). Modeling rater effects using a combination of generalizability theory and IRT. *Psychological Test and Assessment Modeling*, 60(1), 53-80.
- Crocker, L., & Algina J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich Inc.
- Çakıcı Eser, D., & Gelbal, S. (2013). Genellenebilirlik kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlılığın karşılaştırılması [Comparison of interrater agreement calculated with generalizability theory and logistic regression]. *Kastamonu Education Journal*, 21(2), 421-438.
- Çakır, M., & Aldemir, B. (2011). İki aşamalı genetik kavramlar tanı testi geliştirme ve geçerlik çalışması [Developing and validating a two tier mendel genetics diagnostic test]. *Mustafa Kemal University Journal of Social Sciences Institute*, 8(16), 335-353.
- Deliceoğlu, G., & Çıkrıkçı Demirtaşlı, N. (2012). Futbol yetilerine ilişkin dereceleme ölçeğinin güvenilirliğinin genellenebilirlik kuramına ve klasik test kuramına dayalı olarak karşılaştırılması [The comparison of the reliability of the soccer abilities' rating scale based on the classical test theory and generalizability theory]. *Hacettepe Journal of Sport Sciences*, 23(1), 1-12.
- Demir, B. P. (2016). Vee diyagramından elde edilen puanların güvenilirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi [The examination of reliability of vee diagrams according to classical test theory and generalizability theory]. *Journal of Measurement and Evaluation in Education and Psychology*, 7(2), 419-431.
- Demirel, Ö., & Epçaçan, C. (2012). Okuduğunu anlama stratejilerinin bilişsel ve duyuşsal öğrenme ürünlerine etkisi [Effects of reading comprehension strategies on cognitive and affective learning outcomes]. *Kalem International Journal of Education and Human Sciences*, 2(1), 71-106.
- Doğan, C. D., & Anadol, H. Ö. (2017). Genellenebilirlik kuramında tümüyle çaprazlanmış ve maddelerin puanlayıcılara yuvalandığı desenlerin karşılaştırılması [Comparing fully

- crossed and nested designs where items nested in raters in generalizability theory]. *Kastamonu Education Journal*, 25(1), 361-372.
- Doğan, N., & Bıkmaz Bilgen, Ö. (2017). Using generalizability theory in reliability estimation of measurements of higher-order cognitive skills. *The Journal of Academic Social Science*, 44, 1-9.
- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2015). *How to design and evaluate research in education*. McGraw Hill Education.
- Giray, M. D., & Sahin, D. N. (2012). Algılanan örgütsel, yönetici ve çalışma arkadaşları desteği ölçekleri: Geçerlik ve güvenilirlik çalışması [Perceived organizational, supervisor and co-worker support scales: A study for validity and reliability]. *Turkish Psychological Articles*, 15(30), 1-9.
- Güler, M., & Yetim, Ü. (2008). Ebeveyn rolüne ilişkin kendilik algısı ölçeği: Geçerlik ve güvenilirlik çalışması [Self-perception of parental role (SPPR) scale: Validity and reliability study]. *Turkish Psychological Articles*, 11(22), 34-43.
- Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenilirliğin karşılaştırılması [The comparison of reliability according to generalizability theory and classical test theory on random data]. *Education and Science*, 36(162), 225-234.
- Güler, N., Eroğlu, Y., & Akbaba, S. (2014). Genellenebilirlik kuramına göre ölçüt bağımlı ölçme araçlarında güvenilirlik: Yemek yeme becerileri örneğinde bir uygulama [Reliability of criterion-dependent measurement tools according to generalizability theory: Application in the case of eating skills]. *Abant İzzet Baysal University Journal of Faculty of Education*, 14(2), 217-232.
- Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Genellenebilirlik kuramı [Generalizability theory]*. Pegem.
- Gülle, A., Uzun, N. B., & Akay, C. (2018). Ortaokul öğrencilerine yönelik blok flüt icra performansı dereceli puanlama anahtarının güvenilirliğinin genellenebilirlik kuramı ile incelenmesi [The study on the reliability of the grading key measuring the performance of the block flute performance of the secondary school students via generalizability theory]. *Elementary Education Online*, 17(3), 1463-1475.
- Hathcoat, J. D., & Penn, J. D. (2012). Generalizability of student writing across multiple tasks: A challenge for authentic assessment. *Research & Practice in Assessment*, 7, 16-28.
- Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56-64.
- Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, 6, 249-260.
- Kamış, Ö., & Doğan, C. D. (2017). Genellenebilirlik kuramında gerçekleştirilen karar çalışmaları ne kadar kararlı? [How consistent are decision studies in g theory?]. *Gazi University Journal of Gazi Educational Faculty*, 37(2), 591-610.
- Kaplan, A., & Duran, M. (2016). Ortaokul öğrencilerine yönelik matematiksel üstbilgi farkındalık ölçeği: Geçerlik ve güvenilirlik çalışması [Mathematical metacognition awareness inventory towards middle school students: Validity and reliability study]. *Journal of Kazım Karabekir Education Faculty*, 32, 1-17.
- Karasar, N. (2016). *Bilimsel Araştırma Yöntemi*. Nobel.
- Karlsson, J. (2017). *Generalizability theory and a scale measuring emotion knowledge in preschool children* [Master's thesis, Stockholm University]. <http://www.diva-portal.org/smash/get/diva2:1065849/FULLTEXT01.pdf>

- Katrancı, M., & Yangın, B. (2012). Üstbiliş stratejileri öğretiminin dinlediğini anlama becerisine ve dinlemeye yönelik tutuma etkisi [Effects of teaching metacognition strategies to listening comprehension skills and attitude toward listening]. *Adiyaman University Journal of Social Sciences*, 2013(11), 733-771.
- Kaya, A. (2005). Çocuklar için yalnızlık ölçeğinin Türkçe formunun geçerlik ve güvenilirlik çalışması [The validity and reliability study of the Turkish version of the children's loneliness scale]. *Eurasian Journal of Educational Research*, 19, 220-237.
- Kenny, D.A. (1987). *Statistics for the social and behavioral science*. Little, Brown.
- Nitko, A. (2004). *Educational assessments of students*. Pearson.
- Scherbaum, C., Dickson, M., Larson, E., Bellenger, B., Yusko, K., & Goldstein, H. (2018). Creating test score bands for assessments involving ratings using a generalizability theory approach to reliability estimation. *Personnel Assessment and Decisions*, 4(1), 1-8. <https://doi.org/10.25035/pad.2018.001>
- Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation*, 19, 245-263. <https://doi.org/10.1080/13803611.2013.767632>
- Tavşancıl, E. (2005). *Tutumların ölçülmesi ve SPSS ile veri analizi [Measurement of attitudes and data analysis with SPSS]*. Nobel.
- Shavelson, J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.
- Yılmaz Nalbantoğlu, F., & Gelbal, S. (2011). İletişim becerileri istasyonu örneğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması [Comparison of different designs in accordance with the generalizability theory in communication skills example]. *Hacettepe University Journal of Education*, 41, 509-518.

Uncovering the Reasons of EFL Teachers' Unwillingness and Demotivation towards Being More Assessment Literate

Elcin Olmezer Ozturk ^{1,*}

¹Anadolu University, Faculty of Education, Department of Foreign Language Education, Eskişehir, Turkey

ARTICLE HISTORY

Received: Feb. 08, 2021

Revised: Apr. 18, 2021

Accepted: May 29, 2021

Keywords:

Assessment literacy,
Turkish EFL teachers,
Unwillingness,
Demotivation.

Abstract: The current study investigates the reasons of EFL teachers' unwillingness and demotivation towards being more assessment literate. 19 EFL teachers working in the preparatory programs of various state universities took part in the study, and the data were collected via semi-structured interviews. Those 19 teachers were deliberately chosen from 27 teachers based on their negative utterances towards being more assessment literate in relation to the aim of the current study. The data obtained from the utterances of the participants with respect to two interview questions were transcribed, coded and labelled according to the recurrent and common themes according to the qualitative content scheme of Creswell (2012). The findings revealed that why the participating teachers were unwilling and demotivated to be more assessment literate resulted from five factors; a) seeing language assessment as an extra burden, b) the presence of testing office and materials, c) language assessment as an anxiety-provoking factor, d) institutional factors and e) rarity of ways to improve oneself. Apart from shedding light on the unwillingness and demotivation of teachers to learn more about assessment, this study also comes up with implications for language teachers and research suggestions in relation to the findings of the study.

1. INTRODUCTION

1.1. Assessment Literacy and Language Assessment Literacy

Assessment literacy (AL) “is not an initiative, not just another fad or bandwagon to jump on or off, it is a foundational and essential competency for all school leaders, teachers and students” (McCafferty & Baudry, 2018). As is understood from the quotation, assessment literacy is seen as a “sine qua non for today’s competent educator” (Popham, 2009, p. 4), and it is not an extra feature to possess; rather, it is a basic component of education. The definitions of assessment literacy abound in the literature. Stiggins (1991), coining the term, defined assessment literacy as teachers’ skills in the use of assessment. Falsgarf (2005, p. 6) stated that it “is the ability to understand, analyze, and apply information on student performance to improve instruction”. Additionally, for Popham (2018), it is the understanding of basic and important concepts in assessment.

*CONTACT: Elcin Olmezer Ozturk ✉ elcinolmezerozturk@anadolu.edu.tr 📍 Anadolu University, Faculty of Education, Department of Foreign Language Education, Eskişehir, Turkey

e-ISSN: 2148-7456 /© IJATE 2021

Newly coined term rooted in assessment literacy is language assessment literacy (LAL). Though it has similar features with assessment literacy, as it is specifically on language, it also has different characteristics. Davies (2008), with a focus on language, stated that LAL consists of three parts that are knowledge, skills and principles. For Malone (2013, p. 329), language assessment literacy is “language teachers’ familiarity with testing definitions and the application of this knowledge to classroom practices in general and specifically to issues related to assessing language”. Lastly, Inbar-Lourie (2008, pp. 389-390) defined this term as “language assessment knowledge base comprises layers of assessment literacy skills combined with language specific competencies, forming a distinct entity that can be referred to as language assessment literacy”. As is seen, LAL and AL have similar features, both requiring a teacher having sound knowledge in assessment; yet, LAL also requires a language teacher to be knowledgeable in both assessment and language, and language-related assessment.

1.2. The Necessity of Assessment Literacy and Assessment Literate Teachers

Assessment should not be seen as a product or outcome only; rather, many strategies and processes helping learners become better learners and educators are involved in assessment (McCafferty & Baudry, 2018). Though each and every stakeholder in education is into assessment for various reasons, it is the teachers who have major roles in assessment. Language teachers have this role of assessment as a part of their professions (Mertler, 2003), and also Stiggins (1991) argued that teachers spend 50 % of their instructional time with assessment-related activities. What teachers have to know related to assessment varies such as reliability, validity, designing tasks, alternative assessment, scoring, and it is for sure that each and every teacher needs a dose of assessment literacy (Popham, 2011).

When assessment-literate teachers “make educational decisions based on appropriate assessment-elicited evidence, the resultant decisions almost always will be more defensible-meaning, more likely to improve students’ learning” (p. 2), and when good decisions are made, it means avoiding mistakes (Popham, 2018). Moreover, when more valid decisions are made, it is more possible to appeal to learners’ needs more and adapt instruction (Shepard, 2000). On the other hand, when teachers lack adequate knowledge related to assessment, they could make certain mistakes that could be grouped under three categories that are “using the wrong tests, misusing results of the right tests, and failing to improve instructionally useful tests” (Popham, 2018, p. 8). To avoid these kinds of mistakes, assessment literate teachers are needed in teaching and learning process because the power of assessment is rooted in the knowledge of teachers in assessment (Calderhead, 1996).

In spite of the importance of assessment literate teachers in instruction, teachers have limited competency in assessment (Popham, 2018), and teachers are not assessment literate (Alderson, 2005; Mertler and Campbell, 2005). Many teachers do not feel themselves ready for assessment-literate activities including both pre- and in-service teachers. Pre-service teachers stated that they did not expose to sufficient and qualified education in assessment, and many in-service teachers expressed that they are not adequately equipped with assessment knowledge (Plake, 1993). Stiggins (2010, p. 233) drew attention to this problem by stating that “assessment illiteracy abounds”.

1.3. Related Studies

Research into language assessment literacy “is still in its infancy” (Fulcher, 2012, p. 117); however, the number of the studies investigating assessment literacy and language assessment literacy is increasing day by day. While some studies focused on the needs of teachers as the conductors and designers of assessment-related tasks (Vogt & Tsagari, 2014; Volante & Fazio, 2007), some examined language assessment literacy of teachers (Ölmezer-Öztürk & Aydın, 2019; Volante & Fazio, 2007). Thus, various aspects of assessment literacy have been

investigated in several studies.

To start with, Volante and Fazio (2007) studied with 69 pre-service teachers, and their assessment knowledge and needs for assessment were investigated. Though the participants stated that they had taken an assessment course, they still needed to learn more about assessment. The findings also indicated their low level of confidence in assessment-related tasks. In a similar study with different stakeholders as participants, O'Loughlin (2013) examined university administrators' assessment needs since they were responsible for admission decisions. The administrators from two universities in Australia received a survey including questions related to IELTS use, evaluation, etc. The findings revealed that the administrators needed to be more assessment literate and educated for the valid and reliable interpretation of test scores. On the other hand, in Vogt and Tsagari (2014), 153 teachers from seven European countries were asked about their needs in LAL in three aspects that are classroom-focused language assessment, purposes of testing, and content and concepts of language assessment. The results demonstrated that the teachers were not competent enough in some areas such as self and peer assessment, portfolio assessment, reliability, validity and using statistics.

In addition to the studies focusing on LAL needs of teachers, some others investigated the assessment literacy levels of teachers or their perceptions of it. For instance, Lam (2015) focused on the overall language assessment training in five Hong Kong institutions, and pre-service teachers' perceptions of their LAL development. The findings showed that there was insufficient support to foster LAL, and the training for LAL was inadequate based on the perceptions of the participants. Similarly, Baker and Riches (2017) aimed to examine whether a series of workshops contributed to LAL development of 120 Haitian high school teachers. The data were collected via feedback on drafts of revised exams, survey with teachers, and teacher interviews and the results demonstrated that LAL development of the teachers was clear after these workshops, and their LAL levels increased in creating reading comprehension questions, in learning about reliability, validity, and practicality, and increased attention of the connection between teaching and assessment. In another study conducted in Turkish EFL context, Ölmezer-Öztürk and Aydın (2019) investigated language assessment knowledge of 542 language teachers working in higher education by using a scale they developed. The findings revealed that the participant teachers were not assessment literate, and the teachers were the most knowledgeable in assessing reading whereas they had the lowest score in assessing listening.

1.4. The Present Study

As assessment literate teachers play crucial roles in the efficacy and appropriacy of assessment-related activities, the importance of having assessment literate teachers in education is stressed in the literature (Alderson, 2005; Leung, 2014; Malone, 2013; Popham, 2006). Yet, the studies in the literature demonstrated that both pre- and in-service teachers do not feel themselves competent enough and they are not self-confident and knowledgeable in assessment-related activities due to their lack of assessment literacy (Hatipoğlu, 2015; Tsagari & Vogt, 2017). Though few in number, there exist certain studies focusing on the needs of EFL teachers in relation to language assessment (Fulcher, 2012; Inbar-lourie, 2008; Malone, 2013; Volante & Fazio, 2007), language assessment literacy levels of EFL teachers (Tao, 2014), the effectiveness of trainings on their language assessment literacy levels (Campbell, Murphy, & Holt, 2002; Mertler, 2009), and language assessment knowledge of EFL teachers (Ölmezer-Öztürk & Aydın, 2019; Şahin & Hatipoğlu, 2019). As a common point in these studies, there is a special emphasis on the notion that language teachers lack a certain level of language assessment literacy and they need some training on it. However, they do not present the background and reasons for this problem. In other words, the studies in the literature basically describe the

situation by showing how assessment illiterate EFL teachers are and what they need to become more assessment literate. Besides, compared to the number of aforementioned studies, there is a paucity of research focusing on the reasons why language teachers are or feel themselves incompetent in language assessment. Even though, as demonstrated, many in-service teachers are assessment illiterate, why many teachers do not take action and are not willing and motivated enough to be more knowledgeable in language assessment have not been the foci of any studies so far to the best knowledge of the researcher. Examining the underlying reasons of their unwillingness and demotivation with respect to language assessment is of primary concern, because when the underlying reasons of their unwillingness and demotivation have been uncovered, then better conditions and opportunities could be provided for the teachers to be more assessment literate. Within that scope and purpose, the following research question is asked throughout the study.

What are the underlying reasons of EFL teachers' unwillingness and demotivation to being more assessment literate?

2. METHOD

2.1. Research Design

Aiming to identify the main reasons behind EFL teachers' unwillingness and demotivation to be more assessment literate, the current study employs a basic qualitative research perspective which is "concerned with subjective opinions, experiences and feelings of individuals and thus the explicit goal of research is to explore participants' views on the phenomena being studied" (Dörnyei, 2007, p. 38). Since the major focus of the study is to uncover opinions and experiences of participating teachers on the research matter, their language assessment literacy, a qualitative perspective is followed throughout the study.

2.2. Research Context

The research context is the preparatory programs of state universities in Turkey. Language teachers are responsible for teaching English to learners in this program, and while some programs offer English courses in an integrated way, some divide the courses into skills such as reading, writing, speaking and listening. There exist certain offices in these programs such as testing office, material development office, and curriculum office. Teachers take part in these offices either willingly or upon the will of their managers. Testing office members have various duties, and what they are responsible for may differ based on the institutions since there is not a determined program or schedule for testing office members of the institutions in Turkey. Owing to this, it is usual to come across different and various performances of the institutions. Moreover, to exchange ideas and determine assessment-related tasks, testing office members gather and decide on certain issues related to assessment such as the type of exams, the items to be asked in the exams, scoring, etc.

2.3. Participants

The participants include the teachers working at preparatory programs of nine different state universities. 19 teachers in these programs are the participants of this study, and none of them is a member of the testing office in their institutions. Convenient sampling was preferred for this study. At the very beginning, 27 teachers were sent a question asking for whether they had a positive attitude towards being more assessment literate and whether they were making efforts for this. Eight of them stated that they were eager to learn more about language assessment and trying hard to be more assessment literate teachers. As these eight teachers had a positive attitude towards language assessment, and the focus of the study is to find out the reasons of negative attitudes towards it, they were excluded from the actual study. Based on their negative stance, 19 teachers were interviewed by either skype or face-to-face semi-structured questions.

Out of these 19 teachers, 11 were females and eight were males. Besides, their ages ranged from 28 to 47, and their years of experience in teaching varied as well. Their educational background was also various, and they were the graduates of English Language Teaching Department and English Language and Literature. Finally, different universities were preferred so as to hinder the possible problems that may come out because of the contextual factors.

2.4. Data Collection Process

In semi-structured interviews, the teachers were asked two questions which were prepared by the researcher beforehand in the scope of the study. The questions were checked by two colleagues for clarity and wording. Moreover, two academicians in ELT were asked for their opinions regarding the content of the items and whether they served their purposes or not. The questions were in Turkish to be able to get more and richer answers from the participants and help them feel themselves more relaxed while answering without the interference of the target language. The questions asked in the interview are as [Table 1](#):

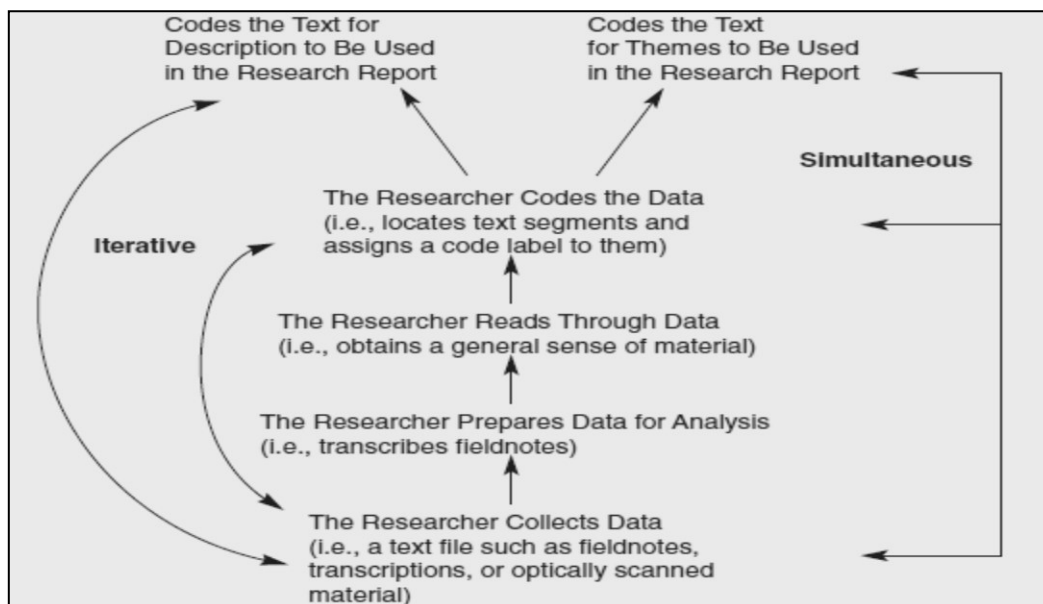
Table 1. *The interview protocol.*

QUESTIONS FOR THE INTERVIEW
1. In the first phase of the study, you stated that you had negative attitudes towards language assessment, and you were not making any/many efforts towards being more assessment literate. What are the reasons of your negative attitude towards language assessment?
2. Do you have any negative memories or experiences related to language assessment? What is this/What are these?

2.5. Data Analysis

The qualitative content analysis scheme of Creswell (2012) was used to analyze the data obtained from 19 participant teachers. All the answers of the participants were transcribed first, and then grouped into codes based on their common and recurrent ideas in the transcriptions. Rooted in these codes, certain themes came out, and these themes were presented in frequencies. Data analysis process is highlighted in the following [Figure 1](#):

Figure 1. *Content analysis scheme.*



3. FINDINGS

The research question of the current study aimed at finding out the reasons of EFL teachers' unwillingness and demotivation to being more assessment literate. The following [Table 2](#) demonstrates the themes and codes that came out based on the investigation of the reasons of their unwillingness and demotivation with respect to the frequencies.

Table 2. Themes and codes derived from the participants' answers.

THEMES	CODES
LA as an extra burden	Requiring extra efforts and time (x13) Not compulsory (x9) Not heavily focused in pre-service education (x8)
The presence of testing office and materials	Testing office's duty, not mine (x14) The presence of ready-made materials (x9)
LA as an anxiety-provoking factor	Not feeling self-confident (x10) Too much terminology (x6) Requiring statistical knowledge (x5)
Institutional factors	Their colleagues and students' harsh criticisms (x14) Absence of support and no appreciation (x10) Rarity of teachers who are role-models and competent in LA (x7) Having the same responsibility with everyone (x5) Objection to changes and novelty in exams (x4)
Rarity of ways to improve oneself	Books focusing on assessment in general (x6) Not enough conferences specifically focusing on assessing language skills (x5)

After the analysis of the data, many codes were identified, and as the next step, these codes were grouped under common themes. The analysis of the data revealed five themes based on the answers of the participants that are language assessment as an extra burden, the presence of testing office and materials, language assessment as an anxiety-provoking factor, institutional factors and rarity of ways to improve oneself. To start with the first theme, the teachers perceived language assessment as not a part of their teaching, but an extra duty or qualification. 13 teachers expressed that language assessment is a demanding field which requires many efforts and hours and days of studying to be competent in.

T7 stated that,

“Assessment is a broad field, and there are many sub-topics of it. To be more assessment literate, I have to study a lot-though I am studying for my courses-, and make many efforts for these.”

T16 expressed that,

“Our duty goes on outside the school as well. When I get home, I check my students' assignments most of the time and give written feedback to their works. Even though these could be regarded as a part of the assessment, I am just giving feedback to my learners. As I do not have much time for my professional development, I cannot find any time to go through the literature on language assessment and learn more.”

In addition to these, T3 mentioned that,

“Whenever I have time, I do my best to improve myself as a teacher such as discovering books on teaching and interesting and motivating activities for students, but not related to assessment. I feel teaching is the primary job of me, and I can survive with the knowledge I have related to assessment.”

One of the participant's utterance (T4) made it clear that she perceived teaching and assessment as too different concepts and not interconnected. She stated that,

“My primary job is to teach, not to assess. I am already busy with teaching, and I have more than 20 hours for teaching per week. I have to prepare materials for the courses, cover the books and select the most appropriate and motivating ones for my students. At the same time I have to keep up with the curriculum. These all take time, and under such conditions, I have no time to be more knowledgeable in assessment.”

Along with requiring many efforts and time, nine teachers added that to be competent in assessment or to be assessment literate is not a must for teachers. T3 mentioned that “he could survive with the knowledge he has related to assessment”, apart from him, T15 stated that,

“What I know is sufficient for me. I am not designing any exams, and what I am expected to do as a teacher is just to check my students' assignments- which is mostly related to grammar and organization- and give feedback to them. Thus, I do not feel the necessity to be better in language assessment.”

T2 voiced that,

“Before we, as teachers, get these positions as teachers at university, some of us are asked theoretical questions about classroom management, students and teaching methodology. Some are asked questions about how to teach an example grammatical unit, some are asked the differences between certain confusing grammatical rules, etc. Yet, I have not heard of a teacher who has been asked any questions about language assessment, how to assess learners best, how to score, or how to design assessment-related tasks. That is, while I was studying for the exam to get my position in my institution, I covered many books related to teaching and learning, but not even a book on assessment.”

Furthermore, T8 added that,

“To be a testing office member does not mean that you are good at assessing learners or you are very knowledgeable in this field. If I am willing to take part in testing office, I then could be a member of it. Also, there is no prerequisite knowledge to have a duty in this office. What I intend to say is that even if when you are having roles in your institution as assessors, your background knowledge in assessment is not important most of the time. For me-non-testing members-, it is naturally not a must as well because I am not having roles in exams.”

Finally, eight teachers expressed that language assessment is ignored in pre-service education as well, if it is of primary importance, it should be given more emphasis throughout pre-service education. In relation to this, T12 stated that,

“For instance, in practicum in my university years, we taught English to learners, we designed materials, we tried to do our best for classroom and time management, but we did nothing related to language assessment. I did not see any sample exams, and I had no idea how the students were assessed.”

Another teacher (T7) expressed that,

“In pre-service education, we had three different methodology courses that were how to teach grammar, how to teach speaking and writing and how to teach listening and reading, but we had only one course in assessment which we took in our fourth year. Fourth year was too late to learn about language assessment, and we were very busy with preparing lesson plans in practicum; thus, that course was not very beneficial for us.”

Opposed to the participants taking language assessment course in their pre-service education, T9 voiced that,

“In pre-service education, we did not have a standalone language assessment course, not even assessment in general. LA was not given enough importance in pre-service education; so, my background is not good enough in relation to language assessment. I do not have any intrinsic motivation to learn more for a subject which is neglected in pre-service education as well.”

Second theme is the presence of testing office and materials. 14 teachers mentioned that assessment-related activities are carried out by testing office, and they are responsible for language assessment. Besides, nine teachers stated that there are ready-made tests and questions that could be used for assessment purposes. Related to these, one participant (T1) uttered that,

“Every teacher has a duty in the institution, and the ones in testing office are responsible for assessment. It is their job, and as I am not a member of testing office, I do not need to be more proficient and knowledgeable in language assessment. Yet, the ones in testing office have to do this.”

Another teacher (T10) verbalized that,

“Testing office members gather and hold long-lasting meetings, they also design questions and negotiate them. If I were in testing office, then I would feel the pressure on my shoulders to search and learn more about language assessment. To do this, I would look for the books and exchange ideas with my friends working in other preparatory programs in relation to their practices. But, now as testing office members prepare everything for me, and what I have to do is to invigilate while students are seated in an exam.”

One more example is related to the ready-made materials, and T5 stated that,

“There exist ready-made questions related to each skill, and these questions are given to the teachers together with their teacher books. These questions are designed by knowledgeable people and they spend a lot of time, and they go through many stages. Thus, is there a real need for designing questions again and again? I guess not.”

Third theme is language assessment as an anxiety-provoking factor. 10 teachers expressed that they do not feel confident enough in language assessment. Six participants thought that there is too much terminology in language assessment, and partly complained about them. Finally, five of the teachers uttered that this field requires statistical knowledge, and a teacher has to be competent in statistics as well. T3 confessed that,

“I feel myself very competent in teaching-related subjects; however, when it comes to language assessment, I get stuck. What I know is not enough to regard myself as an assessment literate teacher. Since I am not self-confident enough, I get more anxious when I have to engage with assessment-related tasks. I cannot even concentrate on what I am doing. So, it is like a chain.”

T2 mentioned that,

“Assessment is a field with too many diverse views; thus, one cannot say an assessment-related task should be done in a certain way most of the time. There are pros and cons of many issues, and due to this situation, I cannot assure myself by saying that what I am doing is totally true. These diverse ideas lead me to be insecure about what I know which directly results in my lack of confidence in language assessment.”

Another teacher (T1) complained about the existence of too many terms stating that,

“Indeed, I am familiar with some kinds of tasks in assessment. But, when I have a look at the books, I come across their names-in other words, terms. Though I may be making use of certain things, I am not very good at remembering their names. Hence, I try to memorize the terms that are too many to memorize, by the way. This memorization process drives me crazy, and I get really stressed.”

T13 voiced that,

“Whenever I open the first page of any books on language assessment, I see the pages loaded with too many terms such as reliability, validity, and their types, etc. They are crucial as well, but seeing all the terms one after another makes me scared, and also anxious.”

T15 also said that,

“Assessment goes hand in hand with statistical knowledge. You have to make calculations, and to be able to do so, you have to have some background statistical knowledge. When I am busy with all the numbers, it is like mathematics and I cannot get the joy of assessment. For instance, what I want to do is only to design questions. I do not want to calculate mean, median, etc. I feel

as if it was not my business. But, you cannot just design questions without including statistical calculations.”

Next theme is institutional factors that were mentioned by many teachers during the interviews. A lot of participants gave some reasons for why they did not want to be more assessment literate, and the existence of certain negative feelings and situations about language assessment were found to be related to their institutions. 14 teachers stated that their colleagues’ and students’ harsh criticisms were the reasons for why they were not very willing and motivated to be more assessment literate. 10 participants expressed that there is no support and appreciation for the teachers who are into language assessment. Furthermore, seven teachers mentioned that the number of teachers who are role-models and competent in language assessment is not enough. Five of them complained about the fact that the ones who have assessment-related duties have the same responsibility with the ones who have no extra office duties. The last one is four participants told that there is an objection to changes and novelty in the designation of exams. Some quotations related to the aforementioned codes are as follows:

To start with, T4 stated that,

“I observe that people are so cruel to the teachers who are in testing office. Students always complain about the quality of the questions and they keep saying that some of the questions are false or do not have the right answers in the options. Let alone the students, teachers in my institution always find a way to imply testing office members that the topic in writing part is not very good or the reading passage is full of unknown words or too easy for students to give correct answers. Whatever they do, people find a way of complaining about the work they have done.”

In parallel with the previous quotation, T6 expressed that,

“One of my friends is a member of testing office. She once told me that she did not even want to go to the canteen to get some tea in the break, because the teachers who came across her in the canteen complained about the questions all the time. She also stated that she gave up drinking tea because of those kinds of teachers murmuring a lot.”

T10 shared a memory as well about these criticisms:

“My roommate was in testing office. One day, just after the exam, a teacher rushed in our room and said that the total of the points did not make a hundred in total with a high pitch of voice. My roommate was trying to be calm saying that we all checked these things again and again, let me check it once more. Then, my roommate counted the points and the total was a hundred. The reaction of the teacher was only “Oh, I miscounted then!”. This example was a good indicator of how other teachers in the institution were unfriendly and intolerant to the teachers in testing office rather than appreciating them.”

In relation to the second code, T1 uttered that,

“As far as I can see and observe, I can say that there is too much to do for testing office members, and nobody helps them just because they are not in this office officially. Some periods are full of work for them, for instance at the beginning and end of the term; but, no support is given to them by the management and the other teachers as well. They are all alone”

T3 mentioned that,

“I do not want to be engaged with language assessment since if the others hear about my interest, then I will be for sure chosen for testing office which is not very good for me. The reason is nobody appreciates what testing office members do, and they all ignore their efforts such as long meetings, negotiations, editing processes, etc.”

The next code is the rarity of role-model teachers, and in relation to this, T11 uttered that,

“The problem is that the teachers in testing office are not more competent and knowledgeable than us, except for one or two teachers. The fact that a teacher has a testing office duty does not mean that I can consult that teacher when I have a question about language assessment.”

About the responsibilities, T14 stated that,

“Testing office members and non-testing members have the same hours of teaching per week. In other words, I teach and then I leave the school; however, they have also testing duties apart from teaching. They have to teach the same hours as me which is surprising. Thus, in our institution, being a testing office member is not given enough value. It would be better if testing office members are given some incentives such as decreasing their workload, or sending them to conferences, etc.

T8 working at a different institution from the previous participant said that,

“In our institution, testing office members teach less hours compared to non-testing teachers. At first, it may sound good; but, they teach two hours less which does not decrease their workload at all during the week.”

Final code is teachers’ not welcoming changes and novelty. With respect to this, T9 mentioned that,

“Once, I was in testing office and read a lot about language assessment; as a result, I learnt many things and saw that some of our practices were not okay. When I voiced this in one of the meetings, all of the members objected to my idea stating that they had already ready-made exams which were controlled and corrected many times up to that time, and there was no need for the things to be mixed up. I was shocked, and I had tons of memories like this unfortunately. Then, I gave up reading more about language assessment, because learning more did not contribute positively to the practices in my institution. If I cannot make use of the knowledge I have regarding language assessment, and cannot implement them in my institution, what is the use of being more assessment literate?”

In parallel to what T9 uttered, T14 said that,

“Once I was reading a book, I read something about language assessment, and I shared it with testing office members believing that they would appreciate it and would make use of that practice in the following exams. Unfortunately, they told me that they had some fixed exams, and they did not want to make big changes on them. Moreover, they expressed that their practice was totally true though I showed them the related parts in the book.”

The last theme is the rarity of ways to improve oneself. Six teachers focused on the absence of books specifically designed for language assessment, and five teachers mentioned that there exist not enough conferences which they can attend and learn to the point practices related to language assessment. Here are some quotations of the participants:

T12 expressed that,

“There are many books in the literature related to assessment; but, when it comes to language assessment and assessing skills separately, there are very few books. They all refer to the first published books, as well-thus, including nearly the same information.”

Another teacher (T6) stated that,

“All the books start with very general terms, and give some statistical information. Thus, it is not very easy to find a book that solely focuses on language assessment and the common practices that will help teachers use in their classrooms and exams. There should be more books covering practical uses of assessment.”

Similar to the books, T8 complained about the conferences by saying that,

“Most of the conferences have many sub-topics, and one of them is assessment. Thus, it is not very easy to find a conference in which there are many speakers who are expert in language assessment and deliver a speech on language assessment that is full of practical issues. Rather, there are some conferences on assessment in general, but the topics are too technical and specific; hence, it does not make any sense whether you attend those or not.”

4. DISCUSSION

The current study shed light on this issue by uncovering the factors leading language teachers to be reluctant and resistant to language assessment literacy. The participants were 19 teachers

working at preparatory programs of state universities, and they did not have duties in testing offices in their institutions. Out of 27 teachers who were sent a question asking for whether they had a positive attitude towards being more assessment literate or not, 19 of them stated that they did not feel very eager to learn more about language assessment, and these participants provided the data for the current study. They were asked three questions, and their answers to these questions were transcribed and code-labeled by the researcher and also a colleague with a Phd in ELT. The data revealed that why the teachers had some resistance to be more assessment literate stems from five main issues that are language assessment as an extra burden, the presence of testing office and materials, language assessment as an anxiety-provoking factor, institutional factors and rarity of ways to improve oneself.

The themes derived from the obtained data were language assessment as an extra burden, presence of testing office and ready-made materials, language assessment as an anxiety-provoking factor, institutional factors and rarity of ways to improve oneself in regard to language assessment. To begin with the first theme, the participants stated that language assessment requires extra efforts and time, it is not compulsory to be more assessment literate, and it is not heavily focused in pre-service education. With respect to language assessment's requiring extra efforts and time and being not compulsory, Purpura (2016, p. 191) stated "rather than seeing assessment as an organic part of applied linguistics, L2 assessment is still often viewed as an afterthought, or as a craft". Besides, Stiggins (1995) drew attention to the fact that language assessment cannot be regarded as an extra thing for teachers since it is an inevitable part of their jobs. What is more, Popham (2006, p. 85) touched upon the necessity and importance of language assessment by saying that "Today, more than ever, assessment plays a pivotal role in the education of the students. That's why educators – and everyone else who has an interest in education- need a dose of assessment literacy". As is seen, assessment literacy is not only necessary for teachers who have an interest in assessment, but also anyone who is an educator. Upon this importance, Mertler (2002) stated that all duties of teachers are important, and should be given great care; but, the most important duty of a teacher is the assessment. Katz (2012) also warned all language teachers that language assessment should not be seen external to teaching and learning; rather, language assessment and instruction cannot be separated and they have to go hand in hand for an effective instructional process (Malone, 2013). As is understood, although the literature presents assessment and learning as a strongly connected process that should go hand in hand, the opinions of the participants show that teachers may perceive assessment as an extra work or duty not closely related with teaching and learning process. This perception of "assessment as an extra burden", which might be originating from several reasons such the lack of supervision at the institutions, turns to be a leading factor in teachers' unwillingness to be more assessment literate.

In relation to the pre-service education, in parallel with these findings, Mertler (2005) and Stiggins (1991, 1995) also stated that education policies are not assuring that the pre-service teachers get adequate training in language assessment before they start their professions. This subject field is still ignored in professional development programs. Furthermore, the participants expressed that they had either no separate language assessment course or just one course covering all the assessment of skills and assessment in general superficially. Schafer (1993) also maintained that half of the teacher education programs do not have a standalone course in language assessment, and added that the ones having the course does not give enough importance to the assessment of each and every skill. For instance, not enough emphasis is given to the teachers for the development of their language assessment literacy in North America (Coombe, Troudi, & Al-Hamly, 2012), and in the similar vein, in our context, Turkey, professional development of teachers is seen as more valuable and given importance day by day, but not in terms of language assessment. Professional development programs are becoming more popular and common day by day, but they include the development of teachers

in terms of their teaching skills. Assessment is still not a part of these programs. This issue was raised by many researchers coming up with the same conclusion that pre-service education should not be restricted to only one course in pre-service education (Hatipoğlu, 2015; Herrera & Macias, 2015; Ölmezer-Öztürk & Aydın, 2019; Popham, 2009). In their study, Mede and Atay (2017) found out that 62% of the participants had a separate language testing and evaluation course in pre-service education, but they found it very insufficient.

Next, in relation to the second theme that is the presence of testing office and ready-made materials, the teachers voiced that language assessment is testing office's duty, not theirs and there already exist ready-made materials to be used for language assessment. In the preparatory programs in Turkey, there are various offices and each of them has a different focus and duty. One of them is testing office, and only the teachers who are the members of testing office are responsible for designing tests and assessment-related tasks, and the teachers who are not the members of testing office are only responsible for invigilating. As is clear, if they do not have the intrinsic motivation to be more knowledgeable in language assessment, they do not need to learn more about assessment. All the things related to language assessment are prepared by testing office, and all the teachers are given the necessary information by this office again. Some teachers may find not having to design any questions easy, and no effort is made as well. In relation to this, Coombe, Troudi, and Al-Hamly (2012) stated that some teachers cannot keep up with the recent changes in the field of language assessment; thus, they just ignore their duties as assessors, and let the others do these duties. As is obvious from the utterance above, some teachers may not feel the necessity to be more knowledgeable, and find it easy when the works are done by more responsible ones.

To go on with the third theme which sees language assessment as an anxiety-provoking factor, it was mentioned in the data of the teachers that they were not feeling themselves self-confident enough in language assessment, there is too much terminology to be covered in this field, and it requires them to know statistical knowledge to be competent in it. This finding is in line with what Coombe, Troudi and Al-Hamly (2012) stated in relation to language assessment that teachers attach unpleasant feelings to language assessment. Moreover, Jacobs and Chase (1992) voiced that the teachers are not very happy while carrying out their assessment-related activities because of the fact that language assessment is seen as one of the unpleasant duties of teachers. In the same vein, Stiggings (1995) maintained that the most important barrier to assessment literacy by teachers is fear of assessment, and this fear is formed owing to the unpleasant experiences teachers had when they were students. Besides, Herrera and Macias (2015) stated that because most of the teachers do not like the assessment part of their jobs, they design test that are not very effective in terms of classroom assessment, and their fear of assessment leads them not to be able to be more assessment literate (Mertler, 2002). Another point leading teachers to unpleasant feelings is the existence of too much terminology and statistical knowledge. This issue was stated by McNamara and Roever (2006) who indeed was drawing attention to the trainings that are all full of applied psychometrics. Maybe owing to this reason, when teachers hear language assessment, one of the first things that comes to their minds is statistics. In other words, they relate language assessment to statistics.

Fourth theme is institutional factors in which the participants expressed that teachers engaging with language assessment in their institutions get harsh criticisms from their colleagues and students, there is no support for these teachers who make efforts to be more assessment literate, there is scarcity of teachers in their institutions who could be regarded as role-models for them and more knowledgeable than them, and finally these teachers trying to be interested in language assessment or testing office duties have the same responsibility with other teachers. What Coombe, Troudi, and Al-Hamly (2012) mentioned in their article is in parallel to the findings of this study, and they stated that some heads do not reduce the workload of teachers who deal with assessment-related tasks, and do not support these teachers. Banat (2018) also

stated similar ideas related to the institutions, and added that though common in institutions, assessment does not always rely on fair and valid basis, and owing to this reason, poor assessment in institutions does not always result from the assessment illiterate teachers, but the heads and institutional policies may be leading to inappropriate measures. The possible reasons of these negative attitudes could be the idea that language assessment is not seen as a must which each and every teacher should get involved in. Since both heads and teachers have similar ideas related to language assessment, they just tend to ignore the efforts of teachers who are actively involved in assessment-related tasks. In a similar vein, the workload of teachers is not reduced and they have the same teaching hours with other teachers, because what they do is not appreciated by others, and language assessment as a field is ignored.

The last one is the rarity of ways to improve oneself in terms of language assessment. The teachers said that there are not enough books which are specifically designed for assessing language skills and give practical information to teachers. Rather, they are mostly loaded with terminology and general information about language assessment. One more thing they stated is that the number of conferences specifically focusing on language assessment is not many in number, especially in our country. In the same vein, Coombe, Troudi, and Al-Hamly (2012) expressed that one of the barriers to assessment literacy is insufficient resources, and they suggested that to have more assessment literate teachers, online assessment resources should be available to all language teachers.

5. CONCLUSION

Assessment literacy is not an option or an extra qualification for today's language teachers in such a world where more and more scholars focus on the necessity and importance of assessment literacy for all teachers. As stated by Purpura (2016), it is not an extra craft, but the indispensable part of teachers' jobs. This study, along with touching upon this term, most specifically investigates the reasons of language teachers' resistance and unwillingness to being more assessment literate. The literature shows that there are many studies displaying that many in-service teachers do not feel themselves competent enough to carry out their assessment-related tasks, and these teachers do not feel ready for their professions (Mertler, 2003; Plake, 1993; Popham, 2006; Stiggins, 1991, 2010). One reason for this is seen as the insufficiency of pre-service education. However, learning and being more equipped with knowledge may stem from many factors, let alone pre-service education.

For language assessment taking extra efforts and time, the participants voiced that language assessment is not compulsory and a teacher does not have to have any skills or trainings to start the profession, and this field is not heavily focused in pre-service education. Another issue raised by the teachers was the presence of testing office and ready-made materials. Owing to these reasons, they thought that assessment-related tasks should be carried out by the teachers who are the members of testing office, and it is not their duty. Furthermore, there are already ready-made materials; thus, it was found awkward and unnecessary by some teachers to design assessment tasks again and again, and not using ready-made materials. Just using the ready-made materials as they is a way of assessing learners; hence, these teachers did not care about being capable of designing them. Moreover, the teachers stated that they feel anxious about language assessment because they do not feel self-confident enough, there is too much terminology and it requires statistical knowledge. Institutional factors were found to be a factor leading the teachers to be more assessment illiterate. They said that there are harsh criticisms by both teachers and students, there is no support or appreciation, there are not enough teachers who could be role-models with the help of their knowledge in language assessment, the testing office members have the same workload with other teachers which the participants found not fair, and changes and novel ideas in the exams are not very welcome by many teachers. The last issue was the rarity of conferences and books solely focusing on the language assessment

and assessment of each skill. These results yielded that why the teachers were not motivated enough to be more equipped with language assessment knowledge is a multi-faceted issue, and there exist many factors leading them to resistance and unwillingness.

5.1. Implications and Future Research

For the implications of the study, to start with, language assessment should be covered in at least two separate courses including practices as well in pre-service education. In practicum, pre-service teachers should be responsible for not only teaching and preparing lesson plans but also assessment parts of teaching. If language assessment is dealt with in detail, and the practice is included in practicum; then the graduates will feel themselves more self-confident in terms of language assessment, and also will have a better understanding about the necessity and importance of this field. Secondly, awareness raising activities that will help teachers gain the importance of assessment in learning and teaching should be organized in institutions, and after that teachers should have more opportunities to receive trainings and workshops on language assessment. The idea that language assessment is not an extra qualification for a language teacher should be transmitted to each and every teacher throughout these trainings and conferences. Last but not the least, the teachers who actively get involved with assessment-related activities should be encouraged, and they should be provided sufficient resources and their workload should be reduced.

In terms of research directions, first of all, a detailed and more extended investigation of EFL teachers' perspectives on language assessment literacy is needed. Identification of such a descriptive picture among EFL teachers that will explore their perspectives, opinions and need on language assessment will also help policy makers make better decisions on teacher training and language assessment policies. Besides, the short term and long term effectiveness of trainings and workshops organized by institutions should also be investigated for the betterment of such practices.

When it comes to the limitations of this study, firstly, the data revealed 19 teachers' ideas, feelings and attitudes regarding language assessment; thus, it should not be generalized to all language teachers. Moreover, the themes and codes derived from the data are restricted to the participant teachers' experiences and context-specific problems in their institutions. It would have been better if more teachers from various universities had participated this study. Finally, it would have been better if the views of all sides could have been investigated such as the heads, colleagues, and students. Then, it would have been a more comprehensive study looking into this phenomenon from various lenses.

Declaration of Conflicting Interests and Ethics

The author 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 author.

ORCID

Elcin Olmezer Ozturk  <https://orcid.org/0000-0001-7743-6361>

6. REFERENCES

- Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. Continuum.
- Baker, B. A., & Riches, C. (2017). The development of EFL examinations in Haiti: Collaboration and language assessment literacy development. *Language Testing*, 35(4), 557-581. <https://doi.org/10.1177/0265532217716732>
- Banat, H. (2018). Policy makers, assessment practices, and ethical dilution. In T. Ruecker, & D. Crusan. (Eds.). *The politics of English second language writing assessment in global contexts*. (pp. 58-65). Routledge

- Calderhead, J. (1996). Teachers: Beliefs and knowledge. In d. C. Berliner, & R. C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 709-725). Macmillan.
- Campbell, Y., Murphy, J. A., & Holt, J. K. (2002). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers*. Paper presented at the meeting of the Mid-Western Educational Research Association, Columbus, OH.
- Coombe, C., Troudi, S., & Al-Hamly, M. (2012). Foreign and second language teacher assessment literacy: Issues, challenges, and recommendations. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoyhoff. (Eds.). *The Cambridge guide to second language assessment*. (pp. 20-29). Cambridge University Press.
- Creswell, J. W. (2012). *Educational Research: Planning, conducting, and evaluating quantitative and qualitative research*. Pearson Education.
- Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, 25 (3), 327-347. <https://doi.org/10.1177/0265532208090156>
- Dornyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.
- Falsgraf, C. (2005, April). Why a national assessment summit? New visions in action. National Assessment Summit. Meeting conducted in Alexandria, Va. Retrieved from: http://www.nflrc.iastate.edu/nva/word_documents/assessment_2005/pdf/nsap_introduction.pdf
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9 (2), 113-132. <https://doi.org/10.1080/15434303.2011.642041>
- Hatipoğlu, Ç. (2015). English language testing and evaluation (ELTE) training in Turkey: Expectations and needs of pre-service English language teachers. *ELT Research Journal*, 4 (2), 111-128. <https://dergipark.org.tr/en/pub/eltrj/issue/28780/308006>
- Herrera, L. & Macias, D. (2015). A call for language assessment literacy in the education and development of teachers of English as a foreign language. *Colomb. Appl. Linguist. J.*, 17(2), 302-312. <http://dx.doi.org/10.14483/udistrital.jour.calj.2015.2.a09>
- Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, 25 (3), 385-402. <https://doi.org/10.1177/0265532208090158>
- Jacobs, L. C. & Chase, C. I (1992). *Developing and using tests effectively: A guide for faculty*. Jossey-Bass.
- Katz, A. (2012). Linking assesment with instructional aims and learning. *The Cambridge guide to second language assessment*. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoyhoff. (Eds.). *The Cambridge guide to second language assessment*. (pp. 66-73). Cambridge University Press.
- Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing*, 32(2), 169-197. <https://doi.org/10.1177/0265532214554321>.
- Leung, C. (2014). Classroom-based assessment issues for language teacher education. In A. J. Kunnan (Ed.), *The Companion to Language Assessment*, (pp. 1510-1519). Wiley Blackwell.
- Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing*, 30 (3), 329-344. <https://doi.org/10.1177/0265532213480129>
- McCafferty, A. S., & Beaudry, J. S. (2018). *Teaching strategies that create assessment-literate learners*. Corwin.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell.
- Mede, E., & Atay, D. (2017). English Language Teachers' assessment literacy: The Turkish context. *Dil Dergisi*, 168 (1), 1-5. <https://dergipark.org.tr/tr/download/article-file/780021>

- Mertler, A. C. (2003). Secondary teachers' assessment literacy: *Does classroom experience make a difference?*. *American Secondary Education*, 33(1), 49-64. <https://www.jstor.org/stable/pdf/41064623.pdf>
- Mertler, C. A., & Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the assessment literacy inventory*. Paper presented at the annual meeting of the American Research Association, Montreal, Quebec, Canada. <https://eric.ed.gov/?id=ED490355>.
- O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, 30 (3), 363-380. <https://doi.org/10.1177/0265532213480336>
- Ölmezer-Öztürk, E., & Aydın, B. (2019). Investigating language assessment knowledge of EFL teachers. *Hacettepe University Journal of Education*, 34(3), 602 -620. <https://doi.org/10.16986/HUJE.2018043465>
- Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher*, 6 (1), 21-27.
- Popham, W. J. (2006). All about accountability / Needed: A dose of assessment literacy. *Educational Leadership*, 63(6), 84-85. <http://www.ascd.org/publications/educationalleadership/mar06/vol63/num06/Needed@-A-Dose-of-Assessment-Literacy.aspx>
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, 48, 4-11. <https://doi.org/10.1080/00405840802577536>
- Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, 46 (4), 265-273. <https://doi.org/10.1080/08878730.2011.605048>.
- Popham, W. J. (2018). *Assessment literacy for teachers in a hurry*. Alexandria, VA: Association for Supervision and curriculum Development.
- Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal*, <https://doi.org/10.1111/modl.12308>
- Schafer, W. D. (1993). Assessment literacy for teachers. *Theory into Practice*, 32 (2), 118-126. https://www.jstor.org/stable/1476329?seq=1#metadata_info_tab_contents
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher*, 29 (7), 4-14. <https://doi.org/10.3102/0013189X029007004>
- Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, 72, 534-539. <https://www.jstor.org/stable/20404455>
- Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, 77 (3), 238-245. <https://www.jstor.org/stable/20405538>
- Stiggins, R. J. (2010). Essential formative assessment competencies for teachers and school leaders. In H. L. Andrade, G. J. Cizek (Eds.), *Handbook of formative assessment*, (pp. 233-250). Taylor & Francis.
- Şahin, S. & Hatipoğlu, Ç. (2019). *An analysis of English language testing and evaluation course in English language teacher education programs in Turkey: Developing language assessment literacy of pre-service EFL teachers*. (Unpublished PhD Dissertation). Middle East Technical University.
- Tao, N. (2014). *Development and validation of classroom assessment literacy scales: English as a Foreign Language (EFL) teachers in a Cambodian Higher Education Setting*. (PhD thesis). Victoria University.
- Tsagari, D. & Vogt, K. (2017). Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects. *Papers in Language Testing and Assessment*, 6(1), 41-64. http://www.altanz.org/uploads/5/9/0/8/5908292/5.si3tsagarivogt_final_formatted_proofed.pdf
- Vogt, K. & Tsagari, D. (2014) Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, 11 (4), 374-402. <https://doi.org/10.1080/15434303.2014.960046>.

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and Professional development. *Canadian Journal of Education*, 30 (3), 749-770. <https://journals.sfu.ca/cje/index.php/cje-rce/article/view/2973>

The Unit Testlet Dilemma: PISA Sample

Cansu Ayan ^{1,*}, Fulya Baris Pekmezci ²

¹Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Educational Measurement and Evaluation, Ankara, Turkey

²Yozgat Bozok University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, Yozgat, Turkey

ARTICLE HISTORY

Received: Sep. 24, 2020

Revised: Mar. 18, 2021

Accepted: May 25, 2021

Keywords:

PISA,

Testlet items,

Local Dependence,

Marginal item parameters.

Abstract: Testlets have advantages such as making it possible to measure higher-order thinking skills and saving time, which are accepted in the literature. For this reason, they have often been preferred in many implementations from in-class assessments to large-scale assessments. Because of increased usage of testlets, the following questions are controversial topics to be studied: “Is it enough for the items to share a common stem to be assumed as a testlet?” “Which estimation method should be preferred in implementation containing this type of items?” “Is there an alternative estimation method for PISA implementation which consists of this type of items?” In addition to these, which statistical model to use for the estimations of the items, since they violate the local independence assumption has become a popular topic of discussion. In light of these discussions this study aimed to clarify the unit-testlet ambiguity with various item response theory models when testlets consist of a mixed item type (dichotomous and polytomous) for the science and math tests of the PISA 2018. When the findings were examined, it was seen that while the bifactor model fits the data best, the uni-dimensional model fits quite closely with the bifactor model for both data sets (science and math). On the other hand, the multi-dimensional IRT model has the weakest model fit for both test types. In line with all these findings, the methods used when determining the testlet items were discussed and estimation suggestions were made for implementations using testlets, especially PISA.

1. INTRODUCTION

PISA (Program for International Student Assessment) is a large-scale examination implemented by the OECD (Organization for Economic Co-operation and Development), which is attended by many countries that evaluate the knowledge and skills acquired by students aged 15 in three-year periods. The main purpose of PISA is to measure students’ ability to transfer the knowledge and skills they have learned at school into daily life. Within this scope, there are three main evaluation areas, namely science, math and reading literacy in the part where cognitive evaluation is made. The concept of “literacy” used in PISA research is defined as the capacity of students to transfer their knowledge into daily life and to make logical

*CONTACT: Cansu AYAN ✉ cnsayan@gmail.com 📍 Ankara University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, Ankara, Turkey

inferences. As can be understood from the definition, this international test administration focuses on the higher-order skills such as analysing and evaluation rather than the cognitive levels such as memorizing or remembering information directly. Many different item types are used to serve this purpose (OECD, 2019c).

One of the PISA item types that make it easy to measure higher-order thinking skills is the item types, which are linked to a common stimulus. Types of items linked to a common stimulus are named “Testlet Items” in the literature. In this item type, many item stems are created from contents such as a picture, a text or a scenario that is used as a common stimulus. Thanks to the content it uses, this format helps make it possible to measure higher-order thinking levels by bringing measurement-evaluation practices closer to real-life problem situations. Furthermore, it can save time by having many items created from the same content (Bao, 2007; DeMars, 2006; Wainer et al., 2000).

In addition to the many advantages of testlet items, which are mentioned above, their limitations have also been a topic of discussion in the related literature. The first one of these discussions is that these items threaten the local independence (LID) assumption, which is one of the main assumptions of the Item Response Theory (IRT). Local independence means that whether a person responds to an item correctly or incorrectly depends only on the ability of that person, and that the items s/he has answered before do not affect this situation (Embretson & Reise, 2000; Hambleton et al., 1991). While in the literature there are many known reasons for local item dependence, one of the most frequently discussed reasons is the dependence arising from the fact that the items are linked to a text (Yen, 1993). The responses to the items may be related to each other in item groups with the same content. For example, for a set of items connected to a reading item, an individual’s interest in the content presented in the reading can be a second factor that will affect answering the items correctly. In this case, it may not be correct to claim that the answers given to these items are independent from each other (Bao, 2007; Fukuhara & Kamata, 2011; Yen, 1993). Many studies that are consistent with this situation have also shown that when a testlet is used in tests, the LID assumption is violated (Lee et al., 2001; Sireci et al., 1991; Wainer & Lewis, 1990; Yen, 1993).

The fact that uni-dimensional IRT models are insufficient in estimating the model parameters, since they violate the LID assumption in the tests where the testlets are used, has become a current issue. Many studies have been conducted on how uni-dimensional IRT estimates affect the results without taking the LID assumption into consideration (Bradlow et al., 1999; Chen & Thissen, 1997; DeMars, 2012; DeMars, 2006; Li et al., 2005; Marais & Andrich, 2008; Sireci et al., 1991; Tuerlinckx & De Boeck, 2001; Wainer & Wang, 2000; Yen, 1993). Overestimation of reliability or information and underestimation of standard errors for ability estimates are possible drawbacks of violation of LID (Wainer & Wang, 2000; Yen, 1993). This also leads to misestimation of parameters. Wainer and Wang (2000) showed that when the local dependence that stemmed from the testlet structure was ignored, item difficulties were still well estimated but lower asymptotes were overestimated, and the discrimination parameters that were overestimated for one test were underestimated for another test. Wainer et al., (2000) proved that by ignoring testlet dependence, discrimination was the most affected parameter among other parameters (trait and difficulty). Wainer and Wang (2000) found that when the testlet dependence was ignored and not modelled, the item discriminations were underestimated for testlet items and overestimated for independent items. Ackerman (1987) found that when the items were locally dependent, item discriminations were underestimated. When a multi-dimensional structure exists, alternative psychometric models should be used for modelling LID. In this context, the issue of which alternative psychometric model to use for measurements using testlets has become a popular topic of discussion. Based on all these research findings, one of the psychometric models proposed for measurements involving testlets is the bifactor

model. The bifactor model is a special version of multi-dimensional IRT developed as an extension of Spearman's bifactor theory (Holzinger & Swineford, 1937). In the bifactor model, it can be possible to load items in two different factors, being one general factor and one or more than one specific factors. In this way, both general and specific factor effects on the items can be estimated and interpreted simultaneously (Canivez, 2016; Houts & Cai, 2013; Reise et al., 2010). This can be considered as a solution for tests in which item sets are used. In the bifactor model, items using the same content are loaded on the same specific factor and also all items are loaded on the general factor. Thus, the properties resulting from the common content of the items that cause the violation of the LID assumption can be modelled in the specific factors (Gibbons & Hedeker, 1992; Houts & Cai, 2013). In the light of all this information, the estimation model of PISA, where item sets are frequently used, can also be discussed.

When the PISA estimation procedure was investigated, it was seen that in the PISA 2018, the uni-dimensional multiple-group IRT model for binary items and the generalized partial credit (GPC) model for the polytomous item responses were used for each of the domains (OECD, 2019c). In this context, when the literature was examined, studies comparing the estimation accuracy of the bifactor model with other IRT models using the PISA items (DeMars, 2006; Yılmaz Koğar, 2016) were found. As a result of these studies, it was seen that the best fitting model was the bifactor model. In the related studies, all items in the same unit were analysed by assuming they were connected to the same common stem. According to PISA, math items are arranged in units that share the stimulus material and it is usually the case that all items in the same unit belong to the same context category (OECD, 2019a). Moreover, PISA science items are arranged in units that are introduced by the specific stimulus material, which may be a brief written passage, or a text accompanying a table, a chart, a graph or a diagram (OECD, 2019b). However, when the reading items released by the PISA 2018 were examined, it was seen that there were three reading passages named "Professor's Blog", "Review of Collapse", and "Did Polynesian Rats Destroy Rapa Nui's Trees?" in the unit named "Rapa Nui". Similarly, there were two reading passages named "Farm to Market" and "Just Say No" in the unit named "Cow's Milk". In this case, it would not be correct to consider all the items in the units "Rapa Nui" and "Cow's Milk" as if they shared the same common stem.

Similarly, when the science items released by the PISA 2015 were examined, in the "Bird Migration" unit, it was seen that there were two different reading passages named "Bird Migration" and "Golden Plovers". When the math items released by the PISA 2012 were investigated, it was seen that in the "Penguins" unit, the first three items partially shared the same passage but the fourth item had its own graph and the student used just that graph to solve that item. In this case, it would not be correct to consider and analyse the items in the aforementioned units as if they shared the same stem. Besides, Baldonado et al., (2015) pointed out the danger that considering items as locally dependent may overestimate the true dependence among the items, even for items sharing the same common stem, without doing any extra investigations. In contrast, they proposed another method, which is based on determining which sentence or information in the passage is used to answer the item correctly, and which requires a detailed examination of the item contents. Underlining that the entire passage is less important than the part needed to answer the item correctly, they state that there is often no dependence for items referring to unique parts of the text.

As stated before, when the items released by the PISA 2018, 2015 and 2012 were examined, units with more than one stimulus were found. However, in the PISA 2018 Framework, it was stated that the items shared a common stimulus (OECD, 2019a; OECD, 2019b). It was seen in the examinations that the fact that the items were from the same unit does not guarantee that they would share the same common stem.

Consequently, it is questionable for the items to be considered as a testlet for all situations where a common stem is used. This situation especially raises more suspicion for situations such as the PISA implementations, where the items are not published and the contents cannot be examined in detail. All these ambiguities make it necessary to conduct more studies on this topic. Due to the advantages they provide, testlet items are a type of item, which is increasingly used in many areas from small-scale classroom implementations to large-scale international implementations. It is thought that having both conceptual and psychometric discussions about this item type is very important for obtaining valid and reliable results from implementations using this item type. This research is an important study, since it aims to help eliminate the unit-testlet ambiguity in PISA in the literature. Within the scope of this study, the estimation results of the bifactor model, uni-dimensional IRT model and multi-dimensional IRT model were compared in the presence of testlets in which both dichotomous and polytomous items existed.

The main purpose of this research is to compare the model estimation results of the bifactor-GPC model with the multi-dimensional-GPC (multi-GPC) model and uni-dimensional GPC (uni-GPC) model for dichotomous and polytomous items from science and math tests in PISA 2018 and to clarify the unit-testlet dilemma.

For this purpose, the following research questions were asked. For science and math;

- (1) Do the items show local dependence for each of the bifactor-GPC, multi-GPC and uni-GPC models?
- (2) What are the model fit indices of the bifactor-GPC, multi-GPC and uni-GPC model estimations?
- (3) What are the item parameters obtained from the bifactor-GPC, multi-GPC and uni-GPC models?
- (4) What are the variance rates explained on the basis of general and specific factors?

2. METHOD

2.1. Participants

The participants of the study were selected from students who participated in the PISA 2018. Among these people, the study was carried out with individuals who took the selected booklets without making a country distinction. In this context, 9365 examinees who completed the selected booklet were selected for the math test. Similarly, 6487 examinees were also selected for the science test with the same method.

2.2. Instrument

The results of the PISA 2018 were used in this study for the real data. Math and science tests were used by selecting a booklet from each. Selected booklets were determined according to its number of polytomous items. The items on the math test came from Booklet 11. Booklet 11 consisted of 24 items in total: two 2-item testlets, three 3-item testlets, one 4-item testlet and seven independent items. Independent items in Booklet 11 were removed, as they were not within the scope of this study. After removing the independent items, 17 items remained. Among these 17 items, four were polytomous (partial credit) and the other 13 were dichotomous items. Polytomous items were coded as follows: 0 for no credit, 1 for partial credit and 2 for full credit. The items on the science test came from Booklet 15. Booklet 15 consisted of 38 items in total: two 5-item testlets, three 4-item testlets, four 3-item testlets and two 2-item testlets. There were no independent items in Booklet 15. Among these 38 items, four were polytomous and 34 were dichotomous items.

2.3. Estimation Procedure

In this study, a mixed item type (dichotomous and polytomous) was used. For both the math and science tests, the items were analyzed according to the GPC model for three IRT models

(bifactor, uni-dimensional and multi-dimensional). Since PISA items are partially scored items, Muraki (1992)'s Generalized Partial Credit (GPC) model was used for parameter estimations. The GPC model is a generalized form of the two-parameter logistic (2PL) model for polytomous data, which describes an examinee's probability of selecting a possible score category among all score categories. When an item has two response categories, the GPC model is equal to the 2PL model.

Chon et al., (2007) found that the GPC model fits mixed data (polytomous and dichotomous) better than 3PL (three-parameter logistic) or 2PL (two-parameter logistic) models. The Metropolitan-Hastings Robbins-Monro (MH-RM) algorithm was used for the parameter estimation method. The MH-RM is ideal for mixing different item response models (dichotomous and polytomous) with many items, many factors and a large sample size (Cai, 2010). Finally, all analyses were made with R-Studio 1.2.5001 and Excel.

2.3.1. Estimation of marginal item parameters

According to Stucky and Edelen (2014), in the bifactor model, slopes on the general trait have an effect of specific traits. So, the inflation of conditional slopes of the general trait is a consequence of the conditional relation between the specific traits and the general trait. Thus, direct comparison should not be made between specific and general slopes (Stucky et al., 2013). Therefore, marginal slopes were calculated to compare the model (uni-GPC, bifactor-GPC, multi-GPC) parameters using equations (Eq.1. Eq.2. Eq.3) (Stucky & Edelen, 2014; Stucky et al., 2013).

$$\lambda_j^{*G} = \frac{\alpha_j^G/D}{\sqrt{1+(\alpha_j^G/D)^2+(\alpha_j^S/D)^2}} \quad (\text{Eq.1})$$

$$(\sigma_j^{*G})^2 = 1 - (\lambda_j^{*G})^2 \quad (\text{Eq.2})$$

$$\alpha_j^{*G} = \left(\frac{\lambda_j^{*G}}{\sqrt{(\sigma_j^{*G})^2}} \right) \quad (\text{Eq.3})$$

According to the equations, D= a scaling constant of 1.7, λ_j^{*G} = marginal loading of item j on the general trait, $(\sigma_j^{*G})^2$ = unexplained (unique) item variance on the general trait, α_j^G = conditional slope for item j on the general trait, α_j^S = conditional slope for item j on a specific trait. The marginal location parameter on the general trait should be calculated according to Eq. 4 (Stucky & Edelen, 2014).

$$b_{j(k)}^* = \frac{-c_{jk}}{\alpha_j^G} \quad (\text{Eq.4})$$

Ip (2010) showed that marginalization of parameters does not affect the b- and c- parameters. However, in this study, all parameters were marginalized for both the general trait and specific traits (see [Appendix B](#)).

2.3.2. Dimensionality analysis

Before the IRT analysis, dimensionality of the data was detected for both math and science tests. If the tests were uni-dimensional, then there would be no significant testlet factors. For dimensionality analysis, parallel analysis was done via psych (Revelle & Revelle, 2015) package in R.

Figure 1. Parallel analysis for science test

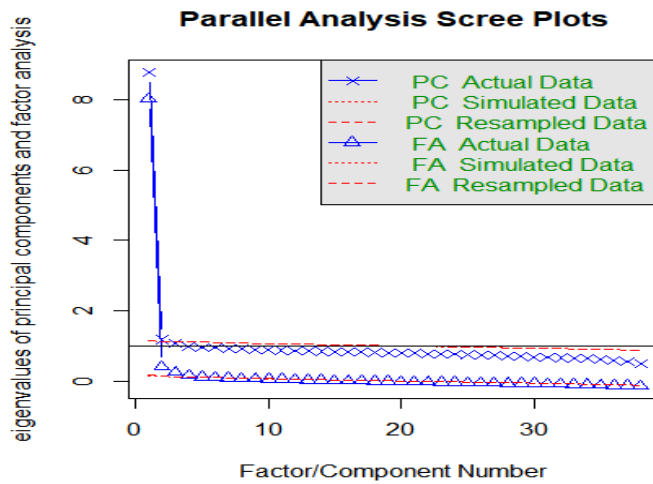
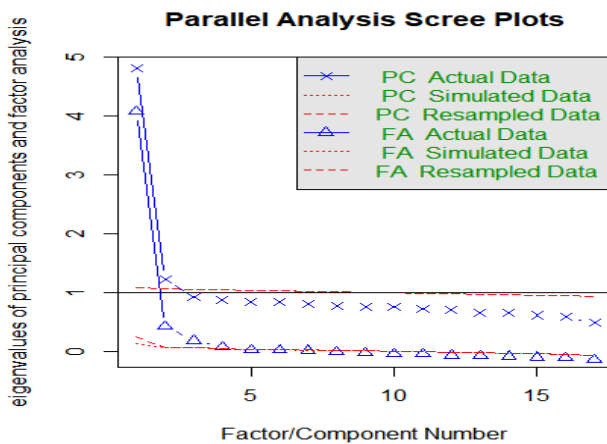


Figure 2. Parallel analysis for math test



Item clusters that had eigenvalues greater than 1 were designated as components and the existence of more than one component showed that data were not uni-dimensional. According to the scree plots for science (Figure 1) and math (Figure 2), it was seen that the data were not uni-dimensional. For math data 4 factors, and for science data 6 factors were extracted.

3. RESULT / FINDINGS

3.1. Evaluation of Local Independence

Local Independence (LID) is examined according to Chen and Thissen’s (1997) standardized local dependence (LD) χ^2 statistics. Large positive LD values indicate that the covariation between item responses is not completely modelled by a given IRT model. Local dependence was calculated via R-Studio 1.2.5001. R computes the local dependence according to Cramer’s V. When an item has two categories, Cramer’s V gives the same output with the phi coefficient. The datasets of this research consisted of mixed items. Therefore, the LD matrix was interpreted according to Cramer’s V coefficient cut-off values, the same as phi, which is > 0.15 for strong association and > 0.25 for very strong association (Akoğlu, 2018). Table 1 summarizes the items, which shows LD for three IRT models (for all LD values see Appendix A).

Table 1. Number of items with LD.

	IRT Models		
	Uni-GPC	Multi-GPC	Bifactor-GPC
Math (17 items)	1	13	None
Science (38 items)	None	34	None

Large positive LD values show that there is an unmodelled covariance between items by a given IRT model (Cai et al., 2015). As seen in Table 1, for math, while only one item (M32 with M33) showed local dependence in the uni-GPC model, for the multi-GPC model almost all items showed local dependence. In addition to this, none of the items showed local dependence in the bifactor-GPC model. As with the math test, similar results were seen for the science test. In the science test, none of the items showed local dependence in the bifactor-GPC and uni-GPC, whereas for the multi-GPC, almost all items showed local dependence. This result reveals that in modeling of item covariance, bifactor-GPC and uni-GPC are better than multi-GPC. Also, it seems possible that this result is due to the unmodelled item covariance regarding the general factor in the multi-GPC.

3.2. Global Model-Data Fit and Comparison

Nested models should be compared in terms of goodness of fit with the deviance statistics. The deviance statistics are calculated by the difference between the more complex model (more parameters) and the reduced model (fewer parameters) and have a χ^2 distribution. In this study, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used for model fit. Also, Cai and Monroe (2014) omnibus limited-information goodness-of-fit statistic, C_2 , was used for model fit. C_2 was chosen over other goodness-of-fit statistics (M_2^* : Cai & Henson, 2013; M_2 : Maydeu-Olivares & Joe, 2005) because it was suitable for the ordinal response data and shows the same performance as M_2 and M_2^* but can be more powerful (Cai & Monroe, 2014). C_2 , M_2 and M_2^* are equal when the items are dichotomous. Because C_2 has a χ^2 distribution, it is sensitive to the sample size. Therefore, model error or misspecification can be computed, such as the root mean square error of approximation (RMSEA), as in the structural equation modelling literature, but it is computed based on the C_2 statistic ($RMSEA_{C_2}$) (Toland et al., 2017). As Toland et al., (2017) emphasize, IRT models are non-linear models and traditional RMSEA is for linear models, so cut-off ($RMSEA \leq 0.08$ - adequate fit) should be interpreted cautiously. Smaller RMSEA values are an indicator of a better model-data fit.

Table 2. Model-data fits for three IRT models.

	IRT models	-2LL	BIC	AIC	C_2 (df)	$RMSEA_{C_2}$
Math	Bifactor-GPC	-92372.3	185247.6	184854.7	251.78(98) ^{***}	0.013
	Uni-GPC	-92739.3	185826.1	185554.6	1190.51(119) ^{***}	0.031
	Multi-GPC	-92513.9	185133.9	185512.6	12156.09(104) ^{***}	0.111
Science	Bifactor-GPC	-145160.2	291356.2	290556.5	1211.47(623) ^{***}	0.013
	Uni-GPC	-145360.5	291423.2	290881.0	1747.55(665) ^{***}	0.016
	Multi-GPC	-153924.4	308682.6	308038.8	15347.26(650) ^{***}	0.059

^{***} $p < 0.001$

Table 2 summarizes the three IRT model comparisons. All fit statistics (AIC, BIC, $RMSEA_{C_2}$) prove that, for both tests, bifactor-GPC has a better fit than the other two IRT models. When $RMSEA_{C_2}$ statistics were compared, for both tests (math and science), the bifactor-GPC model showed the lowest value among the models. The uni-GPC model comes after the multi-GPC model, which had the largest $RMSEA_{C_2}$ value among the models. Since the $RMSEA_{C_2}$ statistics

should be interpreted cautiously for non-linear models, they were interpreted relatively. To understand the models in depth, detailed inspection was made for the bifactor-GPC and uni-GPC based on the item parameters.

3.3. Comparison of Item Parameters / Model parameters

The marginal slopes are the adjusted slopes to compare the uni-GPC and bifactor-GPC models. For the math test, when the conditional and marginal parameters were examined, there were slight differences detected between those parameters for items M13, M53, M61, and M62, which had slopes close to “0” on the specific trait. These differences could have occurred because the specific trait did not affect the probability of responding to the item. Only item M33 showed local dependence in the uni-GPC, and had higher slopes in both the uni-GPC and bifactor-GPC. Also, that item’s marginal and conditional slopes differed greatly. There was a slight difference between uni-GPC and bifactor-GPC slope parameters. When the multi-GPC slopes were compared with those of the uni-GPC and the marginal coefficient for the bifactor-GPC, it was seen that the multi-GPC had larger slopes than both of the other models. Inflation of slopes may have resulted from the larger LD values of the multi-GPC. The larger LD values may have arisen from the undefined latent factor (general factor) underlying the items.

For the science test, similar results were obtained to those of the math test. When the conditional and marginal parameters were examined, there were no differences detected between those parameters for items SC71 and SC94, which had slopes close to “0” on the specific trait. It was seen that when the specific trait slopes became higher, the gap between the marginal and conditional slopes increased. When the slope parameters were compared between the bifactor-GPC and uni-GPC, slight differences were detected. When the multi-GPC slopes were compared with those of the uni-GPC and the marginal coefficients for the bifactor-GPC, it was seen that the multi-GPC had larger slopes than both of the other models. Inflation of slopes may have resulted from the larger LD values and the undefined general trait of the multi-GPC.

3.4 Explained Common Variance

The explained common variance (ECV) index is a useful psychometric measure to determine both the magnitude of the general trait related to a specific trait and essential uni-dimensionality (Reise et al., 2010).

Table 3. Explained common variances for math items.

Item	$IECV_G$	$IECV_S$
M11	0.916	0.084
M12	0.760	0.240
M13	0.988	0.012
M14	0.927	0.073
M21	0.981	0.019
M22	0.773	0.227
M31	0.960	0.040
M32	0.675	0.325
M33	0.796	0.204
M41	0.920	0.080
M42	0.805	0.195
M43	0.718	0.282
M51	0.953	0.047
M52	0.965	0.035
M53	0.997	0.003
M61	0.999	0.001
M62	0.852	0.148

Table 4. Explained common variances for science items.

Item	IECV _S	Item	IECV _S	Item	IECV _S	Item	IECV _S
SC11	0.89	SC63	0.98	SC11	0.11	SC63	0.02
SC12	0.66	SC71	1.00	SC12	0.34	SC71	0.00
SC13	0.52	SC72	0.86	SC13	0.48	SC72	0.14
SC14	0.84	SC73	0.99	SC14	0.16	SC73	0.01
SC21	0.95	SC74	0.97	SC21	0.05	SC74	0.03
SC22	0.88	SC81	0.99	SC22	0.12	SC81	0.01
SC23	0.75	SC82	0.94	SC23	0.25	SC82	0.06
SC31	0.92	SC83	0.80	SC31	0.08	SC83	0.20
SC32	0.78	SC91	0.82	SC32	0.22	SC91	0.18
SC33	0.92	SC92	0.98	SC33	0.08	SC92	0.02
SC34	0.91	SC93	0.91	SC34	0.09	SC93	0.09
SC35	0.85	SC94	1.00	SC35	0.15	SC94	0.00
SC41	0.83	SC101	0.98	SC41	0.17	SC101	0.02
SC42	0.95	SC102	0.80	SC42	0.05	SC102	0.20
SC51	0.90	SC103	0.85	SC51	0.10	SC103	0.15
SC52	0.92	SC104	0.92	SC52	0.08	SC104	0.08
SC53	0.96	SC105	0.81	SC53	0.04	SC105	0.19
SC61	0.71	SC111	0.87	SC61	0.29	SC111	0.13
SC62	0.96	SC112	0.90	SC62	0.04	SC112	0.10

Table 3 and Table 4 summarize The ECV indices, which were calculated for items, and general and specific traits. For the math data, results showed that general trait and specific factors explained respectively 86%, 2%, 1%, 7%, 4%, 0.40% and 0.10% of the common variance. Specific traits explained a small amount of variance in contrast with the general trait except S3, which explained 7% of the variance. That specific factor contained the items (M33 with M32) with LD in the uni-GPC model. This proves that the S3 specific factor had a unique effect on those items. Because other specific factors had a small amount of unique (specific) variance, the uni-GPC model may have shown almost the same slope parameters as the bifactor-GPC.

For the science test, results showed that general trait and specific factors explained respectively 89%, 2%, 1%, 2%, 0%, 1%, 1%, 0%, 1%, 0%, 3% and 1% of the common variance. Specific traits explained a small amount of the variance. As with the math test, because of the low uniqueness, the uni-GPC and bifactor-GPC slope parameter estimates also became closer in the science test.

4. DISCUSSION and CONCLUSION

Within the scope of the study, an attempt was made to determine the most appropriate estimation model for the data by comparing the uni-GPC, multi-GPC and bifactor-GPC model estimations for the two booklets selected from the science and math sections of the PISA 2018. As a result, an effort was made to eliminate the unit-testlet ambiguity in PISA in the literature. Care was taken to ensure that both the testlet item groups and the binary and multiple scored item samples were all together in the selected booklets, and how this situation would affect the estimation results was emphasized. In this context, model-fit indices related to the three models (uni-GPC, multi-GPC, bifactor-GPC), differences in item parameter estimation results, and variance ratios explained within the scope of general and specific traits were examined. Before presenting and discussing the results, it can be said that the first findings were very similar for the science and math data. The discussions within this scope are valid for both areas.

In the literature, it was stated that in addition to interaction among the items, multi-dimensionality can also reveal local item dependence (Embretson & Reise, 2000; Tuerlinckx & De Boeck, 2001). In this context, it was observed in this study that the items were multi-dimensional for both data sets (math and science) in the dimensionality analyses made before starting the estimations. However, when the model-data fit analyses were examined, it was seen that the multi-GPC model indicated the worst fit in both the math and science data set. While the bifactor-GPC model provided the best fit, the uni-GPC model fit was very close to that of the bifactor-GPC. Among the compared models, the bifactor-GPC model was expected to indicate the best model fit, which is a consistent finding with the studies by Demars (2006) and Yılmaz Koğar (2016). On the other hand, the fact that the data set of the uni-GPC provided close results to those of the bifactor-GPC and that the multi-GPC provided the worst fit is an unexpected case. This may be because the data set has minor factors. McDonald (2000) explains that the bifactor model should only be meaningfully applied when definable “content facets” that form well-structured secondary dimensions exist. Additionally, Ackerman et al., (2003) state that if subsets of items are from distinct content areas and/or cognitive skills, these items have the potential of being in distinct dimensions.

In order to make a detailed investigation between the models, the item parameters were also examined. While the slopes were very close to each other in the bifactor-GPC and uni-GPC models, larger slopes were obtained in the multi-GPC model than in the other two models. For this case, it can be said that the unmodelled covariance causes slope parameters to be overestimated. The fact that slopes were larger than actual in item parameter estimations without considering local independence is consistent with many study findings in the literature (Ackerman, 1987; Bradlow et al., 1999; Chen & Thissen, 1997; DeMars, 2006; DeMars, 2012; Lee et al., 2001; Li et al., 2005; Sireci et al., 1991; Tuerlinckx & De Boeck, 2001; Wainer et al., 2000; Wang & Wilson, 2005; Yen, 1993).

Examining the variance rates explained on the basis of general and specific factors was another investigation made on the dataset. Most of the variance explained (about 85%) stemmed from the general trait. The effect of specific traits on the variance was very low.

In the specific trait that had the highest contribution to variance in the math data, it was seen that the uni-GPC model analyses included locally dependent items (M32-M33), which is in fact exactly as expected. Locally dependent items also showed considerable weight in the specific trait. However, when the math data were evaluated as a whole, it was determined that the specific factor weights predominantly were quite small, which means that there was a data set with a dominant general factor. This finding is also consistent with the model fit result. Having a dominant general factor caused the model to be the most compatible with the data, after the bifactor model, to be a uni-dimensional IRT model rather than a multi-dimensional model. This result may mean that the accepted assumption in the literature that analyzing with uni-dimensional models will have erroneous results when there is a testlet item must be rethought, and that its limits must be redrawn. In their study with a data set made up of questions based on a reading passage, Baldonado et al., (2015) pointed out the danger that simple approaches that accept all of the items using the same content as local dependent could overestimate the actual dependence among the items. In order to reach more accurate conclusions about the dependence of the items, they proposed another method in which the “necessary information”, which indicated the information used in the passage to examine the content of each item and answer the item correctly, was identified. They argued that the entire passage is less important than the part, which is required to answer an item correctly, and that the approach that assumes the items as dependent since they belong to the same passage, regardless of whether the items share common “necessary information”, would be an overly general approach. Often, multiple questions associated with the same passage refer to different parts of the text. In such cases, a situation where a common passage causes some dependence among the item response processes

may not occur. This also points to the need to consider to what extent the items with the same content in PISA are testlets in various studies. On the other hand, the method proposed by Baldonado et al., (2015) requires examining the item contents based on expert opinion. Considering that PISA items are not disclosed, this situation becomes quite difficult. In cases in which the effect of specific factors is very low in the data set and a general factor is observed, researchers' analysis with uni-IRT will not cause a large bias in the results.

In the light of all these results and discussions, researchers who are to work on testlet items are recommended not to make decisions based on the use of the same content only and if possible, to examine the contents of the items in detail. If this is not possible, it is suggested that they decide which model is to be used by carefully examining the variance rates (based on general and specific factors) which are explained by the local dependence analysis results.

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

Cansu Ayan: Investigation, Resources, Visualization, Software, Analyze, and Writing. **Fulya Baris-Pekmezci:** Investigation, Methodology, Analyze, Supervision, Validation, and Writing.

ORCID

Cansu Ayan  <https://orcid.org/0000-0002-0773-5486>

Fulya Baris Pekmezci  <https://orcid.org/0000-0001-6989-512X>

5. REFERENCES

- Ackerman, T. A. (1987, April). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence*. Paper presented at the annual meeting of the American Educational Research Association. Washington, DC.
- 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>
- Akoğlu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91-93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Baldonado, A. A., Svetina, D., & Gorin, J. (2015). Using necessary information to identify item dependence in passage-based reading comprehension tests. *Applied Measurement in Education*, 28(3), 202-218. <https://doi.org/10.1080/08957347.2015.1042154>
- Bao, H. (2007). *Investigating differential item function amplification and cancellation in application of item response testlet models* [Doctoral dissertation, University of Maryland]. ProQuest Dissertations and Theses Global.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153–168.
- Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3), 307-335. <https://doi.org/10.3102/1076998609353115>
- Cai, L., du Toit, S. H. C., & Thissen, D. (2015). *IRTPRO: Flexible professional item response theory modeling for patient reported outcomes (version 3.1)* [computer software]. SSI-International.
- Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, 66(2), 245-276. <https://doi.org/10.1111/j.2044-8317.2012.02050.x>

- Cai, L., & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data*. (CRESST Report 839). National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Canivez, G. L. (2016). Bifactor modeling in construct validation of multifaceted tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer & C. DiStefano (Eds.). *Principles and methods of test construction: Standards and recent advancements* (pp. 247-271). Hogrefe Publishers.
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289. <https://doi.org/10.3102/10769986022003265>
- Chon, K. H., Lee, W., & Ansley, T. N. (2007). *Assessing IRT model-data fit for mixed format tests*. (CASMA Research Report 26). Center for Advanced Studies in Measurement and Assessment.
- DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, 43, 145–168. <https://doi.org/10.1111/j.1745-3984.2006.00010.x>
- DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, 36, 104–121. <https://doi.org/10.1177/0146621612437403>
- Embretson, S., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates Inc.
- Fukuhara, H., & Kamata, A. (2011). Functioning analysis on testlet-based items a bifactor multidimensional item response theory model for differential items. *Applied Psychological Measurement*, 35(8), 604–622. <https://doi.org/10.1177/0146621611428447>
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, 57, 423–436.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publications.
- Holzinger, K. J., Swineford, F. (1937). The Bi-factor method. *Psychometrika*, 2, 41–54. <https://doi.org/10.1007/BF02287965>
- Houts, C. R., & Cai, L. (2013). *Flexible multilevel multidimensional item analysis and test scoring* [FlexMIRT R user's manual version 3.52]. Vector Psychometric Group.
- Ip, E. H. (2010). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement*, 34(7), 467-482. <https://doi.org/10.1177/0146621610364975>
- Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. *Educational and Psychological Measurement*, 61, 958-975. <https://doi.org/10.1177/00131640121971590>
- Li, Y., Bolt, D. M., & Fu, J. (2005). A test characteristic curve linking method for the testlet model. *Applied Psychological Measurement*, 29(5), 340-356. <https://doi.org/10.1177/0146621605276678>
- Marais, I. D., & Andrich, D. (2008). Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. *Journal of Applied Measurement*, 9, 105–124.
- Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2" contingency tables: A unified framework. *Journal of the American Statistical Association*. <https://doi.org/10.1198/016214504000002069>
- McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, 24(2), 99-114. <https://doi.org/10.1177/01466210022031552>

- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, 1992(1), i-30. <https://doi.org/10.1002/j.2333-8504.1992.tb01436.x>
- OECD (2019a). “PISA 2018 Mathematics Framework”. in *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. <https://doi.org/10.1787/13c8a22c-en>
- OECD (2019b). “PISA 2018 Science Framework”. in *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. <https://doi.org/10.1787/f30da688-en>
- OECD (2019c). “Scaling PISA data”. in *PISA 2018 Technical Report*. OECD Publishing. <https://www.oecd.org/pisa/data/pisa2018technicalreport/Ch.09-Scaling-PISA-Data.pdf>
- Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, 92(6), 544-559. <https://doi.org/10.1080/00223891.2010.496477>
- Revelle, W., & Revelle, M. W. (2015). Package ‘psych’. *The comprehensive R archive network*, 337, 338.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237-247. <https://doi.org/10.1111/j.1745-3984.1991.tb00356.x>
- Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.) *Handbook of item response theory modelling*. (pp. 201-224). Routledge.
- Stucky, B. D., Thissen, D., & Orlando Edelen, M. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement*, 37(1), 41-57. <https://doi.org/10.1177/0146621612462759>
- Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, 60, 41-63. <https://doi.org/10.1016/j.jsp.2016.11.001>
- Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, 6(2), 181–195. <https://doi.org/10.1037/1082-989X.6.2.181>
- Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.). *Computerized adaptive testing: Theory and practice* (pp. 245–269). Springer, Dordrecht.
- Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, 27(1), 1–14. <https://doi.org/10.1111/j.1745-3984.1990.tb00730.x>
- Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, 37(3), 203–220. <https://doi.org/10.1111/j.1745-3984.2000.tb01083.x>
- Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, 29(2), 126-149. <https://doi.org/10.1177/0146621604271053>
- Yen, W. M. (1993). Scaling performance assessments Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>
- Yılmaz Kogar, E. (2016). *Madde takımları içeren testlerde farklı modellerden elde edilen madde ve yetenek parametrelerinin karşılaştırılması [Comparison of item and ability parameters obtained from different models on tests composed of testlets]* [Doctoral dissertation, Hacettepe University]. Hacettepe University Libraries, <https://avesis.hacettepe.edu.tr/yonetilen-tez/c2ade6a0-6a2d-4147-beb0-8a3feb0642c5/madde-takimlari-iceren-testlerde-farkli-modellerden-elde-edilen-madde-ve-yetenek-parametrelerinin-karsilastirilmesi>

6.2. Appendix B

Table B1. Uni-GPC and Multi-GPC parameters for Math.

Item id.	Uni-GPC			Multi-GPC								
	a1	c1	c2	a1	a2	a3	a4	a5	a6	c1	c2	
M11	1.02	0.19	1.27	1.15							0.31	1.38
M12	1.44	1.39		1.60							1.46	
M13	0.86	-1.59	-1.84	0.95							-1.59	-1.92
M14	0.98	0.30		1.04							0.31	
M21	1.72	-0.15			2.34						-0.16	
M22	0.87	-0.23			0.96						-0.23	
M31	1.39	2.52				1.55					2.66	
M32	1.56	0.45				2.03					0.55	
M33	2.60	-1.36				5.13					-2.32	
M41	1.66	1.30					1.99				1.45	
M42	1.60	-0.64					2.01				-0.71	
M43	0.84	-2.48	-0.38				1.02				-2.42	-0.41
M51	1.06	1.24						1.12			1.27	
M52	1.36	-0.58						1.50			-0.60	
M53	1.19	-2.93	-3.40					1.37			-3.00	-3.74
M61	1.22	1.21							2.00		1.54	
M62	0.42	-0.50							0.42		-0.50	

Table B2. Bifactor-GPC conditional parameters for Math.

Item id.	ag	a1	a2	a3	a4	a5	a6	c1	c2
M11	1.07	0.33						0.27	1.35
M12	1.59	0.90						1.56	
M13	0.88	0.09						-1.59	-1.86
M14	0.99	0.27						0.31	
M21	1.79		0.25					-0.14	
M22	0.93		0.52					-0.24	
M31	1.41			0.30				2.56	
M32	1.95			1.59				0.62	
M33	3.70			2.11				-1.99	
M41	1.70				0.51			1.35	
M42	1.72				0.89			-0.70	
M43	1.01				0.66			-2.37	-0.46
M51	1.09					0.24		1.26	
M52	1.39					0.26		-0.59	
M53	1.23					0.05		-2.94	-3.47
M61	1.26						0.03	1.23	
M62	0.42						0.18	-0.51	

Table B3. *Bifactor-GPC marginal parameters for Math.*

Item id.	ag	a1	a2	a3	a4	a5	a6	b1	b2
M11	1.06	0.28						0.26	1.26
M12	1.40	0.66						0.98	
M13	0.88	0.08						-1.81	-2.12
M14	0.98	0.23						0.31	
M21	1.77		0.17					-0.08	
M22	0.88		0.46					-0.25	
M31	1.39			0.23				1.82	
M32	1.43			1.04				0.32	
M33	2.32			0.88				-0.54	
M41	1.63				0.36			0.80	
M42	1.52				0.63			-0.41	
M43	0.94				0.57			-2.36	-0.46
M51	1.08					0.20		1.15	
M52	1.37					0.20		-0.42	
M53	1.23					0.04		-2.40	-2.83
M61	1.26						0.03	0.97	
M62	0.42						0.17	-1.21	

Table B4. Uni- GPC model parameters for Science.

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	c1	c2
1.37											0.23	
0.94											-1.19	
0.66											0.03	
1.28											-0.73	
	2.01										-0.50	
	1.10										-0.12	
	0.96										1.08	
		0.73									0.48	0.77
		1.59									1.54	
		2.22									0.00	
		1.20									0.53	0.05
		1.42									-2.16	-3.22
			0.94								-0.14	
			1.50								-1.05	
				1.61							-0.09	
				1.10							0.63	
				1.22							0.20	
					0.79						0.51	
					1.31						0.34	
					1.59						1.33	
						0.65					-0.30	
						0.88					0.70	
						1.22					0.21	
						1.19					-0.68	
							1.06				1.47	
							1.81				1.07	
							0.76				-0.77	-1.30
								0.63			-0.53	
								1.62			0.72	
								0.82			-0.07	
								1.54			0.31	
									1.25		-0.65	
									4.32		3.02	
									2.25		0.24	
									2.07		-1.40	
									0.94		-1.70	
										1.08	-0.89	
										1.49	2.32	

Table B5. Multi-GPC model parameters for Science.

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	c1	c2
1.37											0.23	
0.94											-1.19	
0.66											0.03	
1.28											-0.73	
	2.01										-0.50	
	1.10										-0.12	
	0.96										1.08	
		0.73									0.48	0.77
		1.59									1.54	
		2.22									0.00	
		1.20									0.53	0.05
		1.42									-2.16	-3.22
			0.94								-0.14	
			1.50								-1.05	
				1.61							-0.09	
				1.10							0.63	
				1.22							0.20	
					0.79						0.51	
					1.31						0.34	
					1.59						1.33	
						0.65					-0.30	
						0.88					0.70	
						1.22					0.21	
						1.19					-0.68	
							1.06				1.47	
							1.81				1.07	
							0.76				-0.77	-1.30
								0.63			-0.53	
								1.62			0.72	
								0.82			-0.07	
								1.54			0.31	
									1.25		-0.65	
									4.32		3.02	
									2.25		0.24	
									2.07		-1.40	
									0.94		-1.70	
										1.08	-0.89	
										1.49	2.32	

Table B6. *Bifactor-GPC conditional parameters for Science.*

ag	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	c1	c2
1.12	0.39											0.21	
0.82	0.59											-1.22	
0.56	0.55											0.03	
1.03	0.45											-0.71	
1.60		0.35										-0.46	
1.02		0.37										-0.13	
0.89		0.52										1.10	
0.67			0.19									0.46	0.75
1.47			0.78									1.58	
1.96			0.56									-0.02	
1.09			0.35									0.49	0.03
1.38			0.58									-2.20	-3.38
0.80				0.37								-0.14	
1.12				0.25								-0.96	
1.44					0.49							-0.11	
0.95					0.27							0.60	
1.08					0.20							0.19	
0.74						0.49						0.52	
1.15						0.25						0.32	
1.30						0.19						1.22	
0.71							0.00					-0.31	
0.81							0.33					0.69	
1.19							0.14					0.20	
1.22							0.22					-0.71	
1.08								0.12				1.50	
1.49								0.38				1.01	
0.80								0.40				-0.76	-1.40
0.66									0.32			-0.54	
1.42									0.20			0.68	
0.84									0.26			-0.08	
1.63									-0.02			0.30	
1.58										-0.22		-0.74	
3.21										1.60		2.63	
2.09										0.88		0.22	
1.91										0.58		-1.41	
0.93										0.46		-1.75	
1.22											0.48	-0.98	
1.43											0.49	2.33	

Table B7. *Bifactor-GPC marginal parameters for Science.*

ag	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	b1	b2
1.09	0.32											-0.19	
0.77	0.53											1.58	
0.53	0.52											-0.05	
1.00	0.38											0.71	
1.57		0.26										0.29	
1.00		0.32										0.13	
0.85		0.46										-1.30	
0.66			0.18									-0.69	-1.13
1.34			0.59									-1.18	
1.86			0.37									0.01	
1.07			0.29									-0.46	-0.03
1.30			0.45									1.69	2.59
0.78				0.34								0.18	
1.11				0.21								0.86	
1.39					0.37							0.08	
0.94					0.24							-0.64	
1.08					0.17							-0.17	
0.72						0.44						-0.72	
1.14						0.21						-0.28	
1.29						0.15						-0.95	
0.71							0.00					0.43	
0.79							0.29					-0.87	
1.19							0.12					-0.17	
1.21							0.18					0.58	
1.08								0.10				-1.40	
1.46								0.29				-0.69	
0.78								0.36				0.98	1.80
0.65									0.30			0.84	
1.41									0.16			-0.48	
0.83									0.23			0.09	
1.63									-0.02			-0.18	
1.56										-0.16		0.48	
2.34										0.75		-1.12	
1.86										0.56		-0.12	
1.80										0.39		0.78	
0.89										0.40		1.95	
1.17											0.39	0.84	
1.37											0.37	-1.70	

Investigation of Measurement Invariance According to Home Resources: TIMSS 2015 Mathematical Affective Characteristics Questionnaire

Derya Cakici Eser ^{1,*}

¹Ankara Music and Fine Arts University, Faculty of Music and Fine Arts Education, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: Oct. 27, 2020

Revised: May 04, 2021

Accepted: June 07, 2021

Keywords:

Measurement invariance,
Socio-economic status,
Home resources,
Affective characteristics.

Abstract: This study aimed to examine the measurement invariance of the mathematical affective characteristics model obtained from TIMSS 2015 4th grade Turkey administration according to home resources. For this purpose, firstly, the factor structure of the mathematical affective characteristics questionnaire was examined by explanatory factor analysis and Velicer's maximum average partial (MAP) test. It was revealed that the questionnaire had three factors. Then the structure was validated by confirmatory factor analysis. In the next stage, multi-group confirmatory factor analysis was employed with a purpose to examine whether the model displayed measurement invariance across the variables of home resources such as internet connection, heating system, cooling system, and dishwasher. The results showed that the strict measurement invariance of the mathematical affective characteristics model was achieved among the subgroups of each of the internet connection, heating system, cooling system, and dishwasher variables. Accordingly, means, variance, covariances, and item residual variances in the subgroups were found to be similar. According to the results of the study, the comparison of the mathematical affective characteristics model based on the home resources is found to be significant and comparisons made show that possible differences arise from the relevant home resource.

1. INTRODUCTION

Exams provide various information to education stakeholders depending on the purpose of exams being administered at all levels of education. Accordingly, based on the information obtained from the exams, information is acquired on such points as the current situation of students in terms of their relevant characteristics, their need for support, the efficiency of the education programs pursued, and whether the educational methods used meet the needs. In addition, curriculum improvements in education are determined with the comparisons made based on the exam results.

Since the late 1990s, education systems of countries have been compared while student achievements through exams have aimed at specific areas, targeting at specific audience with the participation of many countries. Program for International Student Assessment (PISA),

*CONTACT: Derya CAKICI ESER ✉ deryacakicieser@gmail.com 📍 Ankara Music and Fine Arts University, Faculty of Music and Fine Arts Education, Department of Educational Sciences, Ankara, Turkey

e-ISSN: 2148-7456 /© IJATE 2021

Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) are the leading exams to such ends. PISA is a program administered by the OECD every three years and it focuses on 15-year-old students' reading skills, mathematics literacy, and science literacy. TIMSS and PIRLS are programs run by International Association for the Evaluation of Educational Achievement (IEA). PIRLS is an exam that has been held every 5 years since 2001 to measure the reading skills of 4th grade students. TIMSS is a student achievement research program administered every 4 years for 4th and 8th grade students. TIMSS makes it possible to determine the academic success of students, to direct the change over time, to compare one country's situation with those of other countries, and to monitor the results of the attempts in order to increase the level of success with questions prepared in the field of mathematics and science skills (Ministry of National Education [MNE], 2016).

In TIMSS, the education system is managed together with all its stakeholders and components. Accordingly, in practice, there are questionnaires for the home environment thought to have an effect on the upbringing of children and on their success in school, in the school environment as the determinant of efficiency in achieving educational goals, and in the classroom environment where most of the learning and teaching take place. Besides, since many studies in the literature reveal the relationship between student achievement and student attitude, IEA also makes use of questionnaires to determine attitudes towards mathematics and science in TIMSS administration (IEA, 2019).

Each measurement tool is basically developed with the assumption that it measures the same feature in every group in which it is administered. However, in practice, the results might differ depending on the groups they are administered. Accordingly, results may not have equal/equivalent psychometric qualities and therefore, it would be inaccurate to generalize the results for groups (Başusta & Gelbal, 2015). For these reasons, measurement tools administered in different groups should measure the same construct in each subgroup. If it is shown that the factor loadings, inter-dimensions correlations, and error variances of a measurement model are the same in each group, it indicates that the measurement tool has the same structure in different groups (Jöreskog & Sörbom, 1993). In this context, measurement invariance can determine whether a measurement tool measures in the same way in different groups or not. With measurement invariance studies, researchers obtain evidence about whether or not scales measure the same construct in subgroups (Cheung & Lau, 2012; Millsap & Olivera-Ogilar, 2012; Cheung & Rensvold, 2002). Accordingly, it is stated that measurement tools that are not invariant across groups measure different characteristics in subgroups after the measurement invariance study. This is a validity problem for the measurement tool, and after such a measurement process, mathematical relations between the measurement tool variables will be different in each group. Interpretations regarding the results of group comparisons based on such a measurement tool would also be incorrect (Vandenberg & Lance, 2000). On the other hand, if it is shown that a measurement tool is invariant across groups, in other words, if it is shown that the mathematical relations between its variables are equal between groups, two types of validity proofs are obtained based on the measurement tool. These are (1) proof of construct validity in terms of showing that the measurement tools are used to measure the same structure in each group, (2) proof of the external validity in terms of statistically proving that the results of the comparison across groups can be generalized. In this respect, considering the vital importance of obtaining measurement invariance in interpreting the findings of a study, group comparisons made without demonstrating measurement invariance should be approached with suspicion. Hence, along with the definition of measurement invariance, its theoretical foundations, and how to test it need to be briefly explained.

1.1. Measurement Invariance

Measurement invariance is whether the measurement tool employed corresponds to the same meaning in individuals in different groups. The fact that individuals in different populations but in the same condition in terms of measured constructs get the same observed score in a test means that the measurement is invariant. If the individuals are identical in terms of the measured construct but their scores differ, the test violates the assumption of measurement invariance (Schmitt & Kuljanin, 2008). If measurement invariance is not demonstrated, the results of the comparisons across groups cannot be interpreted with certainty. It cannot be known whether the resulting difference can be attributed to a real attitudinal difference or to the difference in psychometric responses to scale items. Although this point is not instantly obvious, it is a very critical point (Cheung & Rensvold, 2002; Horn & McArdle, 1992). For these reasons, it is important to examine measurement invariance before comparing measurements obtained from two or more groups.

Confirmatory factor analysis (CFA) is one of the methods used to test measurement invariance (Kline, 2011; Schmitt & Kuljanin, 2008; Jöreskog & Sörbom, 1993). Under structural equation modelling, measurement invariance is tested using a series of tests through multi-group confirmatory factor analysis (MGCFA). By using MG-CFA in different ways with various constraints, measurement invariance is tested in four stages in a hierarchical manner (Vandenberg & Lance, 2000, Meredith, 1993).

Configural invariance comes first in the hierarchical order of measurement invariance. Configural invariance is that the construct in the measurement tool is the same across groups. If configural invariance is achieved, it can be concluded that the items in the measurement tool measure the same construct in the groups in which the invariance is investigated (Vandenberg & Lance, 2000). Configural invariance is also called as baseline model. This model reveals that the number of factors in each group and the variables that make up the factors are the same (Millsap & Olivera-Ogilar, 2012). If configural invariance is not achieved, measurement invariance will not be ensured at other stages (Kline, 2011).

When it is shown that configural measurement invariance is achieved, the metric invariance test can be conducted (Cheung & Rensvold, 2002). Metric invariance is also known as weak invariance (Meredith, 1993) or pattern invariance (Millsap, 2011). In this phase of invariance, the answer to the question of "Do common factors mean the same in all groups?" is sought (Gregorich, 2006). In the metric invariance analysis, the invariance of the factor structure of the model and the factor loadings of the items in the model in different groups are tested. While the factor variance of all groups is fixed to one in the configural model, the factor variance of the group selected as a reference in the metric model is fixed to one, and the factor variance restriction of other groups is removed (Millsap & Olivera-Ogilar, 2012). If metric invariance is achieved, the results obtained by quantitative group compares of factor variances and covariances are defensible (Gregorich, 2006).

Once metric invariance is achieved, the scalar invariance test follows it. Scalar invariance test consists of a combination of metric invariance and item intercepts invariance (Millsap & Olivera-Ogilar, 2012). Meredith (1993) called scalar invariance strong factorial invariance. At this stage of invariance, an answer to the question "Is it reasonable to compare group means?" is sought (Gregorich, 2006). In scalar invariance, the factor means of the reference group are set to zero. The means of the other groups are not constrained (Millsap & Olivera-Ogilar, 2012). If it is proven that factor loadings and item interceptions are invariant in groups, in other words, if scalar invariance is achieved, group differences estimated based on factor means are neutral. Also, group differences between observed scores are directly related to factor means (Gregorich, 2006).

Scalar invariance is followed by testing strict invariance or strict factorial invariance (Meredith, 1993). At this stage, the aim is to prove the invariance of item residual variances in addition to those whose invariance was proven in previous stages (Gregorich, 2006). Only factor means and factor covariance matrices are released in the analysis (Millsap & Olivera-Ogilar, 2012). By demonstrating that of strict invariance, which is a difficult invariance stage in practice, measurement invariance is fully ensured.

Comparisons across groups based on large-scale exams will only be reasonable when all four stages of measurement invariance given above are met. TIMSS is an exam that makes sure to obtain a very large data set and also to enable longitudinal evaluations, since it is conducted at two levels of education (4th and 8th grades) and repeated every four years. In order to show that the obtained findings are unbiased and accurate, research studies are needed to ensure the measurement invariance across the groups. In this particular study, the aim was to examine the measurement invariance of the mathematical affective characteristics questionnaire in the TIMSS 2015 4th grade assessment according to home resources.

Home resources have been defined as one of the indicators of socio-economic status (SES), pointing to facilities such as books, computers, study rooms, and educational resources (Sirin, 2005). SES refers to the position of an individual or a family in a hierarchy according to access to welfare, power, and social status (Gustafsson, Nilsen & Hansen, 2018). Parental income, parental education, parental occupation, and home resources are four indicators of SES (Sirin, 2005). White (1982) analysed approximately 200 studies investigating the relationship between SES and academic achievement in his meta-synthesis study. In his study White (1982) reported that the relationship between SES and academic achievement points to a weak relationship ($r=0.22$) contrary to expectations. Sirin (2005) replicated the study of White (1982) about 23 years later. Sirin (1982) conducted a meta-analysis on the studies on SES and academic achievement between the years of 1990 and 2000. According to the results, contrary to White (1982), studies conducted in the following years showed that the intensity of the relationship between SES and academic achievement grew and the value of the correlation changed from medium to high. When the meta-analysis studies are evaluated together, it can be said that the relationship between SES and academic achievement has become stronger in the following years. There are many studies in the literature that examine the relationships between home resources which are under the scope of SES and cognitive and affective characteristics (Yıldırım, 2019; Acar Güvendir, 2017; Bofah & Hannula, 2017; Caponera & Losito, 2016; Bouhlila, 2014; Walzeburg, 2014; Azina & Halimah, 2012; Shen, 2005). However, these past studies do not contain evidence that measurement invariance between groups is achieved. In the invariance studies in the literature, invariance was examined across genders, regions, cultures, and testing language (Kıbrıslıoğlu, 2015; Polat, 2015; Uyar & Doğan, 2014; Segeritz & Pant, 2013; Marsh et al., 2006; Erikan & Koh, 2005). Reviews based on home resources are not available. As in all group-based difference studies, researchers should test measurement invariances before performing studies based on SES variables and demonstrate that test invariance is ensured. Apart from these, measurement invariance studies based on large-scale tests in the literature have mainly investigated the invariance of student achievement/performance (Ölçülüoğlu & Çetin, 2016; Aliverinini, 2011; Teo, 2010; Wu, Li, & Zumbo, 2007), but studies investigating the invariance of affective characteristics are relatively few (Ertürk & Erdiñ-Akan, 2018; Polat, 2019). The present study is important because it focuses on the invariance of the affective characteristics of students towards mathematics and also investigates invariance according to home resources, which is related to education but has not been addressed before.

2. METHOD

In this section, the population and sample of the study are defined; data collection and data analysis are discussed.

2.1. Population and Sample

In TIMSS, the population consists of the participating country's 4th and 8th grade students, and the sample consists of the students who took the exam. Students who take the exam are determined in two stages. Accordingly, in the first stage, the schools are selected by the stratified random sampling method, and in the second stage, the classes that will participate in TIMSS are selected by the random sampling. Since the data obtained from the last administration of TIMSS in 2019 had not been released yet at the time of this particular study, this study was based on the 2015 administration and was limited to the 4th grade level. Within this scope, 6.456 students participated in TIMSS 2015 4th grade from Turkey. Because of multivariable statistical analysis based on assumptions, in this research a data screening and cleaning phase was carried out. At the end of this phase 331 cases were cleaned and the sample of this study consisted of the remaining 6.125 students.

2.2. Data Source

The data were obtained from the database at <https://timssandpirls.bc.edu/>. In the TIMSS administration, student, teacher, school, and house questionnaires are included in addition to mathematics and science achievement tests. In the student questionnaires, students are asked for information such as gender, date of birth, place of birth of their parents, and home resources. In addition, the student questionnaire includes items that examine affective characteristics regarding mathematics and science.

In the TIMSS 2015 4th grade administration, questions regarding 11 home resources in the form of yes/no answers were asked to students. While 7 of the 11 items predict the same home resources in all countries, 4 of them are constructed according to the structure of each country as a country-specific indicator of wealth. Accordingly, the first seven items consist of questions such as whether students have their own room, desk, and PC/tablet. In the next four items, the existence of financial opportunities such as having a piano at home, having a swimming pool, or having water running from the tap is investigated according to the welfare level of the country. The heating system, cooling system, washing machine, and dishwasher facilities were asked as welfare indicators in the 2015 Turkey administration. In the study, it was observed that the number of students who do not have a washing machine (n=295) was significantly smaller than those who have a washing machine (n=6,073). For this reason, the washing machine, which is one of the country-specific indicators, was not included in the study. Also, students in all participating countries were asked whether they have an internet connection at home. Internet connection at home is considered important in accessing educational technologies and educational resources, so it was decided to be included in the study. As a result, the study was carried out based on the four home resources included in the TIMSS 2015 4th grade Turkey administration. The names and definitions of the variables included in the study are given in [Table 1](#).

Table 1. Names and definitions of home resources variables.

Variable name	Variable definition
ASBG05E	Internet connection
ASBG05H	Heating System
ASBG05I	Cooling System
ASBG05K	Dishwasher

In TIMSS 2015 4th grade administration, there are 28 items as scored based on 5-point Likert type related to affective characteristics towards mathematics. These items are organized under three question themes in a test form; namely, mathematics lesson, learning mathematics, and mathematics. Home resources used in the study and affective characteristics data regarding mathematics are included in the file named ASGTURM6.

2.3. Data Analysis

The analyses of the research were carried out in various stages. Accordingly, in the first stage, the data were examined in order to test the assumptions. Individuals whose relevant home resources responses were missing were excluded from the study. The missing data in the responses to the affective items were analysed with missing data analysis, and it was observed that the values obtained were less than 5% and were randomly distributed. Missing data were completed by the item means method. For determining multivariate outliers Mahalanobis distances were examined. Accordingly, it was seen that there was no Mahalanobis value exceeding the critical chi-square value at $p < 0.001$. Descriptive statistics' examination showed the variables normally distributed. Skewness and kurtosis values were in expected range. Tavşancıl (2005) stated Bartlett's test of sphericity can be used for normality, and it was found that chi-square was 27293.564 and $p < 0.00$. This value shows that the data have a multivariate normal distribution. Tolerance, VIF, and condition index (CI) were examined for multicollinearity. Accordingly, the tolerance was found to be =1.00, VIF <5 and CI <30, and it was observed that there was no multi-collinearity problem in the data set (Tabachnick & Fidell, 2007). All these results show that factor analysis is applicable to the data.

After examining the assumptions, exploratory factor analysis was performed. Accordingly, 28 items asked in relation to mathematical affective characteristics were analysed. As a result of the analysis, KMO value was obtained as 0.930. This value is interpreted as perfect and means that the sample size is sufficient for factorability (Tabachnick & Fidell, 2007). According to the results obtained as a result of EFA, the items are collected under three factors with eigenvalues greater than 1. Accordingly, the first three eigenvalues are respectively 6.10; 1.72 and 1.49. Eigenvalues are 0.63 and less from the fourth factor. Based on these results, it can be stated that the items are grouped under three dimensions. When factor loadings are examined, it can be seen that items that have factor loadings in more than one dimension and whose difference between factor loadings are 0.1 or less are accepted as overlapping (Büyüköztürk, 2009). Accordingly, 3 of the 28 items (ASBM01B, ASBM03A and ASBM03D) were excluded from the data set because they were overlapping. As a result, a 3-dimensional structure that accounted for 36.006% of the total variance was obtained. Accordingly, there are 8 items in the dimension called liking mathematics, and the factor loadings of the items vary between 0.309 and 0.790. There are 10 items in the second dimension, called interest in mathematics. Factor loadings of the items in this dimension range from 0.314 to 0.590. In the third dimension, which is called self-confidence in mathematics, there are 7 items and the factor loadings of the items vary between 0.350 and 0.669. Table 2 contains the statistics of the structure reached as a result of EFA.

In addition to EFA, Velicer's maximum average partial (MAP) analysis was used to decide the number of factors. MAP results are included in Table 3. When Table 3 is examined, it is seen that the smallest average squared correlation takes the lowest value in the fourth step. The number of steps up to the fourth step gives the number of factors and it is seen that the number of dimensions according to the TR^2 value is three. O'Connor (2000) stated that the fourth power of partial correlation is an effective criterion. Accordingly, when the TR^4 value is examined, it is seen that it takes its smallest value in the fourth step. In this regard, the TR^4 value shows that the number of dimensions is three. Finally, when EFA and MAP results are evaluated together,

it can be stated that the results support each other and the affective characteristics for mathematics has a three-factor structure.

Table 2. *Questionnaire items, factor loadings and factors.*

Item code	Item	Liking Mathematics	Interest in Mathematics	Self-Confidence in Mathematics
ASBM01A	I enjoy learning mathematics	.790		
ASBM01C	Mathematics is boring	.488		
ASBM01D	I learn many interesting things in mathematics	.309		
ASBM01E	I like mathematics	.798		
ASBM01F	I like any schoolwork that involves numbers	.538		
ASBM01G	I like to solve mathematics problems	.630		
ASBM01H	I look forward to mathematics lessons	.648		
ASBM01I	Mathematics is one of my favorite subjects	.702		
ASBM02A	I know what my teacher expects me to do		.314	
ASBM02B	My teacher is easy to understand		.424	
ASBM02C	I am interested in what my teacher says		.498	
ASBM02D	My teacher gives me interesting things to do		.327	
ASBM02E	My teacher has clear answers to my questions		.590	
ASBM02F	My teacher is good at explaining mathematics		.510	
ASBM02G	My teacher lets me show what I have learned		.476	
ASBM02H	My teacher does a variety of things to help us learn		.499	
ASBM02I	My teacher tells me how to do better when I make a mistake		.576	
ASBM02J	My teacher listens to what I have to say		.587	
ASBM03B	Mathematics is harder for me than for many of my classmates			.669
ASBM03C	I am just not good at mathematics			.692
ASBM03E	Mathematics makes me nervous			.577
ASBM03F	I am good at working out difficult mathematics problems			.350
ASBM03G	My teacher tells me I am good at mathematics			.377
ASBM03H	Mathematics is harder for me than any other subject			.661
ASBM03I	Mathematics makes me confused			.631

Table 3. Eigen Values Regarding Partial Correlations Obtained from the MAP Test.

	TR ²	TR ⁴		TR ²	TR ⁴		TR ²	TR ⁴
0	0.06555	0.00991	9	0.03196	0.00768	18	0.17084	0.07851
1	0.01779	0.00076	10	0.03951	0.01096	19	0.18635	0.07830
2	0.01421	0.00056	11	0.05099	0.01526	20	0.20042	0.08979
3	0.00842*	0.00025**	12	0.06017	0.01774	21	0.26284	0.14586
4	0.01010	0.00040	13	0.06642	0.01780	22	0.38916	0.26106
5	0.01249	0.00108	14	0.07968	0.02544	23	0.54993	0.42485
6	0.01594	0.00296	15	0.09420	0.03611	24	100.000	100.000
7	0.01969	0.00448	16	0.11363	0.04385			
8	0.02518	0.00684	17	0.13486	0.05890			

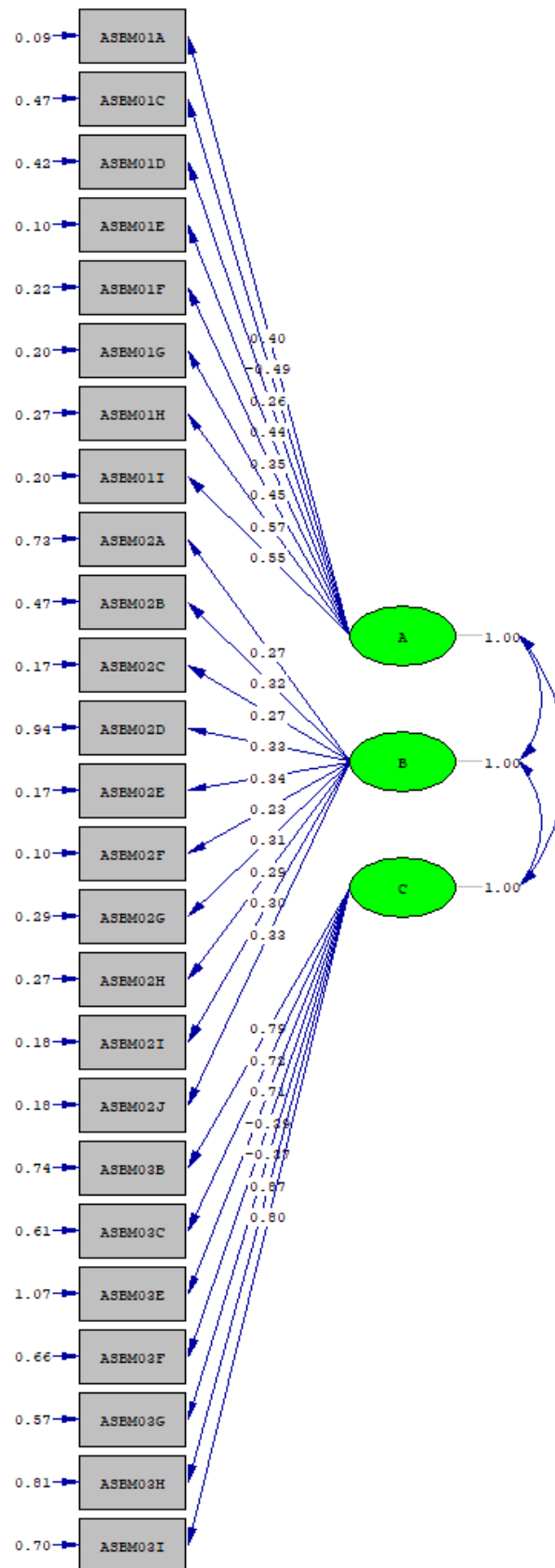
* The smallest average squared correlation

**The smallest average squared correlation's 4th power

Confirmatory factor analysis (CFA) was performed with Lisrel 8.24 to verify the model by EFA. EFA was run with ML algorithm and results showed there is no need for modifications. In the established model, it was found that $\chi^2 = 2605.21$, $df = 272$, $\chi^2/df = 9.57$. In the model established with CFA, the χ^2/df ratio is expected to be ≤ 3.00 . However, χ^2 statistics is sensitive to sample size, and as the sample size increases, this ratio exceeds 3 (Kline, 2011). Therefore, the model χ^2/df value obtained was not interpreted as a model-data misfit and other fit statistics were examined. Accordingly, it was found that RMSEA=0.056, SRMR=0.052, CFI=0.96, and NNFI=0.96. Since all of these values indicated good fit, it was concluded that the model was validated (Kline, 2011). Correlations between factors are $r_{12}=0.54$; $r_{23}=-0.53$ and $r_{13}=-0.34$. The path diagram for the model is presented in Figure 1.

After the mathematical affective characteristics model was verified, measurement invariance tests were carried out. Accordingly, the data set was analysed by MGCFA in configural, metric, scalar, and strict invariance stages separately for each home resources variable. The values of fit statistics, χ^2 , df , RMSEA, SRMR, NNFI and CFI were examined in each invariance stage. In addition, the ΔCFI values revealing the change in the CFI in the transition from the unconstrained model to the constrained model were examined in order to decide if invariance was achieved. In the literature, measurement invariance is examined according to the chi-square difference test and the difference in CFI. In various studies, the lack of significance of chi-square has been shown as evidence for measurement invariance (Hirschfeld & von Brachel, 2014; Brannick, 1995; Kelloway, 1995). However, as the chi-square is sensitive to the sample size, it tends to be significant in large samples. This situation is also valid for this study. Similarly, Cheung and Rensvold (2002) stated that the ΔX^2 test is sensitive to the sample size, the complexity of the model and is less effective in making practical decisions. Cheung and Rensvold (2002) examined 20 different fit indices in their study and stated that the strongest statistics to be examined in the test of intergroup invariance are ΔCFI , Δ Gamma line, and Δ McDonald's NCI. For these reasons, in making the decision about measurement invariance, it is taken as a reference whether the $|\Delta CFI|$ is <0.01 or not as stated by Wu, Li and Zumbo (2007).

Figure 1. Path diagram for mathematical affective characteristics model.



Chi-Square=2827.10, df=272, P-value=0.00000, RMSEA=0.056

3. FINDINGS

The MGCFA method was used for the measurement invariance test with Lisrel 8.54 in the study. In the analysis EM algorithm and covariance matrix were used. However, the validation of the model was tested first in each of the subgroups where invariance would be examined. Accordingly, the mathematical affective characteristics model was validated separately for two groups (according to the responses “yes, I have” and “no, I haven’t”) of the internet connection variable. The same procedure was carried out for the heating system, cooling system, and dishwasher variables. The fit statistics for the model verified in the groups created based on each variable are given in Table 4.

Table 4. CFA fit statistics of the groups based on home resources variables.

Variable	Group	χ^2	df	χ^2/df	RMSEA	SRMR	NNFI	CFI
Internet connection	Yes	3322.49	272	12.215	0.056	0.052	0.96	0.96
	No	2498.51	272	12.862	0.057	0.055	0.94	0.95
Heating system	Yes	2779.62	272	10.208	0.055	0.052	0.96	0.96
	No	2777.77	272	10.212	0.057	0.055	0.95	0.95
Cooling system	Yes	2313.34	272	8.504	0.057	0.055	0.95	0.96
	No	3443.07	272	12.658	0.055	0.053	0.95	0.96
Dishwasher	Yes	3968.37	272	14.589	0.056	0.052	0.96	0.96
	No	1804.79	272	6.635	0.057	0.058	0.94	0.95

According to Table 4, the χ^2/df value was found to be >3 in the model established for each subgroup of variables. Since the χ^2 statistics is sensitive to the sample size, χ^2 , df, χ^2/df were reported in the following phases of the study, but other statistics were taken as a basis to decide if the model was validated. In Table 4, from fit statistics, it was found that RMSEA was <0.06 , SRMR <0.08 , and NNFI >0.90 , and this corresponds to a good fit; also that CFI ≥ 0.95 corresponds to a perfect fit (Hu & Bentler, 1999; Klein, 2011). These results show that the model is validated in the subgroups of each home resources variable. After the model was verified separately in each subgroup, measurement invariance analyses were initiated.

3.1. Measurement Invariance According to Internet Connection Variable

Whether or not there is an internet connection at home is one of the common questions asked regarding home resources in all countries. In Turkey, 58.4% of students ($n=3576$) have an internet connection at home and the remaining 41.6% ($n=2549$) do not. The results of measurement invariance across groups concerning students with and without internet connection are presented in Table 5.

Table 5. Measurement invariance according to internet connection.

Invariance type	χ^2	df	χ^2/df	RMSEA	SRMR	NNFI	CFI	ΔCFI
Configural	5821.00	544	10.700	0.056	0.055	0.95	0.96	-
Metric	5903.51	566	10.430	0.056	0.054	0.96	0.96	0.00
Scalar	6030.11	575	10.487	0.056	0.060	0.95	0.96	0.00
Strict	7082.32	597	11.863	0.060	0.061	0.95	0.95	-0.01

According to Table 5, it is seen that RMSEA is <0.08 , SRMR <0.08 , NNFI ≥ 0.95 , and CFI ≥ 0.95 . ΔCFI was calculated as 0.00 when changing from configural to metric, from metric to scalar and it became -0.01 when switching from scalar to strict. Based on model fit indexes and ΔCFI , the mathematical affective characteristics model ensures all stages of measurement invariance across groups of internet connection variable. According to this result, the factor

structure, item factor loadings, item constants, and error variances of mathematical affective characteristics do not differ depending on whether there is an internet connection at home or not. According to this result, mathematical affective characteristics can be compared significantly concerning the internet connection variable and it can be concluded that the possible differences are due to internet connection.

3.2. Measurement Invariance According to The Heating System Variable

One of the country-specific indicators of wealth concerning home resources is heating systems in TIMSS 2015 4th grade Turkey administration. Accordingly, 49.1% (n=3010) of the participating students have a heating system in their houses, whereas 50.9% (n=3115) do not. The findings regarding the invariance of the mathematical affective characteristics model across the sub-groups of the heating system are given in Table 6.

Table 6. Measurement invariance according to heating system.

Invariance type	χ^2	df	χ^2/df	RMSEA	SRMR	NNFI	CFI	ΔCFI
Configural	5829.38	544	10.716	0.056	0.052	0.95	0.96	-
Metric	5897.45	566	10.420	0.055	0.054	0.95	0.96	0.00
Scalar	6009.45	572	10.506	0.056	0.058	0.95	0.96	0.00
Strict	6857.99	597	11.487	0.059	0.061	0.95	0.95	-0.01

According to the results in Table 6, error indices for RMSEA were found to be <0.08 and for SRMR <0.08; and fit indices for NNFI and CFI were obtained as ≥ 0.95 for all invariance types. ΔCFI was calculated as 0.00 when changing from configural to metric, from metric to scalar, it takes -0.01 value when switching from scalar to strict invariance. These values obtained are within the accepted range indicating that invariance is achieved. The established model ensures all stages of measurement invariance in subgroups of the heating system. Accordingly, the factor structures, factor loadings, regression constants, and error variances obtained in both groups are equal. The differences of mathematical affective properties according to the heating system can be examined and the differences can be explained on the basis of home resource addressed.

3.3. Measurement Invariance According to Cooling System

In TIMSS 2015 Turkey administration, a cooling system is one of the home resources asked as a country-specific indicator of wealth. Students who have air conditioner-like devices as a cooling system account for 37.3% (n=2286) of all participants, and those who do not have a cooling system such as an air conditioner account for 62.7% (n=3839). Findings regarding the mathematical affective characteristics model invariance across groups based on the cooling system are given in Table 7.

Table 7. Measurement invariance according to cooling system.

Invariance type	χ^2	df	χ^2/df	RMSEA	SRMR	NNFI	CFI	ΔCFI
Configural	5756.41	544	10.582	0.056	0.055	0.95	0.96	-
Metric	5811.24	566	10.267	0.055	0.055	0.96	0.96	0.00
Scalar	5857.85	572	10.241	0.055	0.065	0.96	0.96	0.00
Strict	6123.63	597	10.257	0.055	0.068	0.96	0.96	0.00

According to Table 7, RMSEA and SRMR were found to be <0.08, NNFI and CFI were found to be ≥ 0.95 . Since CFI was 0.96 in all invariance models, all values of ΔCFI were equal to 0.00. When the model fit statistics are evaluated together, it is seen that the mathematical affective characteristics model is invariant based on the groups of the cooling system.

Accordingly, it was shown that the factor structure, factor loadings, regression constants, and error variances of the mathematical affective characteristics model were equal in the two groups. Therefore, the mathematical affective characteristics model can be significantly compared and interpreted according to the cooling system variable.

3.4. Measurement Invariance According to The Dishwasher Variable

Dishwasher was considered a country-specific indicator of wealth in TIMSS 2015 4th grade Turkey administration. Accordingly, 71.4% (n=4376) of the participating students had a dishwasher at home, whereas 28.6% (n=1749) did not. The measurement invariance results of the mathematical affective characteristics model based on the dishwasher variable are presented in Table 8.

Table 8. Measurement invariance according to the dishwasher variable,

Invariance type	χ^2	df	χ^2/df	RMSEA	SRMR	NNFI	CFI	Δ CFI
Configural	5773.16	544	10.612	0.056	0.052	0.95	0.96	-
Metric	5860.76	566	10.355	0.055	0.052	0.96	0.96	0.00
Scalar	5955.19	572	10.411	0.055	0.054	0.96	0.96	0.00
Strict	7297.59	597	12.224	0.061	0.055	0.95	0.95	-0.01

According to Table 8, RMSEA and SRMR values were <0.08; fit statistics NNFI and CFI were ≥ 0.95 . Δ CFI was calculated as 0.00 when changing from configural to metric, from metric to scalar, and as -0.01 when changing from scalar to strict. When the statistics in Table 8 are evaluated together, the mathematical affective characteristics model ensures measurement invariance across the dishwasher-based groups. The factor structure, factor loadings, regression constants, and error variances of the model are identical across groups. Mathematical affective characteristics can be meaningfully compared and interpreted based on the dishwasher.

4. DISCUSSION and CONCLUSION

This study investigated whether the mathematical affective characteristics model proposed based on the TIMSS 2015 4th grade Turkey administration showed measurement invariance according to home resources or not. As a result of the study, it was shown that the variables of internet connection, heating system, cooling system, dishwasher, which are considered within the scope of home resources, provide configural, metric, scalar, and strict measurement invariance across the subgroups, respectively. Accordingly, the means, variances, covariances, and item residual variances of the model are identical across the subgroups of each established home resource. The results indicate that the mean of observed scores obtained from the mathematical affective characteristics scale can be compared according to home resources. The results of further research to be obtained by making comparisons are meaningful and possible differences can be attributed to the relevant home resource.

Although there is no similar study in the literature examining measurement invariance based on home resources, there are various measurement invariance studies based on large-scale exams. One of these is the study by Hansson and Gustafsson (2013), which examined whether or not the socio-economic status is invariant according to the ethnic structure using TIMSS 2003 data. In their study, Hansson and Gustafsson (2013) tested the invariance of the latent SES variable between Swedish and non-Swedish groups and found that configural invariance was achieved, but scalar invariance was not. Ertürk and Erdiñç-Akan (2018) and Polat (2019) focused on the mathematical affective characteristics questionnaire of the TIMSS 2015 administration in their studies. Accordingly, Ertürk and Erdiñç-Akan (2018) examined the gender-based measurement invariance of variables related to mathematics achievement based on the 4th grade administration. According to the results of the study, they found that the liking

mathematics scale provides strict invariance, and interest in mathematics and mathematical self-confidence scales provide configural invariance. Polat (2019) investigated the invariance of both mathematical as well as the cultural affective characteristics questionnaire according to cultures (Turkey, Singapore, and Saudi Arabia) and regions (NUTS-Level 1), and gender based on the TIMMS 2015 8th grade administration. The study showed that the established mathematical and science affective characteristics models provide scalar invariance across cultures and regions and strict invariance across genders.

Some of the invariance studies carried out based on PISA are the studies by Kıbrıslıoğlu (2015), Güngör and Kabasakal (2020), and Uyar and Uyanık (2019). Kıbrıslıoğlu (2015) investigated the invariance of the PISA 2012 mathematical learning model across cultures (Turkey, China-Shanghai, and Indonesia) and genders. The study showed that the model provides only configural invariance across cultures. The study examined gender-based measurement invariance based on all of the data set obtained from three cultures, and as a result, the study showed that the mathematical learning model provides strict invariance across genders. Güngör and Kabasakal (2020) investigated the measurement invariance of instrumental motivation and science self-efficacy scales in science teaching according to gender and regions based on PISA 2015 Turkey administration. Güngör and Kabasakal (2020) reported that only configural invariance was achieved based on gender, and metric invariance was achieved across regions. Uyar and Uyanık (2019) established a learning model for science by using the questionnaire in the PISA 2015 administration and investigated the invariance of the established model according to gender in Turkey sample and the invariance of the established model in Turkey-Singapore samples according to cultures. As a result, Uyar and Uyanık (2019) found that across genders metric invariance and across cultures configural invariance was achieved.

When the above-mentioned studies are evaluated together, it is seen that strict invariance based on gender is ensured under certain conditions in large-scale exams, and there are no studies ensuring strict invariance based on cultures. However, there are no studies carried out based on home resources in large-scale exams or SES in general that can be compared with the findings of the present study. In this regard, researchers are recommended that they investigate measurement invariance based on variables such as parental education level, parental income, number of siblings, along with other home resources not included in this study, and to address variables that ensure strict invariance in comparisons across groups.

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

Derya CAKICI ESER  <https://orcid.org/0000-0002-4152-6821>

5. REFERENCES

- Acar Güvendir, M. (2017). Determination of the relationship between the students mathematical literacy and home and school educational resources in program for international student assessment - (PISA 2012). *Mersin University Journal of the Faculty of Education*, 13(1). <https://doi.org/10.17860/mersinefd.305762>
- Alivernini, F. (2011). Measurement invariance of a reading literacy scale in the Italian context: A psychometric analysis. *Procedia Social and Behavioral Sciences*, 15, 436-441. <https://doi.org/10.1016/j.sbspro.2011.03.117>

- Azina, I. N., & Halimah, A. (2012). Student factors and mathematics achievement: Evidence from TIMSS 2007. *Eurasia Journal of Mathematics, Science and Technology Education*, 8(4), 249-255. <https://doi.org/10.12973/eurasia.2012.843a>
- Başusta, N. B., & Gelbal, S. (2015). Examination of measurement invariance at groups' comparisons: A study on PISA student questionnaire, *Hacettepe University Journal of Education*, 30(4), 80-90.
- Bouhlila, D. S. (2014). The impact of socioeconomic status on students' achievement in the Middle East and North Africa: An essay using the TIMSS 2007 database. *International Perspectives on Education and Society*, 24, 199-226 <https://doi.org/10.1108/S1479-367920140000024017>
- Bofah, E. A. T., & Hannula, M. S. (2017). Home resources as a measure of socio-economic status in Ghana. *Large-scale Assessments in Education*, 5(1), 1-15. <https://doi.org/10.1186/s40536-017-0039-5>
- Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, 16(3), 201-213. <https://doi.org/10.1002/job.4030160303>
- Büyüköztürk, Ş. (2009) *Sosyal Bilimler İçin Veri Analizi El Kitabı*, Pegem Akademi
- Caponera, E., & Losito, B. (2016). Context factors and student achievement in the IEA studies: Evidence from TIMSS. *Large-scale Assessments in Education*, 4(1),12. <https://doi.org/10.1186/s40536-016-0030-6>
- Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*, 15(2), 167-198. <https://doi.org/10.1177/1094428111421987>
- Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness of fit indices for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French version of TIMSS. *International Journal of Testing*, 5, 23-35. https://doi.org/10.1207/s15327574ijt0501_3
- Ertürk, Z., & Erdiñç-Akan, O. (2018). The investigation of the variables effecting TIMSS 2015 mathematics achievement with SEM, *Journal of Theoretical Educational Science*, 2, 14-34.
- Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, 44, 78-94. <https://doi.org/10.1097/01.mlr.0000245454.12228.8f>
- Gustafsson, J. E., Nilsen, T., & Hansen, K. Y. (2018). School characteristics moderating the relation between student socio-economic status and mathematics achievement in grade 8. Evidence from 50 countries in TIMSS 2011. *Studies in Educational Evaluation*, 57, 16-30. <https://doi.org/10.1016/j.stueduc.2016.09.004>
- Hansson, Å., & Gustafsson, J.-E. (2013). Measurement invariance of socioeconomic status across migrational background. *Scandinavian Journal of Educational Research*, 57(2), 148–166. <https://doi.org/10.1080/00313831.2011.625570>
- Hirschfeld, G., & Von Brachel, R. (2014). Improving multiple-group confirmatory factor analysis in R—A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research, and Evaluation*, 19(1), 7. <https://doi.org/10.7275/qazy-2946>
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(3), 117-144. <https://doi.org/10.1080/03610739208253916>

- Hu, L., & Bentler, M., P. (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>
- International Association for the Evaluation of Educational Achievement, [IEA], (2019). Chapter 3 TIMSS 2019 context questionnaire framework, <https://timss2019.org/wp-content/uploads/frameworks/T19-Assessment-Frameworks-Chapter-3.pdf>
- Jöreskog, K. G., & Sörbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. IL: Scientific Software International, Inc.
- Kelloway, E. K. (1995). Structural equation modelling in perspective. *Journal of Organizational Behavior*, 16(3), 215-224. <https://doi.org/10.1002/job.4030160304>
- Kıbrıslıoğlu, N. (2015). *The investigation of measurement invariance PISA 2012 mathematics learning model according to culture and gender: Turkey - China (Shanghai) - Indonesia*, [Graduate Thesis, Hacettepe University].
- Kline, R. B., (2011). *Principles and Practices of Structural Equation Modelling*. The Guilford Press.
- Marsh, H. W., Hau, K. T., Artelt, C., Boument, J., & Peschar, J. (2006). OECD's brief selfreport measure of educational psychology's most useful affective constructs: cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6 (4), 311-360. https://doi.org/10.1207/s15327574ijt0604_1
- McGaw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical and Statistical Psychology*, 24(2), 154-168. <https://doi.org/10.1111/j.2044-8317.1971.tb00463.x>
- Ministry of National Education (2016). TIMSS 2015 ulusal matematik ve fen ön raporu 4. ve 8. Sınıflar [TIMSS 2015 national mathematics and sciences preliminary report 4th and 8th grades]. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2017_06/23161945_timss_2015_on_raporu.pdf
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543. <https://doi.org/10.1007/BF02294825>
- Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle, (Ed.) *Handbook of structural equation modeling*, (pp. 380-392), Guilford.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*, Routledge.
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32, 396-402. <https://doi.org/10.3758/BF03200807>
- Ölçüoğlu, R., & Çetin, S. (2016). The investigation of the variables that affecting eight grade students' TIMSS 2011 math achievement according to regions, *Journal of Measurement and Evaluation in Education and Psychology* 7(1), 202-220. <https://doi.org/10.21031/epod.34424>
- Polat, M. (2019). *The investigation of measurement invariance of TIMSS-2015 mathematics and science affective characteristics models according to culture, gender and statistical region*, [Graduate Thesis, Hacettepe University].
- Schmith, N., & Kuljanin, G. (2008). Measurement invariance: review of practice and implication. *Human Resources Management Review*, 18, 210-222. <https://doi.org/10.1016/j.hrmr.2008.03.003>
- Segeritz, M., & Pant, H. A. (2013). Do they feel the same way about math? Testing measurement invariance of the PISA students' approaches to learning instrument across immigrant groups within Germany. *Educational and Psychological Measurement*, 73(4), 601-630 <https://doi.org/10.1177/0013164413481802>

- Shen, C. (2005). How American middle schools differ from schools of five Asian countries: Based on cross-national data from TIMSS 1999. *Educational Research and Evaluation*, 11(2), 179-199. <https://doi.org/10.1080/13803610500110810>
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3), 417-453. <https://doi.org/10.3102/00346543075003417>
- Tabachnick, B. G., & Fidell, L.S. (2007). *Using Multivariate Statistics*, (5. ed.) Pearson Education.
- Tavşancıl, E. (2005) *Tutumların Ölçülmesi ve SPSS ile Veri Analizi [Measuring Attitudes and Data Analysis with SPSS]*, Nobel Yayınları
- Teo, T. (2010). Gender differences in the intention to use technology: A measurement invariance analysis. *British Journal of Educational Technology*, 41(6), 120-124. <https://doi.org/10.1111/j.1467-8535.2009.01023.x>
- Uyar. Ş., & Doğan, N. (2014). An investigation of measurement invariance of learning strategies model across different groups in PISA Turkey sample, *International Journal of Turkish Educational Studies*, 2, 30-43.
- Vanderberg, 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(4), 4-70. <https://doi.org/10.1177/109442810031002>
- Walzebug, A. (2014). Is there a language-based social disadvantage in solving mathematical items?. *Learning, Culture and Social Interaction*, 3(2), 159-169. <https://doi.org/10.1016/j.lcsi.2014.03.002>
- White, K. R. (1982). The relation between socioeconomic status and academic achievement, *Psychological Bulletin*, 91(3), 461-481. <https://doi.org/10.1037/0033-2909.91.3.461>
- Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, 12(3). <https://doi.org/10.7275/mhqa-cd89>
- Yıldırım, S. (2019). Predicting mathematics achievement: The role of socioeconomic status, parental involvement, and self-confidence, *Education and Science*, 44(198), 99-113. <https://doi.org/10.15390/EB.2019.7868>

Examination of Common Exams Held by Measurement and Assessment Centers: Many Facet Rasch Analysis

Gulden Kaya Uyanik¹, Tugba Demirtas Tolaman^{2,*}, Duygu Gur Erdogan²

¹Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Turkey.

²Sakarya University, Faculty of Education, Department of Turkish and Social Sciences Education, Sakarya, Turkey.

³Sakarya University, Faculty of Education, Department of Educational Sciences, Sakarya, Turkey.

ARTICLE HISTORY

Received: May 02, 2020

Revised: May 01, 2021

Accepted: July 08, 2021

Keywords:

Turkish Course,
Common Exam,
Many Facet Rasch
Analysis,
Multiple-Choice Item,
Rater.

Abstract: This paper aims to examine and assess the questions included in the “Turkish Common Exam” for sixth graders held in the first semester of 2018 which is one of the common exams carried out by The Measurement and Evaluation Centers, in terms of question structure, quality and taxonomic value. To this end, the test questions were examined by three specialists with expertise in different fields in terms of structure, content, and taxonomic values. The test questions were then rated by raters with expertise in different fields according to the criteria set by the researchers. Hence, the study employed the descriptive survey model. The data obtained from the assessment of the questions were analyzed using the Many Facet Rasch Model (MFRM). According to the findings, of the 20 questions included in the exam, 5 (five) are in the category of “Remembering”, 12 (twelve) in the category of “Understanding”, 2 (two) in the category of “Analyzing” and 1 (one) in the category of “Evaluating.” Accordingly, the number of questions that measure higher-order thinking skills was lower than the number of lower-level questions. In addition, the study contained three facets: raters, tasks (items), and criteria. There were no differences among the raters (a Turkish Education Specialist, a Program Development Specialist, and a Testing and Assessment Specialist) in terms of severity and leniency: all the raters were in agreement. Finally, in this study, the questions met the criteria measuring the structural features, while they failed to meet the criteria measuring the quality and clarity.

1. INTRODUCTION

The new century requires raising individuals who are not passive receivers of information, (i.e., who do not only obtain information but also question it, translate information into different forms according to changing conditions, use information effectively, and develop higher-order thinking skills, such as creative thinking, critical thinking, and comparison). Hill (2016) defines higher-order thinking skills as the ability to transcend the information provided, adopt a critical stance, make evaluations, develop meta-cognitive awareness, and use problem-solving skills. For this reason, education systems aim to raise individuals equipped with the skills needed in the 21st century, starting from the preschool period. Raising individuals who can meet the

*CONTACT: Tuğba DEMİRTAŞ TOLAMAN ✉ tdemirtas@sakarya.edu.tr 📍 Sakarya University, Faculty of Education, Department of Turkish and Social Sciences Education, Sakarya, Turkey

expectations of the era, keep up with current developments, have a sense of self-confidence, research, question, and realize themselves is what is expected from modern education systems (Anil, 2009). Language skills are an important tool in the acquisition of high-level mental skills. According to Gunes (2007), language is the most important tool for learning as well as developing mental skills. In Turkey, language skills consisting of reading, writing, speaking, and listening skills are acquired and developed by students in Turkish classes. Turkish lesson is not considered a course aiming to give information, but a process aimed at helping students to acquire and develop language skills (Kurudayioglu & Cetin, 2015; Karaduz, 2010; Gunes, 2011). The four basic language skills targeted by the Turkish course are also a basis for other courses. In other words, students' success in Turkish classes in understanding, interpreting, criticizing, and evaluating what they read and making inferences is a prerequisite for success in other courses as well as for overall academic performance. As a matter of fact, according to Cer (2018), one of the most important responsibilities in developing higher-order thinking skills in children falls upon the shoulders of mother tongue programs.

Constructivism-based curricula that have been implemented in Turkey since 2006 target not only basic language skills but also higher-order thinking skills. For example, Turkish Course Curriculum designed by the Republic of Turkey Ministry of National Education (2006) includes higher-order thinking skills such as critical thinking, creative thinking, problem-solving, and decision making. Also, in the revised Turkish Course Curriculum (2019), the structure and hierarchy of the learning objectives have been arranged in a way that contributes to the development of students' high-level cognitive skills as well as basic language skills. The curriculum aims to help students develop skills such as researching, exploring, interpreting, and constructing knowledge as well as accessing information from printed materials and multimedia sources and organizing, questioning, using, and producing information. In addition, it is aimed to help students understand, evaluate, and question what they read from a critical perspective.

All these skills are planned to be conveyed to students through the learning objectives specified in the curriculum. On the other hand, testing and assessment, which reveals whether the stated learning objectives are achieved by students, is carried out by teachers through classroom activities. In addition, the Ministry of National Education or Student Selection and Placement Center conducts testing and assessment on a national and local scale in order to place students in a higher education institution (Kardes-Birinci, 2014; Cepni, Ozsevgenc & Gokdere, 2003). Furthermore, to evaluate the Turkish education system according to international criteria, Turkey has participated in the Programme for International Student Assessment (PISA), a worldwide study by the Organisation for Economic Co-operation and Development (OECD).

Recent years have witnessed developments in cognitive, psychometric, and technological tools, concepts and theories in the assessment of education (Mislevy, 2006). One of the developments in Turkey is the "Turkish Language Test for Four Skills." This test is held to measure students' four basic language skills within the framework of the Turkey's 2023 Education Vision. The pilot implementation of the test, which was held by the Ministry of National Education to measure students' four language skills in an electronic environment and with a standard measurement tool, was carried out with the participation of 7th-grade students in 15 provinces. This test is important in that it was the first nation-wide practice to measure students' basic language skills in the mother tongue in line with international standards (Republic of Turkey Ministry of National Education, 2020).

Regardless of the level and content of education, measuring student learning throughout and at the end of the education process is a necessity (Buyukozturk, 2016). The main tools for educational measurement are tests and exams. They do not only measure students' knowledge and skills related to a particular area (Buyukozturk, 2016) but also are indicators of whether learning objectives specified in a curriculum are achieved. Downing (2006) underlines twelve

steps for effective test development: overall plan, content definition, test specifications, item development, test design and assembly, test production, test administration, scoring examination responses, establishing passing scores, reporting examination results, item banking, and test technical report. These twelve steps provide a structured, systematic process for developing effective exams/tests of all kinds.

The content definition is one of the most important steps of test development. When defining the content, a table of specifications is used. On the other hand, when developing a table of specifications, taxonomies are used. The taxonomies developed to be used in the educational field (Bloom 1956, Haladayna, 1997; Marzano & Kendall, 2007, etc.) are used not only to guide the development of curricula but also for the development of effective test questions suitable for learning outcomes and objectives. These learning taxonomies also provide standardization in education both at the national and international levels.

Revised Bloom's Taxonomy employed in this study is a revision of Bloom's Taxonomy developed by Bloom et al. in 1956. It was published in 2001 by a group of testing and assessment specialists, cognitive psychologists, curriculum theorists, and instructional researchers chaired by Lorin W. Anderson, who was once a student of Bloom (Anderson et al., 2001). Bloom's original taxonomy identified six levels within the cognitive domain, from the simple recall or recognition of facts to increasingly more complex and abstract mental levels. These six levels are (1) *knowledge*, (2) *comprehension*, (3) *application*, (4) *analysis*, (5) *synthesis* (6) *evaluation* (Anderson, 2005). On the other hand, Revised Bloom's Taxonomy contains two dimensions: knowledge dimension (factual knowledge, conceptual knowledge, procedural knowledge, and meta-cognitive knowledge) and cognitive process dimension (applying, analyzing, evaluating). Also, in the revised taxonomy, knowledge was renamed as "remembering," comprehension as "understanding," and synthesis as "creating" (Anderson, 2005). Thus, the original taxonomy was revised and provided with a structure more appropriate for the new century.

Exams/tests are administered through asking questions. The question at the center of learning is generally defined as a statement expressed to extract information from the learner (Hill, 2016). Asking and answering questions means engaging in a mental process. According to Dillion (2006), "one turns to logic, philosophy, and linguistics for analyses of the nature of questions, their relation to answers, and their function in discourse, that is, for a theory of questions." This indicates that questions are a complex but effective tool consisting of many skills.

Asking and answering questions is one of the activities/methods frequently used in communication. Since Socrates (469 BC - 399 BC), Socratic method and questions (Noddings, 2018) have been at the center of learning and teaching activities. Teachers ask questions for different purposes in their educational activities. According to Yildirim (2012), for example, questions are the most important tools to monitor student learning. The competence of teachers and students at this level has an important place in improving students' comprehension of what they read. Gunes (2012) lists the objectives of questions in Turkish teaching as motivating students, increasing their comprehension levels, helping them develop language and mental skills, and effectively conducting and evaluating the learning and teaching process. Andre (1979) argues that questions may be used in at least four different situations to guide student learning: questions can be used in classroom recitation or discussion; they can be inserted in text or other instructional media; they can be used on examinations; finally, students can ask questions of themselves while studying.

The nature of questions has a crucial impact on the progress/development of thought in the classroom. The questions asked by teachers not only define the framework of the lesson but also indicate teachers' expectations from students (Wilén, 1991). In addition, the level of

questions also affects the quality of thinking skills. According to Andre (1979), level-of-question refers to the nature of cognitive processing required to answer a question. A question may ask a learner to repeat or recognize some information exactly as it was presented in instruction. Such a question is typically referred as a knowledge, factual, or verbatim question. Factual questions are believed to involve less complex cognitive processing than questions requiring more than direct memory. Ates et al. (2016) stated that teachers tend to ask lower-level questions, that students' thinking levels are affected by the questions asked by teachers, and students do not usually ask questions at higher levels than those posed by their teachers. The authors also stated that teachers often use questions to measure and assess students' comprehension levels, rather than to improve their comprehension skills or enable them to develop higher-order thinking skills.

Also, Dillon (2006) argues that questions alone are insufficient to foster students' independent thinking and may limit their thinking abilities. To eliminate these limitations, the following methods are recommended: avoiding direct questions, confirming what is said, keeping silent (waiting). Shaunessy (2000) recommends that for students to develop creative, critical, and higher-order thinking skills, teachers should use divergent questions to provoke more questions and new inquiries rather than convergent questions that have one correct answer.

Questions are considered as one of the basic tools of thinking and are often employed in Turkish classes for different purposes. A thorough review of the relevant literature has yielded a number of studies examining the questions asked by teachers in Turkish tests, the questions included in teacher's books and workbooks, and the questions asked by teachers and students during Turkish classes. Kavruk and Cecen (2013), Cintas Yildiz (2015), Gufta and Zorbaz (2008) examined the test questions prepared by Turkish teachers, Bircan (2012), Yesilyurt (2012), and Aktas (2017) examined the test questions prepared by prospective teachers studying in the Turkish language teaching department, and Gocer (2016) examined the questions prepared by Turkish teachers enrolled in postgraduate education. Also, Cayhan and Akin (2015) examined the nation-wide TEOG (transition from primary to secondary education) test, and Demiral and Mensan (2017) compared the test questions developed by teachers with TEOG test questions and PISA test questions.

Besides, many studies have examined the questions in Turkish workbooks and teacher's books as well as reading comprehension questions included in student's books. Gocer (2008), Cecen and Kurnaz (2015) examined the questions in the measurement and evaluation sections at the end of each theme, Ozdemir et al. (2007) and Bozkurt et al. (2015) examined the questions in workbooks, Eroglu and Kuzu (2014) examined the grammar questions in workbooks, Onalan and Zengin (2015), Sarar-Kuzu (2013), and Celikturk-Sezgin and Gedikoglu-Ozilhan (2019) examined reading comprehension questions, and Durukan (2009) examined the questions in teacher's books. All these studies examined the test questions developed by Turkish teachers, the questions in Turkish textbooks, the questions in workbooks, reading comprehension questions, and the questions in teacher's books according to Bloom's Taxonomy or the Revised Bloom's Taxonomy and revealed that the examined questions do not address high-level mental skills, which is an important common finding.

1.1. Many Facet Rasch Model

The main problem of the study was answered by using the Many Facet Rasch Model (MFRM). The Rasch Model, which is a two-facet model based on Item Response Theory (IRT), is used in measurement situations where the Rasch Model is affected by different variability sources (raters, different measurement situations, etc.) other than individual and item facets. MFRM is a measurement model that can overcome the limitations of Classical Test Theory (CTT) (Anshel et al., 2009; Kim et al., 2012; Govindasamy et al., 2019; Uto, 2020). In the MFRM, predictions for each facet (individual, item, rater, situation, etc.) can be made independent of

other variability sources (Engelhard, 1994). For example; item parameters can be estimated independently of the severity/generosity levels of raters or other sources of variability that may affect the measurement results. In CTT, the ability levels of individuals in a test are estimated by the sum of their scores from test items. It is assumed that the difficulty levels of each item in the test (or the likelihood of participating in an item or not) are equal and / or their contribution to the total score is the same. However, if each item has a different contribution in the measured property, accepting the contribution of each item as equal in the total score causes biased results and the statistics based on this acceptance contain errors (Brinthaup & Kang, 2012). Based on the results obtained from raw scores in CTT, individuals can only be ranked according to their ability levels and these scores obtained in the ranking scale cannot be collected. However, the mathematical model on which MFRM is based overcomes this limitation and by taking the natural logarithm of the raw data (log-odds), the measurement results are converted to the interval scale (logit) level. In addition to these, compliance statistics (INFIT and OUTFIT) can be determined for each variability source with a single analysis in MFRM. In addition, parameter estimates for each facet can be interpreted together on a common ruler (logit scale) (Linacre, 1989). Relative places of facets can be examined in this ruler with a common metric. Thus, for example; By observing the distribution of items, it can be determined at what level the item was absent / missing and at what level there were many items throughout the skill level (Brinthaup & Kang, 2012). In addition, MFRM also provides descriptive information about other facets (eg raters) in the study. For example, in a measurement case involving more than one rater, one rater scoring more generous than the others; This "unexpected scoring situation" can be determined where all other raters give a high score and this rater gives a lower score (Linacre, 1989). When examined in the light of all this information, it was thought that MFRM was a suitable method for this study where 3 different raters evaluated based on 20 different criteria.

1.2. Purpose of the Study

In 2017, the “Monitoring, Research and Development Project of Measurement and Evaluation Practices” was launched by the Ministry of Education in our country. In the annual report prepared, the main objectives of this project are stated as follows;

- *Improving the measurement and evaluation capacities of the provinces,*
- *Revealing the acquisition levels of the students and teachers in a way to give feedback,*
- *Ensuring that teachers perform more qualified exams by using the Question Bank software to be created at the end of the project,*
- *Improving the capacity of conducting joint exams across the province.*

Within the scope of the project, measurement and evaluation centers have been established in 81 provinces and common exams have been started in most of these centers and these exams are still ongoing. Besides, these exams expand their content in terms of grade level and lesson each year. In this study, conducted in this regard, it was aimed to examine and assess the questions included in the province-wide “Turkish Common Exam” for sixth graders held in the first semester of 2018 by the Sakarya measurement and evaluation center. To this end, the test questions were examined by three specialists with expertise in different fields in terms of structure, content, and taxonomic values, and the obtained results were evaluated. The study is considered to be important in terms of revealing the structural features of province-wide common examinations and providing suggestions for implementation in line with the opinions of the specialists. On the other hand, it is important in terms of bringing a critical perspective to the exams and revealing the points that should be considered in the exams held in all measurement and evaluation centers throughout the country through the Sakarya Sample.

2. METHOD

This study examined the questions included in the province-wide “Turkish Common Exam” for sixth graders held in the first semester of 2018 by the Sakarya Provincial Directorate of National Education- Testing and Assessment Services Unit. The test questions were then rated by raters with expertise in different fields according to the criteria set by the researchers. Hence, the study employed the descriptive survey model (Karasar, 1998).

2.1. Study Group

The study group consists of a team of three raters (a Turkish Education Specialist, a Program Development Specialist, and a Measurement and Evaluation Specialist). Researchers came together to deal with the questions structurally. Then, they determined the structural criteria that should be included in a question in the context of curriculum development dimensions, assessment and evaluation dimensions and Turkish education program objectives in the light of the relevant literature. As a result of the decision of the researchers and the relevant literature review, 20 criteria have emerged. Each question addressed in the context of these 20 criteria was evaluated separately by the researchers. The researchers independently examined the 20 test questions included in the common exam using the 20-criteria assessment form developed by the researchers.

2.2. Data Collection Tools

In the research, firstly, literature review was conducted on the stages of test development and a list of criteria obtained from the related sources (Downing, 2006; Garden & Orpwood, 1996; Lane et al., 2015; Ozcelik, 2009; Webb, 2007) was developed. The list was examined by the specialists, who added additional criteria suitable for the purpose of the study, and a form for assessing test questions was finally created. The form was then examined by the testing and assessment specialist in terms of its scope and by a grammar specialist in terms of grammar. The form was edited and finalized according to the opinions of the specialists.

As a result of examinations and editions in terms of scope and grammar, a form consisting of 20 items was obtained. The first part of the form consists of 2 items (to find out whether the test questions are positively or negatively worded and to what step in Revised Bloom’s Taxonomy the test questions correspond), aiming to describe the descriptive features of the test question. The second part contains 18 3-point Likert type items. Rating in the second part is as follows: 1=no, 2=partially, and 3=yes. The data collection tool used in the study is included in the appendix.

2.3. Data Analysis

The data obtained from the assessment of the questions were analyzed using the Many-Facet Rasch Model (MFRM). MFRM has a conceptual framework similar to regression analysis (Eckes, 2011), and with this analysis method, groups, raters, and items are categorized in a reliable manner (Basturk, 2009). In this model, when estimating an individual’s ability or levels of items, other variables that may affect the results are taken into account; thus, more objective results are obtained (Stenner, 1990).

In the present study, three raters examined and rated the 20 multiple choice questions included in the common exam using an assessment form developed by the researchers. With MFRM analysis, the appropriateness of the questions, the consistency of the raters in rating, and the reliability of the examination criteria were tried to be determined. The study contained three facets: raters, tasks (items), and criteria. Mathematical formula for MFRM which is used for the study is:

$$\ln[P_{nijk} / P_{nijk-1}] = E_j (B_n - D_i - C_j - F_k) \quad (1)$$

In Equation (1);

- P_{nij} : probability of all items being awarded,
- E_j : a slope for the item characteristic curve associated with rater j .
- B_n : the items trait level,
- D_i : difficulty level of the item,
- C_j : raters attitude:
- F_k : difficulty of observing k 'th category (Myford & Wolfe, 2004)

MFRM analyses were performed using the FACETS computer program developed by Linacre (2007).

3. RESULT / FINDINGS

Test questions were first examined in terms of expressing the item root as positive or negative form. Table 1 presents the findings.

Table 1. *Whether the test questions are positively or negatively worded.*

Question	Positive / No	Question	Positive / No	Question	Positive / No	Question	Positive/ Negative
1	Negative	6	Positive	11	Positive	16	Positive
2	Positive	7	Positive	12	Positive	17	Positive
3	Negative	8	Positive	13	Positive	18	Positive
4	Positive	9	Positive	14	Positive	19	Positive
5	Negative	10	Negative	15	Positive	20	Positive

As can be inferred from Table 1, of the 20 questions included in the common exam, 4 are negatively while 16 are positively worded questions. Among the many suggestions given to question authors to write multiple choice questions, the most common one is to avoid negatively worded questions (Chiavaroli, 2017). In terms of frequency of citation, one review of educational textbooks noted that 31 of the 35 authors specifically advise against negatively-worded multiple choice questions (Haladyna & Downing, 1989a, 1989b) When we look at the studies in the literature, the main reason for avoiding negative questions is the lowering of the validity of the test (Haladyna & Downing, 1989b; Case & Swanson, 2002). Researchers point to an increased risk of emerging associated technical defects, such as heterogeneous options or low cognitive levels that are seen to be encouraged by negatively worded questions (Chiavaroli, 2017). For this reason, it is desirable to write multiple-choice items as positive questions. In this context, negatively worded questions were examined. It was realized that these questions could also have been written as positively worded. For example, the first question (Which of the following statements cannot be inferred from the graph?) seeks to measure students’ ability to read graphs. This question, which covers the learning objective of “interprets the information presented in graphs, tables, and charts”, could, in fact, be asked as a positively worded question. The test questions were also examined in terms of cognitive steps. For this, the raters examined the questions and decided to what step in Revised Bloom’s Taxonomy the questions correspond. Table 2 presents the findings.

Table 2. *Distribution of the Test Questions by the Revised Bloom's Taxonomy.*

	Remembering	Understanding	Applying	Analyzing	Evaluating	Creating
Items	4-8-18-19-20	1-2-5-6-7-9-11- 12-13-15-16-17	-	3-10	14	-

As can be inferred from Table 2, of the 20 questions included in the common exam, 5 correspond to “Remembering” step, 12 correspond to “Understanding” step, 2 correspond to “Analyzing” step, and 1 corresponds to “Evaluating” step. Hence, we can conclude that 85% of the questions correspond to “remembering” and “understanding” steps and that there is an insufficient number of questions for higher-order thinking skills.

The ratings of the raters based on the 18 criteria included in the second part of the assessment form were analyzed by MFRM, and the resulting variable map provided by FACETS (citation) software is given in Figure 1.

Figure 1. Variable Map.

Measr	+Rater	+Task	+Criteria	Scale
3			11 12	(3)
2				
1	3 1 2	17 3 18 15 20 19 9	14 5 16	---
0		2 10 11 4 6 16 12 8 1 14 7 13	1 15 13 2 9 8 3 18 4 7 6	2
-1		5	17	---
-2			10	(1)
Measr	+Rater	+Task	+Criteria	Scale

The logit table in Figure 1 consists of five columns. The first column (Mease) contains the logit, the unit of measurement of the Rasch model. The rater, task (item), and criteria facets are interpreted at this unit level. The second column (Rater) contains the ratings of the raters. When interpreting the rater column, the rater with the highest ratings is considered the severest rater, while the rater with the lowest ratings is considered the most lenient rater. Accordingly, the severest rater was the Turkish Education Specialist (at 1.00 logit) and the most lenient rater was the Program Development Specialist (at .70 logit). The third column lists the items according to the extent to which they meet the specified criteria. Accordingly, item 17 was the item that most met the criteria (at 1.05 logit), and item 5 was the item that least met the criteria (at -0.97 logit). The fourth column lists the criteria according to the extent to which they are met. Accordingly, the 11th criterion (The question does not contain clues for other questions) and

the 12th criterion (The answer to the question is not given in other questions) were the criteria most met by the test questions (at 2.61 logit), whereas the 10th criterion (The question is for higher-order thinking skills) is the criterion least met by the test questions (at -1.72 logit).

With the logit table, the three facets in the study are interpreted over a linear metric. In addition, detailed measurement reports were obtained for each facet by MFRM analysis. Table 3 presents the measurement results for the rater facet.

Table 3. Measurement Results for the Rater Facet.

Raters	Measure	Standard Error	Infit	Outfit
R3	1.00	0.08	1.08	1.09
R2	.94	0.08	.95	.87
R1	.70	0.08	1.00	1.04
Mean	.88	0.08	1.01	1.00
Standard Deviation	0.16	0.00	.07	.12
Reliability= .75 Separation index = 1.27 Chi-square = 8.0 $sd = 2$ $p = .02$				

According to the measurement results given in Table 3, Rater 3 was the severest rater while Rater 1 was the most lenient rater, but there is no big difference between their logit values. The fit statistics show the degree of fit between the model and the data, and a value of 1.00 is considered a perfect fit (Hetherman, 2004). According to Wright and Linacre (1994), the acceptable range of infit and outfit values is between “0.5 and 1.5”. The fact that the values are in this range indicates that the raters’ ratings fit in with the model, in other words, none of the raters disrupted the model-data fit. Since the infit values are between 1.08 and .95 and outfit values between 1.09 and .87, it can be said that the model-data fit was achieved and that no rater disrupted the fit. Also, the raters’ reliability index was calculated as .75. The reliability value refers to the difference in the raters’ ratings and, like the Cronbach alpha reliability value, it ranges between 0 and 1 and is interpreted in a similar way. In addition, the separation index, which refers to the level of difference between the raters’ ratings, is 1.27. Low separation indices indicate that the raters’ ratings are consistent and that there is no big difference between the ratings. Thus, considering these two values, we can say that the severity and leniency of the raters were similar. The chi-square (chi-square = 8.0 $p < .05$) of this difference shows that the difference between the raters is statistically significant. When the logit table in Figure 1 is examined, it is seen that the least spread is among the raters.

Table 4 presents the measurement results the task facet. As can be inferred from Table 4, item 17 was the item that most met the criteria, in other words, it was found to be the most satisfactory item by the raters, whereas item 5 was the item that least met the criteria. The infit values of the items ranged between .63 and 1.34, and the outfit values ranged between .52 and 1.34. This indicates that the infit and outfit values of all the items are in an acceptable range. The average statistical value of the infit and outfit values was found to be 1.00. Accordingly, the average of the fit statistics in the item measurement being equal to 1 indicates that the model-data fit is perfect. The reliability and separation indices of the items were found to be .85 and 2.30, respectively. These values show that the items could be adequately separated in terms of the criteria.

Table 4. Measurement Results for the Task Facet.

Items	Measure	Standard Error	Infit	Outfit
I17	1.05	.28	1.10	1.07
I3	.77	.25	.97	.83
I18	.71	.24	.85	1.30
I19	.45	.22	.79	.65
I9	.40	.22	.88	.67
I15	.35	.22	.84	1.10
I20	.30	.21	1.29	1.04
I2	.09	.20	.92	.76
I4	.01	.20	.94	.78
I10	.01	.20	1.25	1.06
I6	-.03	.20	1.34	1.22
I11	-.03	.20	.96	.82
I16	-.11	.19	.63	.52
I12	-.26	.19	1.02	.93
I8	-.29	.19	1.03	.91
I1	-.36	.19	.94	.81
I7	-.65	.19	1.11	1.14
I14	-.65	.19	.95	.80
I13	-.79	.19	1.05	.93
I5	-.97	.19	1.20	.66
Mean	0.00	.21	1.00	1.00
Stn. Dev.	.54	.02	018	.44
Reliability= .85	Separation index = 2.30	Chi-square = 115.9	<i>sd</i> = 19	<i>p</i> = .00

Table 5 presents the measurement results for the criterion facet. As can be inferred from Table 5, the 11th criterion (The question does not contain clues for other questions) and the 12th criterion (The answer to the question is not given in other questions) were the criteria most met by the test questions, whereas the 10th criterion (The question is for higher-order thinking skills) was the criterion least met by the test questions. The infit values of the criteria ranged between .57 and 1.42, and the outfit values ranged between .60 and 1.35. Considering that the optimal range for fit statistics is between 0.5 and 1.5, all criteria contributed to a perfect model-data fit. The reliability and separation indices of the criteria were found to be .93 and 3.64, respectively. Accordingly, the criteria functioned reliably to separate the items according to the extent to which they met the criteria. In addition, the significant chi-square value (chi-square = 189.9, $p < 0.05$) shows that there is a statistically significant difference between the difficulty levels of the criteria.

Table 5. Measurement Results for the Criterion Facet.

Criteria	Measure	Standard Error	Infit	Outfit
C11	2.61	.69	1.42	.87
C12	2.61	.69	1.42	.87
C14	1.02	.29	1.25	1.35
C5	.66	.25	.94	1.07
C16	.35	.21	.91	.81
C1	.14	.20	.93	1.03
C15	-.15	.18	1.25	1.19
C2	-.21	.18	1.03	1.11
C13	-.21	.18	1.18	1.07
C9	-.34	.17	.86	.84
C8	-.43	.17	1.24	1.19
C3	-.52	.17	.98	1.03
C4	-.57	.17	1.24	1.34
C7	-.57	.17	.75	.73
C18	-.63	.17	1.14	1.10
C6	-.74	.17	.96	.92
C17	-1.30	.17	.84	.88
C10	-1.72	.19	.57	.60
Mean	0.00	.25	1.09	1.00
Stn. Dev.	1.11	.16	.32	.20
Reliability= .93 Separation Index = 3.64 Chi-square = 189,9 <i>sd</i> = 17 <i>p</i> = .00				

Table 6. Measurement Results for the Rating Scale Facet.

Criterion Ratings	Frequency	%	Cumulative %	Average Measurement	Expected Measurement	Outfit
1	226	21	21	-.04	-.11	1.3
2	236	22	43	.30	.44	.5
3	618	57	100	1.44	1.41	.9

Table 6 presents the measurement results for the scale facet (1=no, 2=partially, 3=yes). As can be inferred from Table 6, of all the ratings, 21.2% are 1 (no), 22% are 2 (partially), and 57% are 3 (yes). Accordingly, as the ratings increased (from 1 to 3), their usage rates also increased. The frequency of the ratings at a value of at least 10 indicates that the ratings functioned adequately and have a balanced distribution (Engelhard, 1994). Accordingly, considering the obtained frequency, we can say that the rating data are at the desired level. The outfit values of the criterion rating range between .5 and 1.3, which indicates that the rating fits the model.

4. DISCUSSION and CONCLUSION

In this study, 20 multiple-choice questions in the 6th grade Turkish course common exam conducted throughout the province by the Directorate of National Education Directorate of Assessment and Examination Services were examined by three different field experts through a form consisting of 20 criteria.

Our findings show that of the 20 questions included in the exam, 4 are negatively while 16 are positively worded questions. Using negative questions in a test affects the test reliability;

therefore, it is necessary to avoid negative questions. Negative expressions such as “not,” “except” decrease the comprehensibility of the question, increasing the probability of the student making mistakes due to lack of attention. In addition, it takes more time for the student to answer such items (McMillan, 2013). Therefore, in common exams seeking to measure students’ language skills, to achieve accurate measurement results and to exclude other variables like “attention” from the test results, it is recommended to avoid negative questions.

It was also investigated that, the questions examined in the study correspond to which step in the "Revised Bloom Taxonomy". Of the 20 questions, 5 corresponded to “*Remembering*” step, 12 corresponded to “*Understanding*” step, 2 corresponded to “*Analyzing*” step, and 1 corresponded to “*Evaluating*” step. Accordingly, the number of questions that measure higher-order thinking skills was lower than the number of lower-level questions. Studies in the relevant literature have also reported similar findings. In the study conducted by Kavruk and Cecen (2013), the questions in the 6th, 7th, and 8th-grade tests developed by 38 Turkish teachers were examined according to Bloom’s Taxonomy. As a result of the assessment, it was observed that most of the questions were at the level of knowledge, comprehension, and application that measure lower-level skills. In a similar study conducted by Cintas Yildiz (2015), the questions in the 5th, 6th, and 7th-grade tests were analyzed according to the Revised Bloom’s Taxonomy, and most of the questions were found to be at conceptual knowledge step of the knowledge dimension and at the understanding step of the cognitive process dimension. In addition, studies conducted with taxonomies other than Bloom’s Taxonomy reported similar results. Kocaarslan and Yamac (2018) examined reading comprehension questions in tests developed by Turkish teachers according to the reading comprehension taxonomy developed by Day and Park (2005) and stated that the questions were mainly for literal comprehension. The authors also found that only a few questions triggered learners’ reorganization, prediction, and personal response skills while none aimed to assess inference and evaluation skills. Similarly, Ates (2011) found that teachers most frequently employ the strategy of asking questions and that they do not ask many questions to trigger students’ higher-order thinking processes. This shows that teachers’ questioning skills remain unchanged and continue in a traditional way, even as time progresses.

Lower-level questions that require memorization and conveying existing knowledge instead of generating new knowledge may be beneficial for the learning of disadvantaged children but does not contribute much to the development of normal and gifted children. In contrast, higher-level questions that require students to use higher-order thinking skills contribute to normal and gifted students in terms of cognitive development (Gall, 1984 as cited in Topcu, 2017). According to Akyol et al. (2013), the success (or failure) of Turkish students at questions requiring higher-order thinking skills (such as critical thinking) in international tests such as PIRLS, PISA, and TIMSS may be related to the level of questions they encounter in the teaching process and written materials. Higher-order thinking requires students to go beyond simple recall of facts and manipulate information and ideas. When teachers ask higher-level questions, they may initially see that students have difficulty in answering the questions or that they give answers consisting of only a few words. Therefore, the teacher should model for his/her students how to give a higher-level answer. Though it may take some time to train students to give higher-level answers, it will definitely produce positive outcomes (Peterson & Taylor, 2012). In fact, the Turkish Course Curriculum (Republic of Turkey Ministry of National Education, 2019) underlines the importance of developing tests and exams that contain various types of items that trigger students’ higher-order thinking processes such as making inferences, critical thinking, analysis, visual reading, reasoning, and spatial skills.

The 18 criteria in the assessment form were analyzed by MFRM. The study contained three facets: raters, items, and criteria. There were no differences among the raters (a Turkish Education Specialist, a Program Development Specialist, and a Testing and Assessment

Specialist) in terms of severity and leniency: all the raters were in agreement. Commissions to develop common exams to be conducted through the central examination system should include specialists in different fields: a specialist in the lesson content, a program development specialist, and a testing and assessment specialist. The common exam assessed in this study was developed by a commission of Turkish teachers working in the Sakarya Provincial Directorate of National Education- Testing and Assessment Services Unit. The commission does not include a program development specialist or a testing and assessment specialist. However, the commissions to develop such province-wide common exams that will affect many students should include specialists with expertise in different fields. In addition, in-service training on testing and assessment approaches and tools and developing new types of questions should be given to the teachers in such commissions. As a matter of fact, Maden (2011) stated that Turkish teachers found complex the testing and assessment tools and methods in the 2006 Turkish Course Curriculum, and Erdogan (2017) stated that teachers do not make enough effort to improve their questioning skills. As a result, rather than creating their own questions or tasks to use in tests, teachers use readily available questions included in printed or online resources. In fact, we realized that some of the test questions included in the common exam were taken from other resources.

Considering all the criteria in the assessment form used in the study, of the 20 items, only 8 are in the range of 0 and 1 logit, 2 are at 0 logit, and 10 at a negative logit. This indicates that the questions failed to meet the criteria sufficiently. Also, it was observed that the criteria measuring the structural features of the questions were met while the criteria measuring the quality and comprehensibility of the questions were not met. This shows that though the exam development commission paid attention to the structural features of the test questions, they failed to attach sufficient importance to the quality of test questions. In other words, they took care to include multiple-choice items that seek to measure the learning objectives specified in the curriculum but failed to meet the criteria set for the quality of test questions.

Furthermore, the exam failed to measure the four basic language skills in the mother tongue: though the exam contained questions measuring students' grammatical knowledge and reading comprehension skills, there were no questions to assess students' listening, speaking, or writing skills. Turkish classes are aimed at helping students develop all four basic language skills. For this reason, and in order to develop tests measuring students' four basic language skills, the "Turkish Language Test for Four Skills" developed by Republic of Turkey Ministry of National Education (2020) should be examined thoroughly by teachers.

Overall, the study concludes that the questions included in the common exam was appropriate for the learning objectives specified in the Turkish Course Curriculum (Republic of Turkey Ministry of National Education, 2019) but failed to address higher-order thinking skills. Therefore, we recommend that exam development commissions to develop province-wide common exams that will affect many students should include specialists with expertise in different fields.

In this study, an exam for the Turkish course prepared and administered by the Sakarya Provincial Directorate of National Education- Measurement and Evaluation Center- is examined. In the future researches, it is recommended to examine the exams held in different provinces for both Turkish and different courses, compare the results obtained and thus determine the situation across the country.

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

Gulden Kaya Uyanik: Investigation, Resources, Supervision, Methodology, Development of Data Collection Tool, Analysis, Writing the original draft. **Tugba Demirtas Tolaman:** Investigation, Resources, Development of Data Collection Tool, Writing the original draft. **Duygu Gur Erdogan:** Investigation, Resources, Development of Data Collection Tool, Writing the original draft.

ORCID

Gulden Kaya Uyanik  <https://orcid.org/0000-0002-8100-6994>

Tugba Demirtas Tolaman  <https://orcid.org/0002-6632-9752>

Duygu Gur Erdogan  <https://orcid.org/0000-0002-2802-0201>

5. REFERENCES

- Aktaş, E. (2017). Öğretmen adaylarının farklı metin türlerine yönelik soru sorma becerilerinin Yenilenmiş Bloom Taksonomisine göre değerlendirilmesi [Evaluation of the questioning skills of teachers candidates towards the different text types according to The Renewed Bloom Taxonomy]. *Electronic Turkish Studies*, 12(25), 99-118. <http://dx.doi.org/10.7827/TurkishStudies.12274>
- Akyol, H., Yıldırım, K., Seyit, A., & Çetinkaya, Ç. (2013). Anlamaya yönelik nasıl sorular soruyoruz? [What kinds of questions do we ask for making meaning?]. *Mersin University Journal of Education*, 9(1), 41-56.
- Anderson, L. W. (2005). Objectives, evaluation, and the improvement of education, *Studies in Education Evaluation*, 31, 102-113. <https://doi.org/10.1016/j.stueduc.2005.05.004>
- Anderson, L.W., Krathwohl, D.R., Airaisan, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). A taxonomy for learning, teaching and assessing: a revision of Bloom's Taxonomy of educational objectives. Addison Wesley Longman, Inc.
- Andre, T. (1979). Does Answering higher-level questions while reading facilitate productive learning? *Review of Educational Research*, 49(2), 280-318.
- Anıl, D. (2009). Uluslararası öğrenci başarılarını değerlendirme programı (PISA)'nda Türkiye'deki öğrencilerin Fen Bilimleri başarılarını etkileyen faktörler [Factors effecting Science achievement of science students in Programme for International Students' Achievement (PISA) in Turkey]. *Education and Science*, 34(152), 87-100.
- Anshel, M.H., Weatherby, N.L., Kang M. & Watson, T. (2009). Rasch calibration of a unidimensional perfectionism inventory for sport. *Psychology of Sport and Exercise*, 10(2009), 210-216. <https://doi.org/10.1016/j.psychsport.2008.07.006>
- Ateş, S. (2011). *Evaluation of Fifth-Grade Turkish Course Learning and Teaching Process in Terms of Comprehension Instruction* [Unpublished doctoral dissertation, Gazi University]. Gazi University Libraries.
- Ateş, S., Güray, E., Döğmeci, Y., & Gürsoy, F. F. (2016). Öğretmen ve öğrenci sorularının gerektirdikleri zihinsel süreçler açısından karşılaştırılması [Comparison of questions of teachers and students in terms of level]. *Research in Reading and Writing Instruction*, 4(1), 1-13.
- Baştürk, R. (2009). Applying The Many – Facet Rasch Model to evaluate powerpoint presentation performance in higher education. *Assesment And Evaluation in Higher Education*, 33(4), 431-444. <https://doi.org/10.1080/02602930701562775>

- Bircan, E. (2012). Türkçe öğretmeni adaylarının hazırladığı soruların yeniden yapılandırılan Bloom Taksonomisine göre değerlendirilmesi [Evaluation of the questions prepared by Turkish language teacher candidates according to The Revised Bloom's Taxonomy]. *Kastamonu University Journal of Education*, 20(3), 965-982.
- Bloom, B. (1956). Taxonomy of educational objectives. David Mckay. <https://www.uky.edu/~rsand1/china2018/texts/Bloom%20et%20al%20-Taxonomy%20of%20Educational%20Objectives.pdf>
- Bozkurt, B.Ü., Uzun, G.L., & Lee, Y. (2015). Korece ve Türkçe ders kitaplarındaki metin sonu sorularının karşılaştırılması: PISA 2009 sonuçlarına dönük bir tartışma [A comparison of reading comprehension questions in Korean and Turkish textbooks: A discussion on PISA 2009 results]. *International Journal of Language Academy*, 3(4), 295-313. <http://dx.doi.org/10.18033/ijla.327>
- Brinthaupt, T.M., & Kang, M. (2012). Many-faceted rasch calibraton: an example using the self-talk scale. *Assessment*, 21(2), 241-249. <https://doi.org/10.1177/1073191112446653>
- Büyüköztürk, Ş. (2016). Sınavlar üzerine düşünceler [Thoughts on exams]. *Kalem International Journal of Education and Human Sciences*, 6(2), 345-356.
- Case, S.M. & Swanson, D.B. (2002). *Constructing written test questions for the basic and clinical sciences*. 3rd Ed (rev.) National Board of Medical Examiners.
- Cayhan, C., & Akın, E.(2015). TEOG sınavı Türkçe dersi sorularının Türkçe Dersi Öğretim Programındaki kazanımlar açısından değerlendirilmesi [The evaluation of Turkish lesson questions TEOG examination in terms of Turkish lesson education program objectives]. *Siirt University Journal of Social Sciences Institute*, 5, 106-114.
- Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research, and Evaluation*, 22(1), 3. <https://doi.org/10.7275/ca7y-mm27>
- Çeçen, M. A., & Kurnaz, H. (2015). Ortaokul Türkçe dersi öğrenci çalışma kitaplarındaki tema değerlendirme soruları üzerine bir araştırma [Student workbook of secondary school Turkish course: A research on theme evaluation questions]. *Journal of Karadeniz Social Sciences*, 7(2).
- Çeliktürk Sezgin, Z., & Gedikoğlu Özilhan, Y. G. (2019). 1.-8. sınıf Türkçe ders kitaplarındaki metne dayalı anlama sorularının incelenmesi [Examining text-based comprehension questions in Turkish textbooks of the 1st- the 8st graders]. *Journal of Mother Tongue Education*, 7(2), 353-367. <https://doi.org/10.16916/aded.530191>
- Çepni, S., Özsevgenç, T. & Gökdere, M. (2003). Bilişsel gelişim ve formal operasyon dönem özelliklerine göre ÖSS fizik ve lise fizik sorularının incelenmesi [Examination of SSE physics and high school physics questions according to cognitive development and formal operation period features]. *Journal of National Education*, 157, 30-39.
- Çer, E. (2018). A comparison of mother-tongue curricula of successful countries in PISA and Turkey by higher-order thinking processes. *Eurasian Journal of Educational Research*, 73, 95-112.
- Day, R. R., & Park, J. (2005). Developing reading comprehension questions. *Reading in a Foreign Language*, 17(1), 60-73.
- Demiral, H., & Menşan, N. (2017). Sekizinci sınıf Türkçe dersinin PISA okuma becerilerine göre değerlendirilmesi [Evaluation of the eighth grade Turkish lesson according to PISA reading skills]. *Küreselleşen dünyada eğitim* (Edt: Özcan Demirel, Serkan Dinçer). Pegem Yayıncılık.
- Dillon, J.T. (2006). Effect of questions in education and other enterprises. In *Westbury, I.& Milburn, G. (Eds.), Rethinking Schooling* (pp.145-174). Routledge. <https://doi.org/10.4324/9780203963180>

- Downing, S. M. (2006). Twelve steps for effective test development. In Downing, S.M,& Haladyna, T. M. (Eds.), *Handbook of test development*, (pp.3-25). Routledge.
- Durukan, E. (2009). 7. sınıf Türkçe ders kitaplarındaki metinleri anlamaya yönelik sorular üzerine taksonomik bir inceleme [A taxonomic study on questions to understand texts in 7th grade Turkish Textbooks]. *Journal of National Education*, 181, 84-93.
- Eckes, T. (2011). *Introduction to many-facet rasch measurement*. Frankfurt Am.
- Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of Educational Measurement*. 31(2), 93-112. <https://doi.org/10.1111/j.1745-3984.1994.tb00436.x>
- Erdogan, T. (2017). İlkokul dördüncü sınıf öğrencilerinin ve öğretmenlerinin Türkçe dersine ilişkin sordukları soruların Yenilenmiş Bloom Taksonomisi açısından görünümü [The view of primary school fourth grade students and teachers' questions about Turkish language lessons in the terms of The Revised Bloom Taxonomy]. *Education and Science*, 42(192). <http://dx.doi.org/10.15390/EB.2017.7407>
- Eroğlu, D., & Kuzu, T.S. (2014). Türkçe ders kitaplarındaki dilbilgisi kazanımlarının ve sorularının Yenilenmiş Bloom Taksonomisine göre değerlendirilmesi [The evaluation of the grammar acquisitions and questions in Turkish course books with respect to New Bloom Taxonomy]. *Başkent University Journal of Education*, 1(1), 72-80.
- Garden, R. A., & Orpwood, G. (1996). Development of The TIMSS Achievement Tests. *Third International Mathematics and Science Study. Technical Report, 1*.
- Govindasamy, P., del Carmen Salazar, M., Lerner, J., & Green, K. E. (2019). Assessing the reliability of the framework for equitable and effective teaching with the many-facet rasch model. *Frontiers in Psychology*, 10, 1363. <https://doi.org/10.3389/fpsyg.2019.01363>
- Göçer, A. (2008). İlköğretim Türkçe ders kitaplarının ölçme ve değerlendirme açısından incelenmesi [Analysis of Turkish course books for measurement and evaluation]. *Atatürk University Journal of Social Sciences Institute*, 11(1), 197-210.
- Göçer, A. (2016). Lisansüstü eğitim gören Türkçe öğretmenlerinin yazılı sınav sorularının incelenmesi [Investigation of written exam questions of Turkish teachers who upper graduate education]. *Uşak University Journal of Social Sciences Institute*, 9(27/3), 22-37.
- Güfta, H., & Zorbaz, K. Z. (2008). İlköğretim ikinci kademe türkçe dersi yazılı sınav sorularının düzeyleri üzerine bir değerlendirme [A review regarding levels of written examination questions for Turkish courses of the secondary school]. *Çukurova University Journal of Social Sciences Institute*, 17(2), 205-218.
- Güneş, F. (2007). *Türkçe öğretimi ve zihinsel yapılandırma* [Turkish teaching and mental structuring]. Nobel Yayın Dağıtım.
- Güneş, F. (2011). Dil Öğretim yaklaşımları ve Türkçe öğretimindeki uygulamalar [Language teaching approaches and their applications in teaching Turkish]. *Mustafa Kemal University Journal of Social Sciences Institute*, 8(15), 123-148.
- Güneş, F. (2012). Testlerden etkinliklere türkçe öğretimi [Teaching Turkish from tests to activities]. *Journal of Language and Literature*, 1(1), 31-42.
- Haladyna, T.M., & Downing, S.M. (1989a). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 2(1), 37-50. https://doi.org/10.1207/s15324818ame0201_3
- Haladyna, T.M., & Downing, S.M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 2(1), 51-78. https://doi.org/10.1207/s15324818ame0201_4
- Haladyna, T.M. (1997). *Writing test items to evaluate higher order thinking*. Viacom Company.

- Hetherman, S.C. (2004). *An Application of Multi Faceted Rasch Measurement to Monitor Effectiveness of the Written Composition in English in The New York City Department of Education* [Doctoral dissertation, Columbia University]. Columbia University Libraries.
- Hill, J.B. (2016). Questioning techniques: A study of instructional practice. *Peabody Journal of Education*, 91(5), 660-671. <https://doi.org/10.1080/0161956X.2016.1227190>
- Karadüz, A. (2010). Dil becerileri ve eleştirel düşünme [Language skills and the critical thinking]. *Turkish Studies*, 5(3), 1566-1593. <http://dx.doi.org/10.7827/TurkishStudies.1572>
- Karasar, N. (1998). *Araştırmalarda rapor hazırlama yöntemi* [Research Report Preparation Method]. Ankara: Pars Matbaacılık Sanayi.
- Kardeş-Birinci, D. (2014). Merkezi sistem ortak sınavlarında ilk deneyim: Matematik dersi [The first experience in central system common exams: mathematics]. *Journal of Research in Education and Teaching*, 3(2), 8-16.
- Kavruk, H., & Çeçen, M.A. (2013). Türkçe dersi yazılı sınav sorularının bilişsel alan basamakları açısından değerlendirilmesi [Evaluation of Turkish language class exam questions in point of cognitive field levels]. *Journal of Mother Tongue Education*, 1(4), 1-9. <https://doi.org/10.16916/aded.15990>
- Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly*, 29, 346-365.
- Kocaarslan, M., & Yamaç, A. (2018). Sınıf öğretmenlerinin Türkçe dersi sınavlarında sordukları metne dayalı anlama sorularının incelenmesi [Investigating text-based comprehension questions primary school teachers ask in exams of Turkish course]. *Trakya Journal of Education*, 8(2), 431-448. <https://doi.org/10.24315/trkefd.356769>
- Kurudayıoğlu, M., & Çetin, Ö. (2015). Temel beceriler ve Türkçe öğretimi [Basic skills and Turkish education]. *Journal of Mother Tongue Education*, 3(3), 1-19. <https://doi.org/10.16916/aded.65619>
- Lane, S., Raymond, M.R., & Haladyna, T.M. (2015). *Handbook of test development*. Routledge.
- Linacre, J.M. (1989). *Many-facet Rasch Measurement* [Doctoral dissertation, University of Chicago]. Chicago University Libraries.
- Linacre, J. M. (2007). *A User's Guide to FACETS: Rasch Model Computer Programs*. Chicago, IL.
- Maden, S. (2011). Türkçe dersi öğretmenlerinin ölçme değerlendirmeye ilişkin algıları [Turkish course teachers' perceptions on measurement and evaluation]. *Journal of National Education*, 41(190), 212-233.
- Marzano, R.J., & Kendall, J.S. (2007). *The new taxonomy of educational objectives*. Sage.
- McMillan, J. H. (2013). *Classroom assessment: principles and practice for effective standards-based instruction* (6th Edition). Pearson.
- Mislevy, R.J., & Riconscente, M.M. (2006). Handbook of test development. In Downing, S.M., & Haladyna, T.M. (Eds.), *Evidence-centered assessment design* (pp.61-90). Routledge.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using Many-Facet Rasch measurement: part II. *Journal of Applied Measurement*, 5(2), 189-227.
- Noddings, N. (2018). *Philosophy of education*. Routledge.
- Önalın, K., & Nesrin, Z. (2015). Türkçe ders kitaplarındaki metin altı soruların aşamalı sınıflandırmaya göre incelenmesi [Examining text-based comprehension questions in Turkish textbooks of the 1 st - the 8 st Graders]. *International Journal of Languages' Education and Teaching*, 1527-1533.
- Özçelik, D.A. (2009). *Test hazırlama klavuzu* [Test Guide]. (4. Baskı). Pegem Akademi.

- Özdemir, M., Özdemir, O., & Çetinkaya, Ç. (2007, 15-17 November). *Analysis of the questions in the primary Turkish course workbooks* [Conference presentation]. 1. Ulusal İlköğretim Kongresi, Ankara.
- Peterson, D.S., & Taylor, B.M. (2012). Using higher order questioning to accelerate students' growth in reading. *The Reading Teacher*, 65(5), 295-304. <https://doi.org/10.1002/TRTR.01045>
- Republic of Turkey Ministry of National Education (2006). Turkish Course Curriculum.
- Republic of Turkey Ministry of National Education (2019). Turkish Course Curriculum. <http://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=663>
- Republic of Turkey Ministry of National Education (2020). Turkish Language Exam in Four Skills. https://www.meb.gov.tr/meb_iys_dosyalar/2020_01/20094146_Dort_Beceride_Turkce_Dil_Sinavi_Ocak_2020.pdf
- Sarar Kuzu, T. (2013). Türkçe ders kitaplarındaki metin altı sorularının Yenilenmiş Bloom Taksonomisindeki hatırlama ve anlama bilişsel düzeyleri açısından incelenmesi. [Investigation of the text following questions in Turkish course books with respect to their remembering and understanding cognition levels of The Revised Bloom Taxonomy]. *Sivas Cumhuriyet University Faculty of Letters Journal of Social Sciences*, 37(1), 58-76.
- Shaunessy, E. (2000). Questioning techniques in the gifted classroom. *Gifted Child Today*, 23(5), 14-21. <https://doi.org/10.4219/gct-2000-752>
- Stenner, A. J. (1990). Objectivity: Specific and General. *Rasch Measurement Transactions*, 3(4), 111.
- Topçu, E. (2017). TEOG Tarih sorularının Yenilenmiş Bloom Taksonomisine göre analizi [Analysis of History questions asked in the transition from primary to secondary education according to The Renewed Bloom Taxonomy]. *International Journal of Turkish Education Sciences*, 9, 321-335.
- Turkish Course Common Exam. (2018). The Measurement and Evaluation Center. <http://sakarya.odm.meb.gov.tr>
- Uto, M. (2020). Accuracy of performance-test linking based on a many-facet Rasch model. *Behavior Research Methods*, 1-15. <https://doi.org/10.3758/s13428-020-01498-x>
- Webb, N. L. (2007). Issue related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1), 7-25.
- Wilens, W.W. (1991). *Questioning skills for teachers*. National Education Association.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable Mean-Square Fit Values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 8(3), 370.
- Yeşilyurt, E. (2012). Öğretmen adaylarının bilişsel alanla ilgili sınav durumu soruları yazma yeterliklerinin değerlendirilmesi [Evaluating teacher candidates' competencies on writing testing situation questions related to cognitive domain]. *Kastamonu University Journal of Education*, 20(2), 519-530.
- Yıldırım, K. (2012). Öğretmenlerin öğrencilerin okuduğunu anlama becerilerini değerlendirmede kullanabilecekleri bir sistem: Barrett Taksonomisi [A system to be used by teachers to evaluate students' reading comprehension skills: Barrett Taxonomy]. *Mustafa Kemal University Journal of Social Sciences Institute*, 9(18), 45-58.
- Yıldız, D. Ç. (2015). Türkçe dersi sınav sorularının yeniden yapılandırılan Bloom Taksonomisine göre analizi [The analysis of Turkish course exam questions according to re-constructed Bloom's Taxonomy]. *Gaziantep University Journal of Social Sciences*, 14(2), 479-497.

Adaptation of Statistics Anxiety Scale to Turkish: Validity and Reliability Study

Ismail Durak ^{1,*}, Yalcin Karagoz ²

¹Duzce University, Faculty of Business Administration, Department of Business Administration, Düzce

²Duzce University, Faculty of Business Administration, Department of Health Management, Düzce

ARTICLE HISTORY

Received: Jan. 17, 2021

Revised: May 22, 2021

Accepted: July 06, 2021

Keywords:

Statistics anxiety,
Measuring statistics
anxiety,
Scale adaptation,
Validity,
Reliability.

Abstract: The aim of this study is to adapt the Statistics Anxiety Scale (SAS) developed by Vigil-Colet et al. (2008) to Turkish. This study is expected to fill an important gap in the literature since no valid and reliable specific statistics anxiety scale developed or adapted in Turkish for undergraduate students in the literature is available. The sample consists of a total of 439 university students, 258 women and 181 men, studying at Düzce University. The construct validity of the Turkish form of SAS was examined by EFA and DFA. Also, for the criterion validity, a different statistics anxiety scale whose validity and reliability tested was used. As a result of EFA, a three-dimensional structure was obtained as in the structure of the original scale. According to the CFA results, which is the second analysis for construct validity, all fit index results of the model were at an acceptable level. Thus, the CFA results supported the three-factor structure obtained from EFA findings. As a result of the reliability analysis, the Cronbach's Alpha internal coefficients of the SAS and its subscales and, the Guttman and Spearman-Brown internal consistency coefficients of Split-Half Reliability methods were quite high and above the limit of 0.70. For item discrimination, items have good discrimination by obtaining all values above 0.30 lower limit in the results. When the results of the study are evaluated as a whole, the SAS form adapted to Turkish can be used as a guiding scale to measure the statistics anxiety of undergraduate students.

1. INTRODUCTION

Statistics is a discipline that is very important today as it was in the past. Although there are many reasons for this, some of them are its being a necessary tool for scientific research (Neşe et al., 2019), used for logical inference, critical thinking and right decision (Joe, 2005; Chew, 2016), and to promote in technology and knowledge. Along with these, Basic statistics information is used for many situations in daily life including forecasting weather and growth, crime and unemployment rates, the spread of a disease, keeping various reports as political elections, etc. (Chew, 2016; Paul et al., 2018). The use of statistics techniques especially in big data, data science and various artificial intelligence techniques, which are important components of artificial intelligence, increases the importance of statistics. However, many people are not statistically literate and lack the ability to perform statistics evaluation (Utts,

*CONTACT: İsmail Durak ✉ ismaildurak@duzce.edu.tr 📍 Düzce University, Faculty of Business Administration, Department of Business Administration, Düzce, Turkey

2003; Earp, 2007). In addition, it is stated in various studies that Statistics plays an important role in students' academic careers (Young & Nelson (1994); Parker et al., 1999; Collins & Onwuegbuzie, 2007; Neşe et al., 2019). For these reasons, statistics course is both very important and it is offered as a compulsory course in many programs of universities in both social sciences and natural sciences. For example, all the business program, one of the social sciences fields, at the graduate level is required to take at least one statistics course in Turkey. Similarly, research methods course includes a substantial proportion of statistics is offered among compulsory courses in almost all programs at the undergraduate and graduate education in Turkey. In addition, measurement and evaluation in education courses, which include a certain level of statistics and are included in all programs in education faculties, are among the compulsory courses. The important role of statistics and statistics related subjects in the curriculum in Turkey as it is visible in the curricula of other countries. For example, Stoloff et al. (2009) states that taking at least one statistics course is mandatory in 98% of psychology programs, which is one of the social sciences programs, in the USA Chew (2016), while this rate is 100% in Singapore and Australia.

Although statistics is a compulsory course in many undergraduate programs, due to its quantitative and mathematical based structure, it can create a risky perception and cause anxiety in the minds of students who do not have a numerical background (e.g. social sciences). Even, statistics and statistics-based courses are positioned as a negative situation by graduate students (Collins & Onwuegbuzie, 2007). For this reason, students may exhibit academic procrastination behavior, which can be defined as delaying preparation to an exam, doing homework and other academic obligations for various reasons (Roberts & Bilderbeck, 1980). Moreover, being a known fact that personal characteristics play a role in the academic performance of students, this situation is supported in many studies (Furnham & Chamorro-Premuzic, 2004; Vigil-Colet, 2008; Zhou, 2015; Hazrati-Viari et al.; Sari et al., 2017). Anxiety, one of the personal characteristics, is a forward-looking mood and is defined as an emotional state associated with preparedness for possible upcoming negative events (Spielberger, 1983; Barlow, 2002). Although it is stated that moderate anxiety has a positive effect on the individual (Donnelly, 2009), high levels of anxiety can negatively affect the social, work, psychological, family and educational life of the individual (Zahrakar, 2008). High level of anxiety in education life may occur more especially in some lessons. Since statistics, which is one of these courses, creates anxiety for many students, it is specifically addressed in the literature and is called "statistics anxiety".

Statistics anxiety is a type of situational anxiety and is defined as an emotion that occurs when faced with statistics in any form or time (when a statistics lesson is taken, a statistics analysis or interpretation is required, etc.) (Zeidner, 1991; Onwuegbuzie et al., 1997). This situation prevents learning and negatively affects academic performance as well as creates various psychological problems (Onwuegbuzie & Daley, 1999). Such problems caused by statistics anxiety attracted the attention of researchers and led to the development of various scales to measure statistics anxiety. However, it is stated that some scales developed for statistics anxiety contain items related to attitude towards statistics lesson (For example, STARS) and some of them include items related to mathematics anxiety (For example, Zanakis & Valenzi, 1997), and it is emphasized that the distinction between them should be made (Nasser, 2004; Grajzel, 2019). Unlike statistics anxiety, attitude towards statistics is a multidimensional phenomenon that shows students' learned tendencies to respond positively or negatively to statistics (Emmioğlu & Çapa -Aydın, 2012), while mathematics anxiety is generally a negative emotional reaction to mathematics and defined as a state of tension and discomfort caused by problems (Hembree, 1990). Although it was mostly associated with mathematics anxiety in the past (For example, Zanakis & Valenzi, 1997) and the statistics anxiety scale was first developed based on mathematics anxiety (Pretorius & Norman, 1992), these three concepts, which are clearly

seen to have different meanings by definition, should be distinguished from each other to measure them correctly. Apart from that, Vigil-Colet et al. (2008) stated that to reveal the relationship between academic performance and anxiety more clearly, a specific anxiety scale should be used for the variable aimed to be measured. Supporting this, the work of Rindermann and Neubauer (2001) and Ferrando et al. (1999) found that test anxiety scale is more related to academic performance than a general anxiety scale.

When the literature on statistics anxiety is examined, it is seen that the literature focuses on the relationship between statistics anxiety and academic performance. Accordingly, statistics anxiety can damage students' thinking abilities, thus causing a decrease in learning and academic performance. To put it more clearly, it has been revealed by various researchers that statistics anxiety causes psychological problems such as depression and panic as well as various physical problems such as muscle pain and headache (Onwuegbuzie et al., 1997), decreases focus (Chiesi, Prime, & Marquez, 2011), and causes distraction. (Fitzgerald, Jurs, & Hudson, 1996). It has been found as a result of various studies that such problems caused by statistics anxiety specifically affect statistics success negatively (Hanna & Dempster, 2009) and cause low academic success (Baloglu, 2001; Gal & Ginsburg, 1994; Fitzgerald, Jurs, & Hudson, 1996). In some other studies, statistics anxiety was found to have an indirect negative effect on performance (Chiesi, Prime, & Marquez, 2011). In addition, another negative situation created by statistics anxiety is that it causes academic procrastination. As a matter of fact, a student who took the statistics course for the first time and failed or passed the statistics course with difficulty may prefer academic procrastination in order not to face statistics again. Supporting this, Alexander and Onwuegbuzie (2007) and Vahedi et al. (2012) found that students displayed the behavior of delaying writing the statistics term report, studying for the exam and performing weekly homework.

The second focus of the literature is on the relationship between statistics anxiety and attitude towards statistics, and the relationship between statistics anxiety and mathematics anxiety. Accordingly, many researchers argue that there is a negative relationship between Statistics Anxiety and attitude towards statistics lesson (Chiesi et al., 2011; Mji & Onwuegbuzie, 2004; Watson et al., 2003; Zanakis & Valenzi, 1997). Accordingly, Chiese and Primi (2010) concluded and pointed out a two-way relationship that high statistics anxiety reduced attitude towards statistics, and low attitude towards statistics resulted in high statistics anxiety. When examining the studies on the relationship between statistics anxiety and mathematics anxiety, it was found that students who mostly have a poor mathematics background or take a limited number of mathematics lessons have higher statistics anxiety compared to other students (Baloglu & Zelhart, 2004; Zeidner, 1991; Primi & Chiesi, 2018). On the other hand, there are some studies that have been found to have high statistics anxiety, although mathematics anxiety is low (For example, Onwuegbuzie et al., 1997). However, a few such results in the literature do not cast doubt on the positive correlation between mathematics anxiety and statistics anxiety, which is generally accepted and proved by many studies.

Another focus of the literature is the relationship between statistics anxiety and socio-demographic characteristics. In this direction, the relationship of statistics anxiety with the following demographic characteristics was particularly emphasized; gender, age, mathematics background, social class, ethnicity, personality type, reading ability (Onwuegbuzie & Wilson, 2003; Collins & Onwuegbuzie 2007). One of the most studied socio-demographic features in relation to statistics anxiety is gender. Benson and Bandalos (1989) found that girls have a higher level of statistics anxiety than boys. However, in a similar study, although statistics anxiety of girls was higher than boys, no difference was found between these two groups in terms of statistics success (Bradley & Wygant, 1998). There are also a bunch studies examining the relationship between statistics anxiety and mathematics background, one of the socio-

demographic characteristics. Accordingly, Malik (2015), Becker and Bzhetai (2018) and Grajzel (2019) found that a strong mathematics background has a decreasing effect on statistics anxiety, and in the opposite case, it has an increasing effect. However, there are a few studies that conclude that there is no relationship between mathematics background and statistics anxiety (For example, Sutarso, 1992).

In the past, statistics anxiety, which was mostly associated with mathematics anxiety and measured within that framework, resulted in the development of scales named directly with the statistics anxiety scale over time. For this purpose, the first scale included in the literature under the name of statistics anxiety is Statistics Anxiety Rating Scale (STARS), developed by Cruise et al. (1985). Afterwards, it can be listed as Statistics Anxiety Inventory developed by Zeidner (1991), Statistics Anxiety Scale developed by Köklü (1996), Statistics Anxiety Measure developed by (Earp, 2007), Statistics Anxiety Scale (SAS) developed by Vigil-Colet et al. (2008), and Statistics Anxiety Scale developed by Faber et al. (2018). Although STARS is the most widely used Statistics Anxiety Scale whose psychometric properties have been studied various studies as Baloğlu (2002), Hanna et al. (2008), Chew et al. (2018), this scale has been criticized for having different structures other than anxiety (attitude towards statistics) and being quite long (Vigil-Colet et al., 2008; Chew & Dillon, 2014; Grajzel, 2019). Therefore, Vigil-Colet et al. (2008) developed a shorter scale and specifically measuring statistics anxiety whose sub-dimensions taken from the STARS.

In addition, until recently, there were no other scales other than the scale developed by Köklü (1996), which was developed in Turkish directly to measure statistical anxiety. However, it has been determined that the scale developed by Köklü (1996) is neither in the archive of the journal in which the study was published nor in Google Scholar etc. databases. On the other hand, as a result of the literature review conducted within the scope of this study and as stated by Güler et al. (2019), although there are Turkish studies (for example Baloğlu & Zelhart, 2004) using STARS, no studies have been found that adapt this scale to Turkish. Apart from that, in the Turkish literature, there is an adapted scale (adapted by Güler et al., 2019) to measure graduate students' statistics anxiety. However, although this scale measures the statistics anxiety of undergraduate students to a certain extent, it cannot be able to measure comprehensively since it was developed for graduate students.

Finally, it has just been noticed due recently published that Bektaş et al. (2021) adapted the SAS of Vigil-Colet et al. (2008) to Turkish. Our study includes some advantages over the study of Bektaş et al. (2021). Firstly, we also apply confirmatory factor analysis, which is a crucial analysis for testing the construct validity of the scale, as it was performed on many developed scales. Furthermore, there was no item lost in the scale we adapted but in theirs were. Therefore, it is aimed to fill an important gap by introducing a different reliable and validated scale for measuring Statistics Anxiety to the Turkish literature with this adaptation study. In addition, with this adaptation study, a scale that could both directly measure statistics anxiety and be considered relatively short to the STARS etc. scales would be introduced to the Turkish literature.

2. METHOD

This study, which aims to adapt Statistics Anxiety Scale developed by Vigil-Colet et.al (2008) into Turkish, is a quantitative descriptive research. Detailed information on participants, data collection tools, adaptation process of the scale to Turkish, data collection Process and analysis are presented below.

2.1. Participants

The sample of the study consists of 439 undergraduate students who are studying at the Faculty of Business Administration of Düzce University and enrolled in the Statistics course in Fall 2019. In this context, data were collected from the students of Business Administration, International Trade and Health Management departments of the Faculty of Business Administration using the random sampling method. In this regard, a questionnaire was shared in social media course groups in which all students participated, within the framework of the ethics committee's permission. The average age of the participants was 21.18 and the standard deviation was 1.13. The characteristics of the participants' department and gender variables are given in Table 1.

Table 1. *Frequencies of the Participants by Department and Gender.*

Department	Frequency	Percentage (%)	Gender	Frequency	Percentage (%)
Business Administration	188	42.6	Female	260	59.0
International Trade	137	31.1	Male	181	41.0
Health Management	116	26.3	Female	260	59.0

2.2. Data Collection Tools

Statistics Anxiety Scale-SAS (Vigil-Colet et al., 2008): In order to purify statistics anxiety from the different structures as included in SARS scales and thus measure it more specifically and accurately, Vigil-Colet et al. (2008) developed a Statistics Anxiety Scale (SAS). This scale includes a total of 24 items and has a three-dimensional structure. In this scale, the participants were asked to state their opinions on a 5-point Likert-type scale ranging from 1 = No Anxiety (*Absolutely Disagree*) to 5 = Considerable Anxiety (*Strongly Agree*). The dimensions of this scale are Exam Anxiety-EA, Interpretation Anxiety-IA and Anxiety to Ask for Help-AAH. One of the sample items in the Exam Anxiety dimension is " Studying for an examination in a statistics course ". One of the items in the Interpretation Anxiety dimension is " Interpreting the meaning of a table in a journal article". One of the items in the Asking Anxiety dimension is " Going to ask my statistics teacher for individual help with material I am having difficulty understanding". Vigil-Colet et al. (2008) used three dimensions and 11 questions from the STARS scale while developing SAS. The remaining 13 questions were originally obtained from faculty members who teach statistics. In Table 2, questions in each dimension of original scale are given.

Table 2. *Subscales of SAS and Corresponding Items.*

Subscales	Items
Examination anxiety	1*, 4, 9*, 11*, 13, 14*, 15, 20
Interpretation anxiety	2*, 6*, 8, 10*, 16, 18*, 19, 22*
Asking for help anxiety	3*, 5, 7, 12, 17*, 21, 23, 24

Note: *=Items obtained from the STAR

The validity of SAS has been proved in Spain, Italy, Australia, Singapore, Bangladesh, and the USA and adapted into these languages. Accordingly, the internal consistency coefficient results obtained in these adapted studies are given in Table 3.

Table 3. Internal Consistency Coefficients Obtained from SAS in Various Studies.

Studies	Examination anxiety	Interpretation anxiety	Asking for help anxiety	Total Score
Vigil-Colet et al. (2008)	0.87	0.82	0.92	0.91
Chiesi et al. (2011)	I	0.87	0.84	0.92
	S	0.91	0.83	0.93
Chew and Dillon (2014)	0.90	0.89	0.95	0.93
O'Bryant (2017)	0.90	0.82	0.92	0.93
Paul et al. (2018)	0.78	0.73	0.82	0.87
Grajzel (2019)	0.91	0.84	0.95	0.94

Note. I=Italy, S=Spain

Statistics Anxiety Scale-SAS (Faber et al., 2018): Faber et al. (2018) developed a statistics anxiety scale, aiming to measure the statistic anxiety of graduate students. This scale includes 17 items in total and has a three-dimensional structure. In this scale, participants were asked to state their opinions on a 4-point Likert-type scale such as 1 = *Strongly Disagree*, 2 = *Somewhat Agree*, 3 = *Strongly Agree* and 4 = *Fully Agree*. The dimensions of this scale and the number of items in the dimensions are Worry (8 items), Avoidance (4 items), and Emotionality (5 items). One of the sample items in the worry dimension is "Despite careful preparation for a statistics exam, I would worry about not passing it". One of the items in the avoidance dimension is "If I could, I would rather take two other courses than do one statistics course". One of the items in the emotionality dimension is "I would be very uncomfortable if I had to work on a statistical problem". This scale was adapted to Turkish by Güler et al. (2019). The Turkish Internal Consistency Coefficients of the scale were 0.91 for the worry dimension, 0.83 for the avoidance dimension, 0.91 for the emotionality dimension, and 0.96 for the overall scale.

2.3. The Adaptation Process of the Scale to Turkish

The scale adaptation process generally consists of the following; obtaining the necessary permission from the authors for the adaptation of the scale, adapting the scale to the target language, piloting the adapted scale, performing validity analysis of the adapted scale, and finally performing the reliability analysis of the adapted scale.

a) Obtaining permissions for adapting the scale: Primarily adapting the SAS scale to Turkish culture, Andreu Vigil-Colet, one of the authors who developed the original scale, was contacted via e-mail. The adaptation process was started after the e-mail reported by Urbano Lorenzo-Seva, the one of the other authors in the study, stated that they approved and welcomed the adaptation of the scale to Turkish.

b) Adapting the scale to the relevant language: First, it should be known that two frequently confused concepts, "translation" and "adaptation", are different from each other. Translation is only one of the stages in the adaptation process and includes linguistic conversion from a language to a language. However, adaptation is a much more comprehensive concept and requires taking into account the cultural, psychological and linguistic differences of the scale to be adapted (International Test Commission-ITC, 2017). Two basic methods are used in the literature in the process of adapting the scale to the relevant language. These are forward and backward translation methods.

i) Forward translation: At this stage, one or more translators translate the relevant scale from its original language to the target language. Then, these translations are compared, and a form is created to reflect the common view. In this framework, the items of the scale were translated into Turkish by five experts, one in assessment and evaluation, one in psychological counseling and guidance, one in English and two in statistics. A common Turkish form was created by

comparing the scale items obtained from these experts. (At this stage, if there are items that do not comply with Turkish culture, write that you revise them. Benefit from Neşe Güler)

ii) *Back translation*: At this stage, the items of the scale translated into the target language are translated into the original language of the scale by other translators, and by comparing these translations, a form in the original language that reflects the common opinion of the translators is obtained. Then, the similarity of the scales is compared by comparing this form obtained in the original language with the back-translation method with the scale form in the original language. In this direction, the scale translated into Turkish was given to a group of three people who are experts in the language of the original scale and independent from the experts in the second stage, and these experts were asked to translate the scale from Turkish into the original language of the scale. Then, the original expressions of each item and the expressions resulting from this translation were compared one to one. As a result of the comparison, it was seen that the translation and the original scale were generally equivalent to each other and the translation process was completed. In this scale, five-point grading was adopted as in the original form, and the scale categories were named as 1 = No Anxiety (*Strongly Disagree*) and 5 = Considerable Anxiety (*Strongly Agree*).

c) *Pilot study of the adapted scale*: After this process, the scale was applied to a group of 35 people in the sample to get feedback on the comprehensibility of the translations. It is aimed to identify problematic questions by adding a question such as "If there are questions you have difficulty in understanding, please specify" at the end of the questionnaire. It was stated that there was no problem with any understanding in line with the feedbacks. In addition, in the pilot study, Cronbach's Alpha coefficient was .78 and item-total correlations were .33 (item 19) and .68 (item 14). In this context, the Turkish form of the SAS, which was prepared for application and given in Appendix-A, was created in order to test its psychometric properties (The fifth item of the scale was removed from the scale as it did not meet the conditions specified in the test of construct validity. Thus, the scale adapted to Turkish consists of 23 items).

2.4. Data Collection and Analysis

The data of the study were collected online between 07 December 2019 and 02 January 2021. In the study, within the framework of the psychometric properties of the measurements obtained with the Turkish form of the SAS; construct validity, criterion validity, internal consistency reliability and item discrimination were tested.

3. RESULT / FINDINGS

Please This part covers outputs of data analysis for the psychometric properties of the adapted form of the SAS. The findings of the statistical analyses for construct validity, reliability, criterion validity and discrimination is presented below under related headings.

3.1. Construct Validity

For construct validity, exploratory factor analysis (EFA) was performed initially and then factors found by exploratory factor analysis were checked by confirmatory factor analysis (CFA). First, in the EFA results, the suitability of the data for factor analysis was examined. In this direction, Kaiser Meyer Olkin (KMO) coefficient and Bartlett test results were examined. The KMO value exceeded the lower limit of .60 (Büyüköztürk, 2010) and was obtained as .94. This result indicates that the sample size is sufficient for factor analysis, and even very good. In addition, Bartlett test detects whether there are high correlations between variables and checks compliance with factor analysis. While the correlation between the obtained factors is desired to be minimum, the intra-factor correlation value should be maximum (Eş & Durak, 2018). In the analysis, Bartlett test was found statistically significant ($\chi^2 = 6362.336$, $sd = 253$, $p < .001$). These results show that the data are suitable for factor analysis. Later, EFA analysis

was done and the principal axis factor method was preferred in the analysis (Tan, 1999). In EFA, the value of .32 was obtained as the determining criterion based on Tabachnick and Fidell (2007) as the lower limit of factor loads. The total explained variances and eigenvalues obtained as a result of EFA are given in Table 4.

Table 4. Explained Variance and Eigenvalues as a Result of EFA.

Components	Eigenvalues	Explained Variance (%)	Total Explained Variance (%)
1	9.434	22.030	22.030
2	3.516	21.652	43.683
3	1.363	18.546	62.229

Table 4 shows that three factors with eigenvalues greater than 1 were formed as a result of EFA. The first factor explains 22.03% of the total variance, the second factor explains 21.652% of the total variance, and the third factor explains 18.546%. The cumulative amount of variance explained by the eigenvalues is 62.23% of the total variance. It can be said that this value is quite good (Karagoz, 2016). The structure of the components obtained as a result of EFA is presented in Table 5.

Table 5. Components Matrix for the Turkish Form of SAS, obtained from EFA.

Items	Components		
	1	2	3
Item-9	.830		
Item-15	.830		
Item-13	.810		
Item-14	.768		
Item-20	.756		
Item-4	.736		
Item-1	.702		
Item-11	.674		
Item-23		.823	
Item-17		.822	
Item-21		.810	
Item-12		.748	
Item-7		.715	
Item-3		.672	
Item-24		.660	
Item-6			.756
Item-10			.720
Item-22			.716
Item-2			.672
Item-18			.615
Item-16			.556
Item-8			.541
Item-19			.532

Extraction Method: Principal axis factoring

Table 5 shows that the scale consists of 23 items and 3 dimensions. Since the factor load of item-5 was obtained less than .32 as a result of EFA, this item was removed from the factor analysis. The structure obtained from the Turkish form as a result of EFA is very similar to the structure in the original language of the scale. The first factor consists of items numbered 1, 4, 9, 11, 13, 14, 15 and 20. The factor loads of these items ranged from .830 (item 9) to .674 (item 11). The second factor consists of 3, 7, 12, 17, 21, 23, and 24 items. The factor loads of these

items ranged from .823 (item 23) to .660 (item 24). The third factor is composed of items 2, 6, 8, 10, 16, 18, 19, and 22. The factor loads of these items ranged from .756 (item 6) to .532 (item 19). The dimensions obtained by considering the meaning of the items in the factors, as in the original scale, has been named respectively factor 1: exam anxiety (EA,) factor 2: asking for help anxiety (AAH), factor 3: interpretation anxiety (IA).

The structure obtained in the exploratory factor analysis was controlled by confirmatory factor analysis. Accordingly, the construct validity analysis was made with the confirmatory factor analysis and the obtained model fit indices are given in Table 6.

Table 6. Result of CFA Fit Indices for three-Factor Structure.

		Fit Indices						
χ^2	df	χ^2/df	GFI	IFI	TLI	CFI	RMSEA	SRMR
743.950	220	3.382	0.865	0.916	0.903	0.916	0.073	0.0639
		(≤ 5)	(≥ 0.85)	(≥ 0.90)	(≥ 0.90)	(≥ 0.90)	(<0.08)	(<0.08)

The fit values show that the data fit the model well. In addition, the diagram showing the model fit and factor loadings obtained as a result of CFA is given in Figure 1.

Figure 1. CFA Results of the Three-Dimensional Structure of the Turkish Form of SAS.

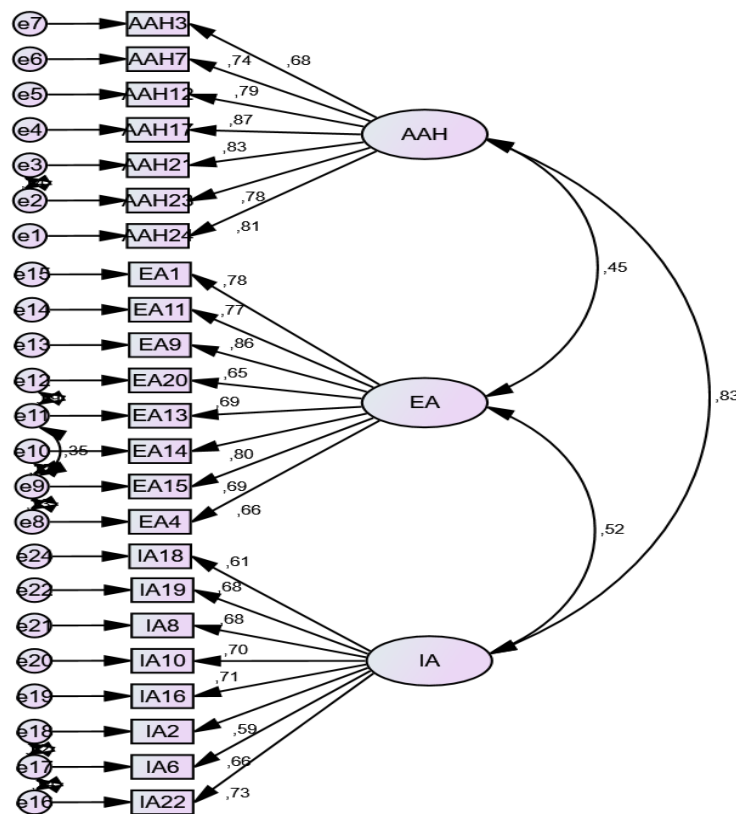


Figure 1 shows that the factor loads of the items in the asking for help anxiety (AAH) dimension ranged between .68 and .87, the factor loads of the items in the test anxiety (EA) dimension ranged between .65 and .86, and the factor loads of the items in the interpretation anxiety (IA) dimension ranged between .61 and .73. Also, Figure 1 shows that item-2 and 3 in AAH; item-8 and 9, item-9 and 10, item-9 and 11, item-11 and 12 in EA dimension; item-16 and 17, item-17 and 18 in the IA dimension were correlated each other and modified. The modified items were examined, and it was seen that the modifications made statistically were supported theoretically.

3.2. Criterion Related-Validity

The SAS scale developed by Faber et al. (2018) was used to test the criterion validity of the adapted Turkish form. The relationship between the SAS scale adapted to Turkish and the SAS scale of Faber et al. (2018) was determined by applying Pearson moments correlation analysis over the total score averages. In this direction, a positive relationship ($r = .73, p < .01$) was determined between the SAS scale adapted to Turkish and the SAS scale developed by Faber et al. This is a good level to ensure criterion validity. As a matter of fact, it can be said that the higher the level of correlation between the scales, the higher the criterion validity.

3.3. Reliability Analysis

A reliability analysis was performed by calculating the internal consistency coefficient both for the overall scale and the subscales. In addition, both Guttman and Spearman-Brown coefficients from Split-Half methods and Cronbach's Alpha were calculated as internal consistency coefficients. In this direction, the internal consistency coefficients for the overall and subscales of the Turkish form were calculated and presented in [Table 7](#).

Table 7. Internal Consistency Coefficients of Turkish Form of SAS and Three Subscales.

Scale and Subscales	Cronbach's Alpha	Split-Half Reliability	
		Guttman Coefficient	Spearman-Brown Coefficient
SAS	.934	.904	.905
AAH	.920	.899	.913
EA	.912	.873	.873
IA	.874	.830	.830

SAS= Statistics Anxiety Scale, AAH= sking for Help Anxiety, EA= Exam Anxiety, IA= Interpretation Anxiety

According to [Table 7](#), Cronbach's Alpha internal consistency coefficients were found between .874 and .934, Guttman coefficients between .830 and .904, and Spearman-Brown coefficients between .830 and .913. These values show that the SAS, adapted to Turkish, has a good level of internal consistency.

3.4. Item Analysis

In the adapted Turkish form of the OIC, the corrected item total correlations (r_{jx}) calculated to determine whether the items are discriminative or not are given in [Table 8](#). [Table 8](#) shows that item correlations take values varying between .446 and .691 (Büyüköztürk, 2010).

Table 8. Discrimination Values of the Items in the SAS.

Items	Corrected item-total correlation	Cronbach's Alpha (if item deleted)	Items	Corrected item-total correlation	Cronbach's Alpha (if item deleted)
Item-1	.650	.930	Item-14	.647	.930
Item-2	.557	.931	Item-15	.478	.933
Item-3	.600	.931	Item-16	.595	.931
Item-4	.481	.932	Item-17	.691	.929
Item-6	.568	.931	Item-18	.522	.932
Item-7	.604	.931	Item-19	.585	.931
Item-8	.660	.930	Item-20	.446	.933
Item-9	.616	.930	Item-21	.684	.929
Item-10	.585	.931	Item-22	.610	.930
Item-11	.644	.930	Item-23	.634	.930
Item-12	.679	.929	Item-24	.691	.929
Item-13	.461	.933			

3.5. Interpretation of Scores Obtained from SAS

The results EFA and item total correlation values showed that all but one of the items in the Turkish form of the SAS had sufficient factor loadings and discrimination values. Therefore, only item 5 with low factor load was removed from the adapted scale. Thus, the scores that can be taken from the scale vary between 23 and 115. Low scores from the scale indicate less anxiety, while high scores indicate a high level of anxiety. Similarly, high scores obtained from the sub-dimensions of the scale indicate a high level of asking for help anxiety, exam anxiety and interpretation anxiety.

4. DISCUSSION and CONCLUSION

In this study, SAS developed by Vigil-Colet et al. (2008) was adapted to Turkish. There is only one statistics anxiety scale developed for undergraduate students in the Turkish literature. However, it was determined that this scale was not found in any database or in the archive of the journal in which the study was published. Thus, this study contributes to the literature by filling an important gap in Turkish literature. In the study, scale adaptation has been made, generally considering the framework recommended by ITC (2017).

In the research, the construct validity of the Turkish form of SAS was examined by EFA and DFA. In addition, for the criterion validity of the adapted form, the scale developed by Faber et al. (2018) for graduate students and adapted to Turkish by Güler et al. (2019) was used. As a result of EFA using the Turkish scale form, a three-dimensional structure was obtained as in the original structure of the scale. The explained variance rate as a result of EFA was 62.229%. In the literature, there are different opinions about the explained variance rate. While Büyüköztürk (2010) stated that the explained variance rate should be at least 30%, Scherer et al. (1988) defined values of 40% and above as acceptable for the explained variance. Accordingly, it can be said that the variance ratio explained obtained as a result of this study is good. The factor loads of all items were above the lower limit of .32 (Tabachnick & Fidell, 2007) as a result of EFA. In addition, as a result of AFA, it was obtained that the number of dimensions obtained in the Turkish form is three, which is the same number as the original form of the scale. When the items that make up the dimensions as a result of EFA are examined, all items except one item gave the same result with the structure in the original form of the SAS. Only the fifth item “Asking a private teacher to explain a topic that I have not understood at all” associated with the asking for help dimension, was not included in the AFA-resulting structure. For this reason, the item five was not included in the SAS form adapted to Turkish. In this direction, it was determined that all the results obtained from EFA, which is the first of the two analyzes made for construct validity, constitute evidence for construct validity. According to the CFA results, which is the second analysis for construct validity, it was found that all fit index results of the model created for CFA were at an acceptable level. Thus, it was determined that CFA results support the three-factor structure obtained as a result of EFA.

As a result of the reliability analysis, the Cronbach's Alpha internal coefficients of the SAS and its subscales and, the Guttman and Spearman-Brown internal consistency coefficients of Split-Half Reliability methods were quite high and above the limit of .70 (Karagöz, 2016). For item discrimination, it can be said that items have good discrimination by obtaining values above .30 lower limit in the results. When the results of the study are evaluated as a whole, the SAS form adapted to Turkish can be used as a guiding scale to measure the statistics anxiety of undergraduate students, since the validity and reliability of the Turkish form of the statistics anxiety scale was ensured.

4.1. Limitations and Recommendations

In this study, as O'Bryant (2017) and Grazjel (2019) did in their study, some words in the original of the scale were slightly changed due to cultural differences and with the approval of

experts. In addition, such small changes may occur due to cultural differences as Baloglu et al. (2011) and Liu et al. (2011) revealed in their study. Using the test-retest method in future scale adaptation studies will provide even stronger support for the reliability results. In addition, for item discrimination, it is recommended to perform item discrimination not only with one method as in this study, but also by using other methods such as Ferguson Delta. Finally, it is recommended that the academicians who teach statistics should use scales that specifically measure statistics anxiety to understand whether students have high statistics anxiety and to take the necessary measures in this direction.

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:** Duzce University, 2020-272.

Authorship Contribution Statement

Ismail Durak: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing -original draft. **Yalcin Karagoz:** Methodology, Supervision, and Validation. Authors may edit this part based on their case.

ORCID

Ismail Durak  <https://orcid.org/0000-0002-8898-9639>

Yalcin Karagoz  <https://orcid.org/0000-0001-5642-6498>

5. REFERENCES

- Alexander, E. S., & Onwuegbuzie, A. J. (2007). Academic procrastination and the role of hope as a coping strategy. *Personality and Individual Differences*, 42(7), 1301-1310. <https://doi.org/10.1016/j.paid.2006.10.008>
- Baloglu, M. (2001). *An application of structural equation modelling techniques in the prediction of statistics anxiety among college students*. [Unpublished doctoral dissertation.] Texas A & M. University, Commerce, TX. Retrieved from <https://iase-web.org/documents/dissertations/01.baloglu.pdf>
- Baloglu, M., Deniz, E. M., & Kesici, S. (2011). A descriptive study of individual and cross-cultural differences in statistics anxiety. *Learning and Individual Differences*, 21, 387-391.
- Baloğlu, M. (2002). Psychometric properties of the statistics anxiety rating scale. *Psychological Reports*, 90(1), 315-325. <https://doi.org/10.2466/pr0.2002.90.1.315>
- Baloğlu, M., & Zelhart, P. (2004). Üniversite öğrencileri arasında yüksek ve düşük istatistik kaygısının ayrıştırıcıları [Discriminants of the Low-and-High Statistics Anxiety Among College Students]. *Eğitim ve Bilim*, 29(133).
- Barlow, D. H. (2002). *Anxiety and its disorders: The nature and treatment of anxiety and panic*. Guilford press. Retrieved from books.google.com
- Bektaş, H., Akman, S., & Yeşilaltay, E. (2021). İstatistik Kaygı Ölçeğinin (SAS) Psikometrik Özelliklerinin İncelenmesi [An Examination of the Psychometric Properties of the Statistical Anxiety Scale (SAS)]. *İstanbul Gelişim Üniversitesi Sosyal Bilimler Dergisi*, 8(1), 1-14. <https://dx.doi.org/10.17336/igusbd.627197>
- Benson, J., & Bandalos, D. (1989). Structural model of statistical test anxiety in adults. *Advances in test Anxiety Research*, 6, 137-154.
- Bradley, D.R., & Wygant, C.R. (1998). Male and female differences in anxiety about statistics are not reflected in performance. *Psychological Reports*, 82, 245-246. <https://doi.org/10.2466/pr0.1998.82.1.245>

- Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı [Data Analysis Handbook for Social Sciences]*. Pegem Akademi
- Chew, P. K. (2016). *An absence of attentional bias: Statistics anxiety is unique among anxieties*. [Doctoral dissertation, James Cook University].
- Chew, P. K. H. and Dillon, D. B. (2014). Reliability and validity of the Statistical Anxiety Scale among students in Singapore and Australia. *Journal of Tropical Psychology*, 4, e7. <https://doi.org/10.1017/jtp.2014.7>
- Chiesi, F., Primi, C., & Marquez, C. J. (2011). Measuring Statistics Anxiety: CrossCountry Validity of the Statistical Anxiety Scale (SAS). *Journal of Psychoeducational Assessments*, 29(6), 559–569. <https://doi.org/10.1177/0734282911404985>
- Collins, K. M. T., & Onwuegbuzie, A. T. (2007). I cannot read my statistics textbook: The relationship between reading ability and statistics anxiety. *The Journal of Negro Education*, 76(2), 118-129. Retrieved from <http://www.jstor.org/stable/40034551>
- Condrón, D. J., Becker, J. H., & Bzhetaj, L. (2018). Sources of students' anxiety in a multidisciplinary social statistics course. *Teaching Sociology*, 46(4), 346-355. <https://doi.org/10.1177/0092055X18780501>
- Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985 August). *Development and validation of an instrument to measure statistical anxiety*. Paper presented at the annual meeting of the Statistical Education Section, Chicago, IL.
- Donnelly, R. (2009). Embedding interaction within a blend of learner centric pedagogy and technology. *World Journal on Educational Technology*, 1(1), 6-29.
- Earp, S.M. (2007). *Development and validation of the statistics anxiety measure*. [Unpublished Doctorate Dissertation, The University of Denver, College of Education]. Retrieved from <https://iase-web.org/documents/dissertations/07.Earp.Dissertation.pdf>
- Emmioglu, E., & Capa-Aydin, Y. (2012). Attitudes and achievement in statistics: A meta-analysis study. *Statistics education research journal*, 11(2), 95-102.
- Eş, A., & Durak, H. S. (2018). Meslek Yüksekokulu Öğrencilerinin İş Bulma Kaygılarına Yönelik Ölçek Geliştirme: Abant İzzet Baysal Üniversitesi Örneği [Scale Development Oriented to Employment Apprehension of the Students in Vocational Schools: Abant İzzet Baysal University Example]. *Ekonomik ve Sosyal Araştırmalar Dergisi*, 14, 115-127.
- Faber, G., Drexler, H., Stappert, A., & Eichhorn, J. (2018). Education science students' statistics anxiety: Developing and analyzing a scale for measuring their worry, avoidance, and emotionality cognitions. *International Journal of Educational Psychology*, 7(3), 248-285. <http://dx.doi.org/10.17583/ijep.2018.2872>
- Ferrando, P.J., Varea, M.D., & Lorenzo, U. (1999). A psychometric study of the Test Anxiety Scale for Children in a Spanish sample. *Personality and Individual Differences*, 16, 26-33. [https://doi.org/10.1016/S0191-8869\(98\)00227-X](https://doi.org/10.1016/S0191-8869(98)00227-X)
- Fitzgerald, S. M., Jurs, S., & Hudson, L. M. (1996). A model predicting statistics achievement among graduate students. *College Student Journal*, 30, 361-366.
- Furnham, A., & Chamorro-Premuzic, T. (2004). Personality and intelligence as predictors of statistics examination grades. *Personality and individual differences*, 37(5), 943-955. <https://doi.org/10.1016/j.paid.2003.10.016>
- Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, 2(2). <https://doi.org/10.1080/10691898.1994.11910471>
- Grajzel, K. (2019). *Validation of the Statistical Anxiety Scale Among College Students in the United States*. [M. S. Dissertation, University of Colorado Colorado Springs, Kraemer Family Library]. Retrieved from <https://mountainscholar.org/handle/10976/167108>

- Güler, N., Teker, G. T., & Ilhan, M. (2019). The Turkish Adaptation of the Statistics Anxiety Scale for Graduate Students. *Journal of Measurement and Evaluation in Education and Psychology*, 10(4), 435-450.
- Hanna, D. & Dempster, M. (2009). The effect of statistics anxiety on students' predicted and actual test scores. *Irish J. Psych.*, 30, 201-209. <https://doi.org/10.1080/03033910.2009.10446310>
- Hanna, D., Shevlin, M., & Dempster, M. (2008). The structure of the statistics anxiety rating scale: A confirmatory factor analysis using UK psychology students. *Personality and Individual Differences*, 45(1), 65-74. <https://doi.org/10.1016/j.paid.2008.02.021>
- Hazrati-Viari, A., Rad, A. T., & Torabi, S. S. (2012). The effect of personality traits on academic performance: The mediating role of academic motivation. *Procedia-Social and Behavioral Sciences*, 32, 367-371. <https://doi.org/10.1016/j.sbspro.2012.01.055>
- Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, 21, 33-46. <https://doi.org/10.5951/jresmetheduc.21.1.0033>
- International Test Commission. (2017). The ITC guidelines for translating and adapting tests (2nd ed.). Retrieved from https://www.intestcom.org/files/guideline_test_adaptation_2_ed.pdf
- Joe A.I. (2005). *Fundamental Statistics for Education and the Behaviour Sciences*: Ibadan; Kraft Books Limited.
- Karagoz, Y. (2016). *SPSS 23 ve AMOS 23 uygulamalı istatistiksel analizler [Statistical Analyses with Application of SPSS 23 and AMOS 23]*. Nobel Akademik Yayıncılık.
- Köklü, N. (1996). İstatistik kaygı ölçeği: Psikometrik veriler [Statistics Anxiety Scale: Psychometric Data]. *Eğitim ve Bilim*, 20(102), 45-49.
- Liu, S., Ownhuebuzie, A. J., & Meg, L. (2011). Examination of the score reliability and validity of the statistics anxiety rating scale in a Chinese population: Comparisons of statistics anxiety between Chinese college students and their Western counterparts. *Journal of Educational Enquiry*, 11(1), 29-42.
- Malik, S. (2015). Undergraduates' Statistics Anxiety: A Phenomenological Study. *Qualitative Report*, 20(2), 120-133.
- Mji, A., & Onwuegbuzie, A. J. (2004). Evidence of score reliability and validity of the statistical anxiety rating scale among technikon students in South Africa. *Measurement and Evaluation in Counseling and Development*, 36(4), 238-251. <https://doi.org/10.1080/07481756.2004.11909745>
- Nasser, F. M. (2004). Structural model of the effects of cognitive and affective factors on the achievement of arabic-speaking pre-service teachers in introductory statistics. *Journal of Statistics Education*, 12(1). <https://doi.org/10.1080/10691898.2004.11910717>
- O'Bryant, M. J. (2017). *How attitudes towards statistics course and the field of statistics predicts statistics anxiety among undergraduate social science majors: A validation of the Statistical Anxiety Scale* [Unpublished doctoral dissertation, University of North Texas. Denton, TX]. Retrieved from <https://eric.ed.gov/?id=ED584089>
- Onwuegbuzie, A.J., & Daley, C.E. (1999). Perfectionism and statistics anxiety. *Personality and Individual Differences*, 26, 1089-1102.
- Onwuegbuzie, A.J., & Wilson, V.A. (2003) Statistics anxiety; Nature, etiology, antecedents, effects and treatments – a comprehensive review of the literature. *Teaching in Higher Education*, 8,195-209. <https://doi.org/10.1080/1356251032000052447>
- Onwuegbuzie, A.J., DaRos, D. & Ryan, J. (1997). Perfectionism and statistics anxiety: a phenomenological study. *Focus Learning Problems Math.*, 19(4), 11–35.

- Parker, R. S., Pettijohn, C. E. and Keillor, B. D. (1999), The nature and role of statistics in the business school curriculum. *Journal of Education for Business*, 75(1), 51-54. <https://doi.org/10.1080/08832329909598990>
- Paul, L., Parveen, T., Ahmed, O., & Aktar, R. (2018). Adaptation Study of The Statistical Anxiety Scale on A Bangladeshi Sample. *Bulgarian Journal of Science & Education Policy*, 12(2), 380-401.
- Pretorius, T. B., & Norman, A. M. (1992). Psychometric data on the [Statistics](#) Anxiety Scale for a sample of South African students. *Educational and Psychological Measurements*, 52, 933-937. <https://doi.org/10.1177/0013164492052004015>
- Primi, C., & Chiesi, F. (2018, July). *The role of mathematics anxiety and statistics anxiety in learning statistics*. Paper presented at the 10th International Conference on Teaching Statistics, Kyoto, Japan. Retrieved from https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_5E2.pdf
- Rindermann, H., & Neubauer, A.C. (2001). The influence of personality on three aspects of cognitive performance: Processing speed, intelligence and school performance. *Personality and Individual Differences*, 30, 829-842. [https://doi.org/10.1016/S0191-8869\(00\)00076-3](https://doi.org/10.1016/S0191-8869(00)00076-3)
- Roberts, D. M., & Bilderbeck, E. W. (1980). Reliability and validity of statistics attitude survey. *Educational and Psychological Measurement*, 40(1), 235-238. <https://doi.org/10.1177/001316448004000138>
- Sari, M. H., Arıkan, S., & Yıldızlı, H. (2017). Factors Predicting Mathematics Achievement of 8th Graders in TIMSS 2015. *Journal of Measurement and Evaluation in Education and Psychology*, 8(3), 246-265. <https://doi.org/10.21031/epod.303689>
- Scherer, R. F., Luther, D. C., Wiebe, F. A., & Adams, J. S. (1988). Dimensionality of coping: Factor stability using the ways of coping questionnaire. *Psychological Reports*, 62(3), 763-770. <https://doi.org/10.2466/pr0.1988.62.3.763>
- Spielberger, C.D. (1983). *Manual for the state-trait anxiety inventory (STAI)*. Palo Alto: Consulting Psychologists Press.
- Stoloff, M., McCarthy, M., Keller, L., Varfolomeeva, V., Lynch, J., Makara, K., Simmons, S., Smiley, W. (2009). The undergraduate psychology major: An examination of structure and sequence. *Teaching of Psychology*, 37(1), 4-15. <https://doi.org/10.1080/00986280903426274>
- Sutarso, T. (1992). Some Variables in Relation to Students' Anxiety in Learning Statistics. (ERIC Document Reproduction Service No. ED 353 334)
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Pearson Education, Inc.
- Tan, Ş. (2009). Misuses of KR-20 and Cronbach's alpha reliability coefficients. *TED Education and Science*, 34(152), 101-112.
- Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician*, 57(2), 74-79. <https://doi.org/10.1198/0003130031630>
- Vahedi, S., Farrokhi, F., Gahramani, F., & Issazadegan, A. (2012). The relationship between procrastination, learning strategies and statistics anxiety among Iranian college students: a canonical correlation analysis. *Iranian journal of psychiatry and behavioral sciences*, 6(1), 40.
- Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema*, 20(1), 174-180.
- Young, V. E. and Nelson, C. V. (1994). A survey of the impressions of economics departments of the Quantitative courses required of economics majors. *Annual Meeting of the Mid-Western Educational Research Association, Chicago, IL*.

- Zaharakar, K., (2008). *Stress Consultant*. (1st ed). Tehran: Bal University Publication, (chapter 1).
- Zanakis, S. H., & Valenzi, E. R. (1997). Student anxiety and attitudes in business statistics. *Journal of Education for Business*, 73(1), 10-16. <https://doi.org/10.1080/08832329709601608>
- Zeidner, M. (1991). Statistics and mathematics anxiety in social science students-some interesting parallels. *British J. Educ. Psych.*, 61, 319-328. <https://doi.org/10.1111/j.2044-8279.1991.tb00989.x>
- Zhou, M. (2015). Moderating effect of self-determination in the relationship between Big Five personality and academic performance. *Personality and Individual Differences*, 86, 385-389. <https://doi.org/10.1016/j.paid.2015.07.005>

6. APPENDIX

Appendix-A. Turkish Version of Statistics Anxiety Scale*

Açıklama: Lütfen istatistikle ilgili aşağıdaki durumların her birinde hissettiğiniz kaygı miktarını 1 ile 5 arasında derecelendirin. 1= “Kaygılanmam”, 5= “Çok fazla kaygılanırım” anlamına gelmektedir.

Faktör 1: Sınav Kaygısı						
1	Bir istatistik dersinin sınavı için çalışırken	1	2	3	4	5
4	Sınavdan bir gün önce, kolay sandığım bazı soruları yapamadığımı fark ettiğim zaman	1	2	3	4	5
8	Bir istatistik dersinin final sınavındayken	1	2	3	4	5
10	İstatistik sınavı olmak için sınıfa doğru giderken	1	2	3	4	5
12	Derste işlediğimiz tüm konuları gözden geçirmeden sınavdan önceki gün geldiğinde	1	2	3	4	5
13	İstatistik sınavının olacağı günün sabahında uyandığimde	1	2	3	4	5
14	Sınava girmeden hemen önce belli bir konuya hazırlanmadığımı fark ettiğimde	1	2	3	4	5
19	Çalışmak için yeterince zaman bulamadan bir istatistik sınavına girdiğimde	1	2	3	4	5
Faktör 2: Yorumlama Kaygısı						
3	İstatistik hocasından anlamakta zorlandığım bir ders kitabı/ders notu hakkında bireysel yardım isterken	1	2	3	4	5
6	Bir olasılık tablosunun (z, t, ki kare vb.) nasıl kullanılacağını dersin hocasına sorduğum zaman	1	2	3	4	5
11	İstatistikle ilgili bir alıştırmamın nasıl yapılacağını dersin hocasına sorarken	1	2	3	4	5
16	İstatistiksel bir analizin sonuç çıktısını anlamak için istatistik hocamdan yardım istediğimde	1	2	3	4	5
20	Bir istatistiksel sonuç tablosunu yorumlamak için istatistik hocamdan yardım istediğimde	1	2	3	4	5
22	Soru sormak için istatistik hocamın odasına giderken	1	2	3	4	5
23	Bir istatistiksel alıştırmamın nasıl yapılacağını anlatması için bir istatistik uzmanından ücret karşılığı yardım istediğimde	1	2	3	4	5
Faktör 3: Yardım İsteme Kaygısı						
2	Bir ders kitabındaki/ders notundaki tablonun anlamını yorumlarken	1	2	3	4	5
5	İstatistiksel analizler içeren bir ders kitabı/ders notu okurken	1	2	3	4	5
7	Bir matematiksel formülü anlamaya çalışırken	1	2	3	4	5
9	Bir otomobil reklamında yakıt tüketimi, yasal düzenlemelere uygunluk vb. özelliklerle ilgili şekilleri/oranları incelerken	1	2	3	4	5
15	İstatistik hocamın tahtaya yazılan matematiksel bir alıştırmayı açıkladığı esnada onu deftere geçirirken	1	2	3	4	5
17	Bir şans oyununda (piyango, zar vb.) kazanma olasılıklarını anlamaya çalışırken	1	2	3	4	5
18	Çözdüğü bir problemin sonuç tablosunu dikkatle inceleyen bir sınıf arkadaşımı gördüğümde	1	2	3	4	5
21	Bir gazete, kitap, makale vb. kaynakta yer alan istatistiksel analizleri anlamaya çalışırken	1	2	3	4	5

* Permission from the authors is not required for the use of the scale. Citing the source is sufficient

Validity Evidence for the Perceptions of Secondary School Students of ‘What Research is’ Scale and Measurement Invariance

Nurullah Eryilmaz ^{1,*}

¹Bath University, Faculty of Education, Department of Education, Bath, UK

ARTICLE HISTORY

Received: Jan. 22, 2021

Revised: June 24, 2021

Accepted: July 09, 2021

Keywords:

Research,
What research is,
Student perception,
Scale validation,
Measurement invariance.

Abstract: Research is a concrete action in academia which has uplifted societies’ prosperity. Although researchers have given particular attention to student perceptions about what research is in a higher education context, little attention has been given to secondary school students’ perceptions about this issue. To fill this gap, Yeoman et al. (2016) qualitatively developed an instrument measuring secondary school students’ perceptions of what research is. The present study quantitatively validates this scale using the dataset originally used to qualitatively validate it. The factor structure of the ‘what research is’ scale and measurement invariance across gender, school type, and key stage was examined. The sample is composed of 2634 secondary school students in seven schools located in East Anglia, UK. The data from this original sample showed a relatively acceptable fit to the four-factor structure after omitting some items. The result also highlighted that whilst there was evidence on configural and metric level invariance (i.e. the factor structures and the factor loadings of the scale are equivalent across gender, school type, and key stage), scalar level invariance was not met (i.e. the item intercepts of the scale are not equivalent across gender, school type, and key stage). Recommendations for future studies and future directions for research are discussed.

1. INTRODUCTION

Over the last few decades, with the expeditious advancement and development of technology, societies and organizations have become dependent on research to keep up with these changes (Bazley, 2019; Nishimura et al., 2019). Ensuring the education system's capability to integrate research-related activities to keep abreast of the advancements taking place in the world has been indispensable (Mosher, 2018; Saleem et al., 2020). Encouraging young people to give importance to research from an early age is of a growing importance in order to broaden the participation of research-related activities in the future (Moore & Hooley, 2012). Accordingly, having students participate in research related activities in their early school years is crucial to whether they choose a research related career in the future (Archer et al., 2013; Archer et al., 2020). Therefore, it would seem logical to acknowledge students' perceptions of what research

*CONTACT: Nurullah Eryilmaz ✉ ne331@bath.ac.uk 📍 Bath University, Faculty of Education, Department of Education, Bath, UK

is and how students perceive research as a potential future career choice during early school years.

Although societies have become progressively more reliant on science and technology, previous studies found that very few students are choosing subjects related to STEM (science, technology, engineering, and mathematics) or considering these areas for a future career (Archer et al., 2020; Donghong & Shunke, 2008; Moore & Hooley, 2012; Mejía-Rodríguez, 2020). There are many reasons why students do not consider science and technology as a future career choice. These include students' attitudes to science at school and parental attitudes (DeWitt & Archer, 2015; Toma & Greca, 2018). There are few studies which examine whether young students have sufficient knowledge about what research is and why people do research (Yeoman et al., 2016; Yeoman et al., 2017).

Furthermore, there is little clarity on whether there is a relationship between students' perceptions about what the research is and the education level they are at (Griffioen, 2019). For example, in studies of undergraduate students, Pearson et al., (2017) stated that students considered research experiences to be beneficial but also found them to be time-consuming. Studies such as this help us understand student perceptions of the research experience and can provide useful information for faculty that are interested in engaging students in the research process. Santos et al., (2017) highlighted that most students did not intend to pursue an academic career. For this reason, it has become more important to learn how the concept of research is shaped by young students and their attitudes towards research (Griffioen, 2019; Griffioen, 2020). This is important in determining whether students become good researchers in the future (Griffioen, 2019; Griffioen, 2020). Therefore, students' perceptions of, and attitudes towards, research at this early stage (secondary school), influence their future career choices (Yeoman et al., 2017).

The present study aims to validate a measurement instrument developed by Yeoman et al. (2016) on research attitudes and research integration that can be used with secondary school students. In this study, the researchers attempt to prove the psychometric properties of students' perception of the 'what research is' scale quantitatively. This questionnaire has been extensively validated using qualitative methods (through piloting and through building each item out of existing studies on public perceptions of research), however, the factor structure has not been validated quantitatively. This research aims to evaluate the psychometric properties of this scale and test its validity and reliability.

In the following section, we present a brief conceptual summary of the 'what research is' scale and its conceptualization. Later, we summarise previous studies that examine students' perceptions of what research is in different educational settings. Next, we illustrate our sample, variables, analytical strategy and present our findings. Lastly, we discuss our results and present implications for both policy-making and future research.

1.1. Conceptualizations of the 'What research is' scale

In this section we briefly present concepts used by Yeoman et al. (2016) to develop the 'what research is' scale –who does research?, the value of research, the process of research and myself and research.

1.1.1. Who does research?

Research is a collection of activities that includes systematically collecting, analyzing, interpreting and evaluating data, and presenting results in a consistent manner, in order to contribute to science and humanity. Scientific research is defined by Santos et al. (2017, p.45) as “a process that occurs in all areas of knowledge and therefore society depends on it”. As a general definition, people who carry out these processes are also called researchers (Çaparlar & Dönmez, 2016). According to the OECD (2015, p.162) “researchers are professionals

engaged in the conception or creation of new knowledge. They conduct research and improve or develop concepts, theories, models, techniques instrumentation, software or operational methods”.

1.1.2. The value of research

The value of research is defined by Georghiou (2015, p.4) as “consumption through its intrinsic value as a cultural good and symbol of human achievement”. There are several ways to ensure that research is valuable, effective, and of high quality (Salter & Martin, 2001). This includes increasing the stock of useful knowledge, training skilled people, creating new scientific instrumentation and methodologies, and collaborating in research projects and networks with users (Georghiou, 2015).

Specifically, in the United Kingdom, the Research Excellence Framework (REF) has given substantial attention to the assessment of the research performance of universities and is being used to make funding allocation decisions (Georghiou, 2015). Another evaluation criterion is, at the institutional level, measuring the effect of universities on the UK economy as if it were an industrial establishment (Kelly et al., 2014). As a result, it is expected that the value of research will be reflected in society economically, socially, and culturally either in the short or long term.

1.1.3. The process of research

Generally, research begins with choosing the research topic. Then, based on this, the researcher proposes the research aim, objectives, and research questions. The researcher will then comprehensively investigate what has been done so far in this area, decide on data collection methods, collect the data, and carry out data analysis. Conclusions are then drawn and lastly research papers are prepared (Brew, 2001).

1.1.4. Myself and research

Self-efficacy is defined as someones’ belief in their potential to complete a time-bound task (Bandura, 2006). ‘Myself and research’ refers to someones’ capability or self-efficacy to conduct research by him/herself (Griffioen & De Jong, 2015). In the literature, some studies have been administered which dealt with students’ capability(self-efficacy) across different subjects such as mathematics and science (Britner & Pajares, 2006; Butz & Usher, 2015) but these were limited to within a higher education context (Webb-Williams, 2017).

1.3. Studies That Examine Students Perceptions of What Research is

Considerable research has been conducted investigating the perceptions of research of undergraduate students (Ommering et al., 2020), postgraduate students (Meyer et al., 2005, 2007; Pitcher, 2011), postdoctoral researchers (Pitcher & Åkerlind, 2009), experienced researchers (Åkerlind, 2008; Brew, 2001), and postgraduate supervisors (Bills, 2004; Kiley & Mullins, 2005).

Recently, Griffioen (2019) conducted a study to examine the relationship between students’ intention to use research in their future professional practice and their perceptions of and attitudes toward research. A sample of 2192 undergraduate students in an applied sciences university in the Netherlands was used. It was found that there was a high association with students’ intention to use research in their future professional practice and their perceptions of and attitudes toward research. Furthermore, another study by Griffioen (2020) examined differences in students’ experiences of research involvement by study year (grade) and disciplines (majors) using the same sample as above. The study’s findings revealed that research involvement showed a different pattern for students across study years and disciplines. Therefore, these studies showed how the relationship between students’ perceptions of what research is and their intention to benefit from research in their professional lives might change

depending on their year of study and their discipline. It is therefore important that year of study and discipline are not overlooked.

There are very few studies that have explored the perceptions of research of secondary school pupils and the value they place on research for their future careers (Yeoman et al., 2016). Grever et al. (2008) conducted a study in the Netherlands and England with 400 young people concerning students' views on history at school and identity. The study showed that there were substantial differences between young peoples' opinions about identity and history. Another study was carried out by Schmidt et al. (2019) with 306 middle school students about their perceptions concerning the field of science and its applicability to daily life situations. They pointed out the importance and critical role of teachers in students' perception of sciences' utility for their daily activities. The more teachers make the connection with daily life, the more students consider science as useful and practical. Thus, these studies provide information about students' comprehension of school subjects and how the perception varies between young people.

With regard to measurement instrument, Visser-Wijnveen et al. (2016) developed the Student Perception of Research Integration Questionnaire (SPRIQ) to capture how students conceive research integration with 221 undergraduate students at a research-intensive university in the Netherlands. Another questionnaire developed by Griffioen (2019) the Research Attitudes in Vocational Education Questionnaire (RAVE-Q) to assess undergraduate students' attitude towards research, which consists of perceptions of research in profession, cognitive attitude towards research, positive affective attitude towards research, negative affective attitude towards research, self-efficacy towards research, the importance of research, and intuition to show research related behaviour dimensions. Moreover, Griffioen (2020) designed a questionnaire to compare lecturers' and students' higher education research integration experience based on the RAVE-Q (Griffioen, 2019) and Research Experience scale (Verbugh & Elen, 2011).

So far, to the best of our knowledge, a measurement instrument has not been developed to capture secondary school students' perceptions of what research is. To fill in the secondary school context gap, the University of East Anglia's research team conducted a project as a potential contributor to this under-researched area by exploring how pupils currently conceive research and science (Yeoman et al., 2016). As part of this project, they designed a questionnaire to gauge secondary school pupil's perception of what research is (Yeoman et al., 2016). However, this questionnaire has not been validated using quantitative methods.

1.4. The Present Study

As mentioned previously, the main purpose of the current study is to investigate the psychometric characteristics of Secondary School Students' Perception of the 'what research is' Scale, quantitatively. To this end, the scale (Yeoman et al., 2016) that was developed with secondary school students was validated using the sample of secondary school students originally used to qualitatively validate the scale. In this paper we attempt to verify the four dimensions of the 'what research is' scale – who does research, the value of research, the process of research, and myself and research – quantitatively in secondary school students (Hypothesis 1).

Research Question 1: Can the structure of this scale be confirmed quantitatively?

Some researchers investigate the relationship between students' gender, school type, and grade and their perceptions of what research is. To do this, this questionnaire should show measurement invariance across groups (Gender (male or female), School Type (state or independent), Key Stage (KS3, KS4, and KS5)). Otherwise, making comparisons between these sub-groups is problematic and researchers should be cautious about making such comparisons.

The secondary purpose of this study is to test whether the factor structure of secondary students' perception of what research is has measurement invariance across gender groups (Hypothesis 2a), across school type (Hypothesis 2b) and across Key stage (Hypothesis 2c). It is recommended that to generalise to all secondary school students the measurement invariance of the scale should be investigated for different sub-groups such as gender, school type, and grade. For this purpose, measurement invariance of the questionnaire across gender groups, school type, and key stage is examined in this study.

Research Question 2: Does this scale satisfy measurement invariance across gender, school type, and key stage?

2. METHODOLOGY

2.1. Participants

The data was gathered from secondary school students from seven schools located in East Anglia in the UK during 2014. The questionnaire was completed by 2634 secondary school students studying in these seven schools. Properties of the seven schools are presented in [Table 1](#).

There are four possible Ofsted ratings[†] (Ofsted Grade 1: Outstanding, Ofsted Grade 2: Good, Ofsted Grade 3: Requires Improvement, Ofsted Grade 4: Inadequate) that a school can receive. These Ofsted grades are based on inspectors' judgements across four Ofsted categories – quality of education, behaviour and attitudes, personal development of pupils, leadership and management as set out in their Education Inspection Framework last updated in 2019.

Table 1. School type and Ofsted rating of schools taking part in the study. ¹Rating is as determined by the Office for Standards in Education, Children's Services and Skills (Ofsted).

School	Type	Description	Key Stages Taught	Current Ofsted rating ¹
<i>A</i>	<i>State</i>	Small, mixed, rural location	KS3 and 4	Good
<i>B</i>	<i>State</i>	Large, mixed, town location	KS3, 4 and 5	Requires Improvement
<i>C</i>	<i>State(Academy status)</i>	Large, mixed, city location	KS3, 4 and 5	Requires Improvement
<i>D</i>	<i>State</i>	Large, mixed, coast location	KS5	Good
<i>E</i>	<i>Independent</i>	Small, mixed, city location	KS3, 4 and 5	Outstanding
<i>F</i>	<i>State(Academy status)</i>	Large, mixed, rural location	KS3, 4 and 5	Special Measures
<i>G</i>	<i>State(Academy status)</i>	Large, mixed, town location	KS3, 4 and 5	Good

(Adapted from Yeoman *et al.* 2016)

[†] More information can be found in (<https://thirdspacelearning.com/blog/ofsted-ratings-reports/#4-what-are-the-ofsted-ratings>).

The split between male and female participants is almost equal (1134 female, 1259 male). The majority of participants were from state schools (2200 state, 434 independent). Almost an equal number of student participants were from KS3, KS4 and KS5 (see [Table 2](#)).

Table 2. Descriptive statistics in terms of gender, school type and key stage.

Variables	Categories	Sample(n)	Percentage
Gender	<i>Male</i>	1134	%47.38
	<i>Female</i>	1259	%52.62
School Type	<i>State</i>	2200	%83.52
	<i>Independent</i>	434	%16.48
Key Stage	<i>KS3(aged 11-14) Years 7, 8 and 9</i>	928	%35.23
	<i>KS4(aged 14-16) Years 10 and 11</i>	845	%32.08
	<i>KS5(aged 16-18) Years 12 and 13</i>	861	%32.69

(Adapted from Yeoman *et al.* 2016)

All data used in this study is publicly available. Anyone who is interested in conducting research using this data or wants to check data characteristics can access the data, without permission, on this website (<http://dx.doi.org/10.5256/f1000research.7449.d108247>).

2.2. Instrument

The researchers (Yeoman *et al.* 2016) explained how they developed this questionnaire as ‘*A questionnaire was designed in a series of research team meetings in the early months of the study. Starting from one of the widely-used and reliability-tested Fennema-Sherman Mathematics Attitudes Scales (Fennema & Sherman, 1976; Wikoff & Buchalter, 1986), 25 items were constructed around the four themes who does research, the value of research, the process of research, and myself and research (6, 4, 9 and 6 items respectively). Attention was given to the inclusion of both positive and negative statements. Seven schools located in East Anglia participated (Table 1). The questionnaire was piloted to about 600 pupils in School C’ (p.4).*

The final version of the questionnaire consists of 25 items that are divided into four main themes who does research, the value of research, the process of research, and myself and research (6, 4, 9, and 6 items respectively). The questionnaire uses a 5-point Likert-type scale (*1 = strongly agree, 3 = neither agree nor disagree, 5 = strongly disagree*). The researchers included both positive and negative statements together. Q4, Q5, Q8, Q11, and Q18 behaved as negative statements in the questionnaire (see Appendix [Table A1](#)). In [Table 3](#), certain information including factors, number of items, and sample items within each factor are provided.

Table 3. Factors, number of items, and sample items.

Factors	Number of items (N)	Sample item
<i>who does research</i>	6 items (Q1, Q7, Q10, Q17, Q21, Q24)	Q1. Scientists do a lot of research.
<i>the value of research</i>	4 items (Q2, Q3, Q5, Q18)	Q2. Research is a worthwhile activity.
<i>the process of research</i>	9 items (Q9, Q12, Q13, Q14, Q15, Q16, Q19, Q20, Q22)	Q14. Research involves collecting new data.
<i>myself and research</i>	6 items (Q4, Q6, Q8, Q11, Q23, Q25)	Q6. I am confident that I can do research.

2.3. Analytical Strategy

Our analytical strategy in this study is divided into three steps: confirmatory factor analysis (CFA), correlation and internal consistency (reliability), and multi-group confirmatory factor analysis (MG-CFA).

Confirmatory factor analysis: In the main study, the scale consisted of a four-factor structure, consisting of who does research, the value of research, the process of research, and myself and research. To verify this four-factor structure, confirmatory factor analysis (CFA) was implemented on the original dataset of the sampled secondary school student groups. CFA model fit was evaluated using four traditional fit indexes. These indexes are commonly used to assess the latent construct of variables. Firstly, we use the Comparative Fit Index (CFI) and the Tucker-Lewis index (TLI) as goodness of fit statistics [the traditional cut-off value for a good model fit, (CFI) and (TLI) should be taken into account as 0.90 or higher]. We also use the root-mean-squared error of approximation (RMSEA) and the standardized root-mean-squared residual (SRMR) as residual fit statistics [the traditional threshold value for an acceptable model, (RMSEA) and (SRMR) is 0.80 or less] (Hu & Bentler, 1999; Kline, 2011).

Correlation and internal consistency: To investigate the patterns between factors and reliability within each factor, correlation, and internal consistency were tested. For correlation, Cohen (1988) suggested the cut-off point as $r \geq .224$ to identify if the correlation effect size is at least moderate. For internal consistency, reliability (internal consistency) was evaluated using Cronbach's alpha coefficient (Cronbach, 1951). This coefficient ranges from 0 to 1, with values close to 1 indicating high levels of reliability.

Multigroup invariance tests: The measurement invariance of secondary students' perception of the 'what research is' measurement model was examined across gender, school type, and key stage using multi-group confirmatory factor analysis technique (MG-CFA) (Jöreskog, 1971). In the present study, measurement invariance was investigated by running a series of statistical analyses in the subsequent order –configural, metric, and scalar invariance (Meredith, 1993; Vandenberg and Lance, 2000). This is to test if the same construct is measured and if the items of the construct are treated in the same way across subgroups (gender, school type, and key stage). The first level is configural invariance which means the same items load on the same latent variables across sub-groups. The second level is metric invariance which means factor loadings of the latent variables are constrained to be equal across sub-groups. The third and last level is scalar invariance which means the items are constrained to have the same intercepts across sub-groups. He and van de Vijver (2012, p.12) stated that “individuals who have the same score on the latent construct would obtain the same score on the observed variable regardless of their groups”.

Scalar invariance is the required condition to make valid comparisons of means of the latent construct across sub-groups.

In the literature, there are two acknowledged approaches commonly used to examine measurement invariance – one is the chi-square (χ^2) test and the other is changes in CFI and RMSEA statistics (Δ CFI and Δ RMSEA) (Cheung & Rensvold, 2002)– Employing the chi-square test to determine the overall model fit is maintained to be unsuitable due to being very sensitive to large sample sizes (Tabachnick & Fidell, 2001). Thus, the Δ CFI and Δ RMSEA values were taken into account to assess measurement invariance. The cut-off criteria (Δ CFI \leq 0.01, Δ RMSEA \leq 0.015) recommended by Chen (2007) and Cheung and Rensvold (2002) were used to test metric and scalar invariance.

In this study, all analyses were run in R statistical software (R Core Team, 2019) using the *lavaan* (Rosseel, 2012) and *semPlot* (Epskamp, 2015) packages.

3. RESULTS

3.1. Preliminary Analysis

First, as Q4, Q5, Q8, Q11, and Q18 behaved as negative statements in the questionnaire (see Appendix Table A1), these questions were reverse coded. Data were screened to check multivariate assumptions (normality, linearity, homogeneity, and homoscedasticity). For missing data value analysis, according to the suggestion by Tabachnick and Fidell (2001), cases with more than 5% item non-responses were extracted. This resulted in the removal of 275 cases. The rest of the missing data values were replaced using multiple imputation chained equations with the *mice* package (Buuren & Groothuis-Oudshoorn, 2010). This technique has flexibility in dealing with different types of variables such as binary, categorical, and continuous (Hughes *et al.*, 2014). A Mahalanobis distance ($\chi^2(28) = 56.89$) was used to detect multivariate outliers. One hundred and sixteen cases were removed using these criteria. All other assumptions were met although there were slight problems with heteroscedasticity. For further analyses, this study continued with observations from 2243 participants.

3.2. Confirmatory Factor Analysis

In Table 4, the CFA results showed that the hypothesized four-factor structure was not verified with the original secondary school student sample (Hypothesis 1) as the CFI and TLI values are less than the 0.90 cut-off suggested by Hu and Bentler (1999). The CFA unstandardized and standardized factor loadings are shown in Table 5. For standardized factor loadings, Tabachnick and Fidell (2001) stated that the correlation should be at least 0.30 or higher as lower would suggest a very weak relationship between the variables. Most standardized factor loadings were higher than 0.30, with the exception of Q4 and Q8 on the ‘myself and research’ factor; Q12, Q15, Q20, and Q22 on the ‘process of research’ factor. Therefore, these questions were removed, and a confirmatory factor analysis was then run with the remaining items.

Table 4. Confirmatory Factor Analysis Model Fit.

Fit statistics	Chi-sqaure	df	CFI	TLI	RMSEA	SRMR
Secondary school students ($n=2243$)	2204.611	269	0.732	0.701	0.057	0.059

Note: df = degree of freedom; CFI = Comparative Fit index; TLI = Tucker-Lewis index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

Table 5. *Unstandardized and Standardized Factor Loadings for Confirmatory Factor Analysis.*

Factors	Items	Unstandardized Factor Loadings	Standardized Factor Loadings
who does research	Q1	1	0.579
	Q7	1.077	0.602
	Q10	0.736	0.306
	Q17	0.911	0.427
	Q21	1.112	0.522
	Q24	0.713	0.321
the value of research	Q2	1	0.501
	Q3	1.095	0.575
	Q5	1.223	0.512
	Q18	1.070	0.557
the process of research	Q9	1	0.525
	Q12	0.669	0.254
	Q13	0.861	0.364
	Q14	0.828	0.332
	Q15	0.468	0.169
	Q16	1.055	0.490
	Q19	1.061	0.550
	Q20	0.703	0.247
	Q22	-0.258	-0.085
myself and research	Q4	1	0.117
	Q6	3.474	0.611
	Q8	0.645	0.089
	Q11	2.390	0.382
	Q23	2.776	0.434
	Q25	3.352	0.559

A new confirmatory factor analysis was executed to investigate the factor structure of the ‘what research is’ scale in this sample. In [Table 6](#), the CFA results indicated that the four-factor structure was confirmed with this sample (Hypothesis 1) as the CFI and TLI were just about within an acceptable range, around 0.90. The RMSEA and SRMR were less than the 0.80 cut-off suggested by Hu and Bentler (1999). Overall, the results of the confirmatory factor analysis indicated that the fit indexes were within an acceptable range. The CFA unstandardized and standardized factor loadings are presented in [Table 7](#). The standardized factor loadings of each item were higher than 0.30 as suggested by Tabachnick and Fidell (2001). Lastly, the measurement model including parameter estimates is provided in [Figure 1](#).

Table 6. *Confirmatory Factor Analysis Model Fit.*

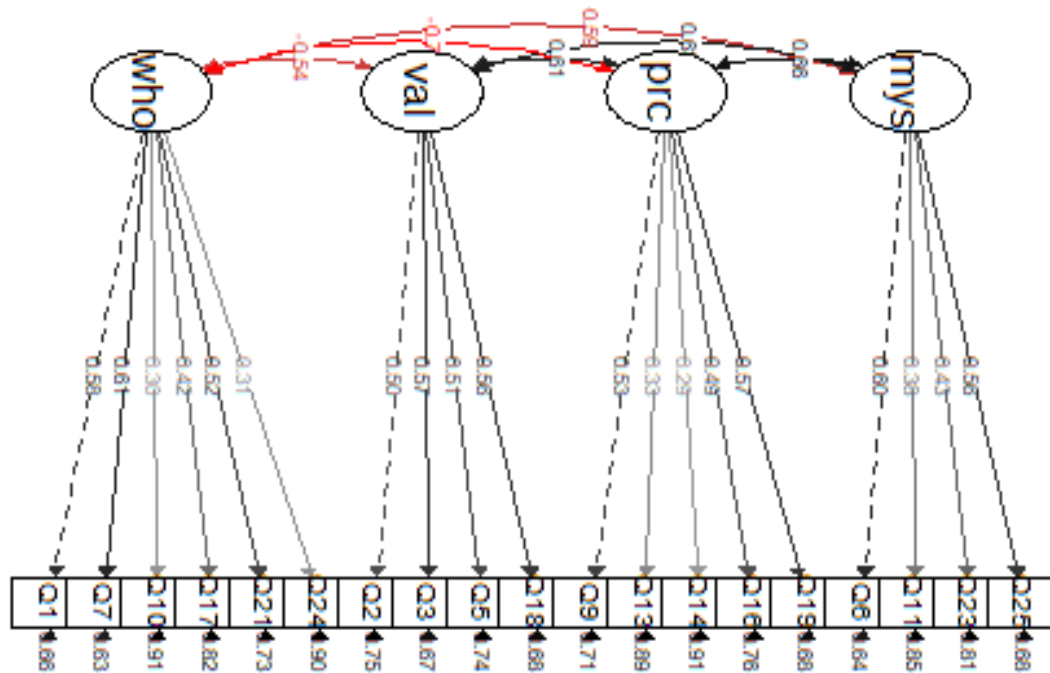
Fit statistics	Chi-square	df	CFI	TLI	RMSEA	SRMR
Secondary school students ($n=2243$)	756.264	146	0.891	0.873	0.043	0.037

Note. df = degree of freedom; CFI = Comparative Fit index; TLI = Tucker-Lewis index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

Table 7. Unstandardized and Standardized Factor Loadings for Confirmatory Factor Analysis.

Factors	Items	Unstandardized Factor Loadings	Standardized Factor Loadings
who does research	Q1	1	0.581
	Q7	1.085	0.609
	Q10	0.723	0.302
	Q17	0.892	0.420
	Q21	1.110	0.523
the value of research	Q24	0.693	0.313
	Q2	1	0.498
	Q3	1.098	0.574
	Q5	1.230	0.512
the process of research	Q18	1.085	0.561
	Q9	1	0.535
	Q13	0.773	0.333
	Q14	0.714	0.292
myself and research	Q16	1.038	0.491
	Q19	1.079	0.570
	Q6	1	0.598
	Q11	0.702	0.381
	Q23	0.816	0.434
	Q25	0.994	0.563

Figure 1. Measurement model including parameter estimates.



3.3. Correlation and Internal Consistency

In Table 8, Cronbach's alpha values of the four dimensions were within a somewhat reasonable range (ranging from 0.53 to 0.63), and factor correlations for the secondary school students sample were within a moderate range (0.21 to 0.37).

Table 8. Factor Correlations, Descriptive Statistics and Cronbach's Alpha Values of the sub-scales.

	who does research	the value of research	the process of research	myself and research
who does research	1			
the value of research	0.22	1		
the process of research	0.29	0.34	1	
myself and research	0.21	0.37	0.35	1
Max	24	17	19	16
Min	9	4	5	4
Mean	15.86	8.08	10.19	7.1
SD	2.31	2.28	2.15	1.93
Cronbach alpha (α)	0.63	0.62	0.53	0.56

3.4. Multigroup Invariance Tests

In addition to confirmatory factor analysis, measurement invariance analysis was performed to investigate if the 'what research is' measurement model was identical for gender, school type, and key stage groups (see Table 9, Table 10, Table 11).

In Table 9 we first examine configural invariance for gender groups. Configural invariance tests whether the same factor structure holds across gender. The results indicated that fit indexes were within an acceptable range. In our second step of measurement invariance, metric invariance was investigated to see if the factor loadings were identical across gender groups. The results revealed that the general adjustment measures were within acceptable ranges. In the metric invariance model, the changes in the CFI and RMSEA values were within acceptable criteria as specified by Chen (2007) and Cheung and Rensvold (2002). This result suggested that the factor loadings were identical across gender groups. CFI reduced from 0.89 to 0.86 when moving from the metric invariance model to the scalar invariance model, which is higher than the expected values. This finding indicated that the intercepts were not invariant across gender groups in the gender model.

Table 9. MGCFA Results across Gender.

Level of invariance	Chi-Square	df	CFI	TLI	RMSEA	SRMR	Δ CFI	Δ RMSEA
Baseline model	756.264	146	0.891	0.872	0.043	0.037		
Configural invariance	912.735	292	0.89	0.871	0.043	0.038		
Metric invariance	923.39	307	0.891	0.878	0.042	0.039	0.001	-0.001
Scalar invariance	1106.859	322	0.861	0.852	0.046	0.043	-0.03	0.004

Note: CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; Δ CFI = Change in values of CFI; Δ RMSEA = Change in values of RMSEA.

In Table 10, our first step is to examine configural invariance for school type groups. We use configural invariance to test whether the same factor structure holds across school type groups. Configural invariance results showed that fit indexes were within an acceptable range. For our second step of measurement invariance, metric invariance was examined to see if the constraining factor loadings were equal across school type groups. This result showed that the general adjustment values were within acceptable ranges. For metric invariance we found that changes in the CFI and RMSEA measures were within acceptable criteria set out by Chen (2007) and Cheung and Rensvold (2002). This finding suggested that the factor loadings were equal across school type groups. The CFI value decreased from 0.89 to 0.87 from the metric

invariance model to scalar invariance model, which was not within an acceptable range. This finding revealed that the thresholds were not invariant across school type groups in the school type model.

Table 10. *MGCFA Results across School Type.*

Level of invariance	Chi-Square	df	CFI	TLI	RMSEA	SRMR	ΔCFI	ΔRMSEA
Baseline model	756.264	146	0.891	0.872	0.043	0.037		
Configural invariance	887.003	292	0.894	0.876	0.042	0.037		
Metric invariance	902.531	307	0.894	0.882	0.041	0.038	0	-0.001
Scalar invariance	1027.59	322	0.874	0.867	0.044	0.04	-0.02	0.003

Note: CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; ΔCFI = Change in values of CFI; ΔRMSEA = Change in values of RMSEA.

In [Table 11](#) we first examine configural invariance across key stage groups. Configural invariance is used to test whether the same factor structure holds across key stage groups. Configural invariance results indicated that fit indexes were within an acceptable range. As the second step of measurement invariance, metric invariance was investigated to see if the factor loadings were identical across key stage groups. These results showed that the general adjustment measures were within acceptable ranges. Moving from the configural invariance model to the metric invariance model, the changes in the CFI and RMSEA measures were within the acceptable criteria set out by Chen (2007) and Cheung and Rensvold (2002). This finding indicated that the factor loadings were identical across key stage groups. CFI was reduced from 0.87 to 0.84 when moving from the metric invariance to the scalar invariance model, which was not within acceptable criteria. This finding revealed that the intercepts were not invariant across key stage groups in the key stage model.

Table 11. *MGCFA Results across Key Stage.*

Level of invariance	Chi-Square	df	CFI	TLI	RMSEA	SRMR	ΔCFI	ΔRMSEA
Baseline model	756.264	146	0.891	0.872	0.043	0.037		
Configural invariance	1118.68	438	0.88	0.86	0.045	0.042		
Metric invariance	1159.21	468	0.879	0.867	0.044	0.045	-0.001	-0.001
Scalar invariance	1405.52	498	0.841	0.836	0.049	0.05	-0.038	0.005

Note: CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; ΔCFI = Change in values of CFI; ΔRMSEA = Change in values of RMSEA.

4. DISCUSSION

The present study aimed to quantitatively validate secondary students’ perception of the ‘what research is’ scale developed by Yeoman et al. (2016) using the dataset originally used to qualitatively validate the scale. The scale was comprehensively developed qualitatively at the beginning of its development process but had not yet been quantitatively validated. In order to empirically validate the scale the factor structure was investigated, reliability analyses of the sub-scales were carried out, and measurement invariance for gender groups, school type groups, and key stage groups were examined.

Providing quantitative evidence for the proposed four-factor structure model of secondary school students’ perceptions of the ‘what research is’ scale was essential to improve the scale’s robustness and validity. First, the data from the original sample of secondary school students’

did not fully fit to the proposed four-factor structure. Nineteen items out of twenty-five were loaded reasonably acceptably on relevant unobserved factors with all factor loadings higher than 0.30. Four items –*Q12. Research involves coming up with new theories, Q15. Research always involves investigating a question, Q20. You do research to confirm your own opinion, Q22. Research is carried out solely through experiments in a laboratory* – from the *process of research* factor and two items –*Q4. People around me would not take me seriously if I said I was interested in a career in research and Q8. Doing research is challenging* – from the *myself and research* factor loaded very weak (lower than 0.30) factor loadings on their factors. Second, after omitting these items from the questionnaire, new confirmatory factor analysis was performed to verify the proposed four-factor structure model with the rest of the nineteen items. The secondary school student data demonstrated an acceptable fit to the proposed four-factor structure. The revised version of the secondary school student perception of the ‘what research is’ scale was provided in Appendix [Table A2](#).

We found reasonably moderate correlations between factors for the factor correlations patterns. The highest factor correlations between *myself and research* and *the value of research* ($r = 0.37$) and the lowest factor correlation patterns were found between ‘who does research’ and ‘myself and research’ ($r = 0.21$). The reliability analysis of each dimensions showed that every factor demonstrates a relatively moderate alpha (ranging from 0.63 to 0.53) probably due to having relatively moderate item factor loadings to some degree. Future studies may investigate this issue by either deleting some items or checking other combinations in relation to the theoretical foundations of what research is.

Supporting evidence concerning the equivalence of the factor structure of secondary school students' perception of the ‘what research is’ scale with its original dataset would increase the feasibility of the ‘what research is’ scale in comparing students’ perceptions about what research is across gender groups, school type groups, and key stage groups. To examine the measurement invariance at three hierarchical levels across gender, school type, and key stage, respectively, the configural, metric, and scalar invariance models were compared. The findings revealed that there were no significant differences between fit indexes at configural and metric level invariance, but there was scalar level invariance across gender, school type, and key stage. Thus, the British sample data satisfied the full configural and metric level invariance model but did not satisfy the scalar invariance model. These findings showed that whilst the scale had the same pattern structure and factor loadings it did not show the same item intercepts across gender, school type, and key stage. The scale allows comparisons of associations, for example, correlation and regression coefficients within gender, school type, and key stage groups. However, the mean of the scale (the average of secondary students’ perception of what research is) cannot be compared between gender groups, school type groups, and key stage groups.

There is a growing body of research that has examined students’ perceptions of what research is at higher education level or higher. However less attention has been paid to this at secondary level. The validation of the secondary school students’ perception of the ‘what research is’ scale provides insights for researchers concerning secondary school students’ perceptions of research. Accordingly, longitudinal studies could be designed to observe students’ future career paths.

4.1. Limitations

This study has limitations that should not be ignored. In the current study, all analyses were performed using the original sample of 2634 secondary school students from seven schools located only in East Anglia in the UK. Furthermore, in this sample, the majority of participants were state school students (2000 state school students, 434 independent school students), this distribution might give rise to some difficulties in terms of generalizability. To increase the generalizability of these research findings in England, the secondary school students' perception

of the ‘what research is’ scale should be implemented in larger samples drawn from other parts of England using an even school type distribution.

Although this current study contributed some concrete evidence that this scale is reliable and valid in an English sample, the secondary school students' perceptions of the ‘what research is’ scale should be investigated to determine its reliability and validity in different cultures and countries. In the present study, measurement invariance was investigated regarding gender, school type, and key stage. Future research should investigate measurement invariance across age groups, ethnicity, culture, and country as well as gender, school type, and key stage to provide more robust and valid evidence on this scale.

4.2. Conclusion

In this study, secondary school students’ perceptions of the ‘what research is’ scale were validated using the original dataset that was used to comprehensively validate the scale qualitatively. The reliability results showed that the ‘what research is’ scale can be used to assess secondary school students' perception of what research is as a moderately reliable measurement instrument. The validity results demonstrated a good fit for the “what research is” scale, which confirms the four-factor structure. The structure includes, who does research, the value of research, the process of research, and myself and research after extracting some items. Furthermore, measurement invariance results indicated that the ‘what research is’ scale has equivalence at metric invariance level across gender, school type, and key stage. Therefore, comparisons should be made cautiously across gender, school type, and key stage regarding secondary school students’ perceptions of what research is.

In conclusion, the current study should be considered as a starting point to paying attention to early years students' perceptions of what research is and whether it can predict future career aspirations.

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.

Authorship Contribution Statement

Author 1: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing - original draft.

ORCID

Nurullah Eryilmaz  <https://orcid.org/0000-0003-1916-8295>

5. REFERENCES

- Åkerlind, G. S. (2008). An academic perspective on research and being a researcher: An integration of the literature. *Studies in Higher Education*, 33(1), 17-31. <https://doi.org/10.1080/03075070701794775>
- Archer, L., Osborne, J., DeWitt, J., Dillon, J., Wong, B., & Willis, B. (2013). ASPIRES: Young People’s Science and Career Aspirations, age 10-14. <https://www.kcl.ac.uk/ecs/research/aspires/aspires-final-report-december-2013.pdf>
- Archer, L., Moote, J., MacLeod, E., Francis, B., & DeWitt, J. (2020). ASPIRES 2: Young people’s science and career aspirations, age 10-19. UCL Institute of Education. https://discovery.ucl.ac.uk/id/eprint/10092041/15/Moote_9538%20UCL%20Aspires%20%20report%20full%20online%20version.pdf

- Bandura, A. (2006). Adolescent development from an agentic perspective. In: Pajares F and Urdan T (eds) *Self-Efficacy Beliefs of Adolescents*. Information Age Publishing, pp. 1–43.
- Bazley, S. (2019). Ensuring Societal Advancement through Science and Technology: Pathways to Scientific Integration. *CUSPE Communications* <https://doi.org/10.17863/CAM.38893>
- Bills, D. (2004). Supervisors' conceptions of research and the implications for supervisor development. *International Journal for Academic Development*, 9(1), 85-97. <https://doi.org/10.1080/1360144042000296099>
- Brew, A. (2001). Conceptions of research: A phenomenographic study. *Studies in higher education*, 26(3), 271-285. <https://doi.org/10.1080/03075070120076255>
- Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching*, 43, 485-499. <https://doi.org/10.1002/tea.20131>
- Butz, A. R., & Usher, E. L. (2015). Salient sources of self-efficacy in reading and mathematics. *Contemporary Educational Psychology*, 42, 49-61. <https://doi.org/10.1016/j.cedpsych.2015.04.001>
- Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-68. <https://doi.org/10.18637/jss.v045.i03>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464-504. <https://doi.org/10.1080/10705510701301834>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modelling*, 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. <https://doi.org/10.1007/BF02310555>
- Çaparlar, C. Ö., & Dönmez, A. (2016). What is scientific research and how can it be done?. *Turkish Journal of Anaesthesiology and Reanimation*, 44(4), 212. <https://doi.org/10.5152/TJAR.2016.34711>
- DeWitt, J., & Archer, L. (2015). Who aspires to a science career? A comparison of survey responses from primary and secondary school students. *International Journal of Science Education*, 37(13), 2170-2192. <https://doi.org/10.1080/09500693.2015.1071899>
- Donghong, C., & Shunke, S. (2008). The more, the earlier, the better: Science communication supports science education. In *Communicating science in social contexts* (pp. 151-163). Springer, Dordrecht.
- Epskamp, S. (2015). semPlot: Unified visualizations of structural equation models. *Structural Equation Modeling: a Multidisciplinary Journal*, 22(3), 474-483. <https://doi.org/10.1080/10705511.2014.937847>
- Fennema, E., & Sherman, J.A. (1976). Fennema-Sherman Mathematics Attitudes Scales: Instruments designed to measure attitudes toward the learning of mathematics by females and males. *Journal Research Mathematics Education*, 7(5), 324-326. <https://doi.org/10.2307/748467>
- Georghiou, L. (2015). Value of research. *Policy Paper by the Research, Innovation, and Science Policy Experts (RISE), European Commission*. https://ec.europa.eu/research/innovation-union/pdf/expert-groups/rise/georghiou-value_research.pdf

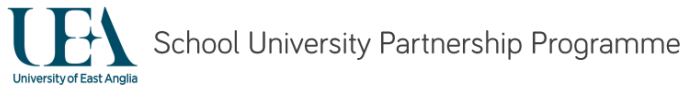
- Grever, M., Haydn, T., & Ribbens, K. (2008). Identity and school history: The perspective of young people from the Netherlands and England. *British Journal of Educational Studies*, 56(1), 76-94. <https://doi.org/10.1111/j.1467-8527.2008.00396.x>
- Griffioen, D. M. (2020). A questionnaire to compare lecturers' and students' higher education research integration experiences. *Teaching in Higher Education*, AHEAD-OF-PRINT, 1-16. <https://doi.org/10.1080/13562517.2019.1706162>
- Griffioen, D. M. (2019). The influence of undergraduate students' research attitudes on their intentions for research usage in their future professional practice. *Innovations in Education and Teaching International*, 56(2), 162-172. <https://doi.org/10.1080/14703297.2018.1425152>
- Griffioen, D. M. (2020). Differences in students' experiences of research involvement: study years and disciplines compared. *Journal of Further and Higher Education*, 44(4), 454-466. <https://doi.org/10.1080/0309877X.2019.1579894>
- Griffioen, D. M., & de Jong, U. (2015). Implementing research in professional higher education: Factors that influence lecturers' perceptions. *Educational Management Administration & Leadership*, 43(4), 626-645. <https://doi.org/10.1177/1741143214523008>
- He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, 2(2), 2307-0919. <https://doi.org/10.9707/2307-0919.1111>
- 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>
- Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., & Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14(1), 28. <https://doi.org/10.1186/1471-2288-14-28>
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36(4), 409-426. <https://doi.org/10.1007/BF02291366>
- Kelly, U., McNicoll, I., & White, J. (2014). The impact of universities on the UK economy. <http://www.universitiesuk.ac.uk/highereducation/Documents/2014/TheImpactOfUniversitiesOnTheUkEconomy.pdf>
- Kiley, M., & Mullins, G. (2005). Supervisors' conceptions of research: What are they?. *Scandinavian Journal of Educational Research*, 49(3), 245-262. <https://doi.org/10.1080/00313830500109550>
- Kline, R.B., (2011). *Principles and Practices of Structural Equation Modelling*. 3rd ed. The Guilford Press.
- Mejía-Rodríguez, A. M., Luyten, H., & Meelissen, M. R. (2020). Gender Differences in Mathematics Self-concept Across the World: an Exploration of Student and Parent Data of TIMSS 2015. *International Journal of Science and Mathematics Education*, Advance online publication,1-22. <https://doi.org/10.1007/s10763-020-10100-x>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543. <https://doi.org/10.1007/BF02294825>
- Meyer, J. H., Shanahan, M. P., & Laugksch, R. C. (2005). Students' Conceptions of Research. I: A qualitative and quantitative analysis. *Scandinavian Journal of Educational Research*, 49(3), 225-244. <https://doi.org/10.1080/00313830500109535>
- Meyer, J. H., Shanahan, M. P., & Laugksch, R. C. (2007). Students' conceptions of research. 2: An exploration of contrasting patterns of variation. *Scandinavian Journal of Educational Research*, 51(4), 415-433. <https://doi.org/10.1080/00313830701485627>
- Moore, N., & Hooley, T. (2012). Talking about career: the language used by and with young people to discuss life, learning and work. Derby: iCeGS, University of Derby.

- <https://derby.openrepository.com/bitstream/handle/10545/220535/Final%20Talking%20about%20career%20iCeGS%20Occasional%20Paper%2015062012%20NPM.pdf?sequence=8&isAllowed=y>
- Mosher, D. A. (2018). *The effect of mode of presentation, cognitive load, and individual differences on recall* [Doctoral dissertation, University of Reading]. <http://centaur.reading.ac.uk/84822/>
- Nishimura H., Kanoshima E., Kono K. (2019). *Advancement in Science and Technology and Human Societies*. In: Abe S., Ozawa M., Kawata Y. (eds) *Science of Societal Safety. Trust (Interdisciplinary Perspectives)*, vol 2. Springer, Singapore. https://doi.org/10.1007/978-981-13-2775-9_2
- OECD (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities*, OECD Publishing, Paris. <http://dx.doi.org/10.1787/9789264239012-en>
- Ommering, B. W., Wijnen-Meijer, M., Dolmans, D. H., Dekker, F. W., & van Blankenstein, F. M. (2020). Promoting positive perceptions of and motivation for research among undergraduate medical students to stimulate future research involvement: a grounded theory study. *BMC Medical Education*, 20(1), 1-12. <https://doi.org/10.1186/s12909-020-02112-6>
- Pearson, R. C., Crandall, K. J., Dispennette, K., & Maples, J. M. (2017). Students' Perceptions of an Applied Research Experience in an Undergraduate Exercise Science Course. *International Journal of Exercise Science*, 10(7), 926-941.
- Pitcher, R. (2011). Doctoral students' conceptions of research. *The Qualitative Report*, 16(4), 971-983. Retrieved from <http://www.nova.edu/ssss/QR/QR16-4/pitcher.pdf>.
- Pitcher, R., & Åkerlind, G. S. (2009). Postdoctoral researchers' conceptions of research: A metaphor analysis. *The International Journal for Researcher Development*, 1, 42-56.
- R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org/>
- 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/10.1108/1759751X201100009>
- Saleem, M. A., Eagle, L., Akhtar, N., & Wasaya, A. (2020). What do prospective students look for in higher degrees by research? A scale development study. *Journal of Marketing for Higher Education*, 30(1), 45-65. <https://doi.org/10.1080/08841241.2019.1678548>
- Salter, A. J., & Martin, B. R. (2001). The economic benefits of publicly funded basic research: a critical review. *Research Policy*, 30(3), 509-532. [https://doi.org/10.1016/S0048-7333\(00\)00091-3](https://doi.org/10.1016/S0048-7333(00)00091-3)
- Santos, M. S., Martins, J. V., Silva, A. P. F., Paula, F. G., Domingos, Á., & dos Santos, W. J. (2017). Analysis of the Influence of Undergraduate Research on the Engineering Formation from the Point of View of Students. *International Journal of Science and Engineering Investigations*, 66(6), 45-51.
- Schmidt, J. A., Kafkas, S. S., Maier, K. S., Shumow, L., & Kackar-Cam, H. Z. (2019). Why are we learning this? Using mixed methods to understand teachers' relevance statements and how they shape middle school students' perceptions of science utility. *Contemporary Educational Psychology*, 57, 9-31. <https://doi.org/10.1016/j.cedpsych.2018.08.005>
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Allyn & Bacon.
- Toma, R. B., & Greca, I. M. (2018). The effect of integrative STEM instruction on elementary students' attitudes toward science. *Eurasia Journal of Mathematics, Science and Technology Education*, 14(4), 1383-1395. <https://doi.org/10.29333/ejmste/83676>

- Vanderberg, 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), 4-70. <https://doi.org/10.1177/109442810031002>
- Verburgh, A., & Elen, J. (2011). The role of experienced research integration into teaching upon students' appreciation of research aspects in the learning environment. *International Journal of University Teaching and Faculty Development*, 1(4), 1-14.
- Visser-Wijnveen, G. J., van der Rijst, R. M., & van Driel, J. H. (2016). A questionnaire to capture students' perceptions of research integration in their courses. *Higher Education*, 71(4), 473-488. <https://doi.org/10.1007/s10734-015-9918-2>
- Webb-Williams, J. (2018). Science self-efficacy in the primary classroom: Using mixed methods to investigate sources of self-efficacy. *Research in Science Education*, 48(5), 939-961. <https://doi.org/10.1007/s11165-016-9592-0>
- Wikoff, R.L., & Buchalter, B.D. (1986). Factor analysis of four Fennema-Sherman mathematics attitude scales. *International Journal Mathematics Education Science Technology*, 17(6), 703-706. <https://doi.org/10.1080/0020739860170605>
- Yeoman, K., Bowater, L., & Nardi, E. (2016). The representation of scientific research in the national curriculum and secondary school pupils' perceptions of research, its function, usefulness and value to their lives [version 2; peer review: 2 approved]. *F1000Research*, 4, 1442. <https://doi.org/10.12688/f1000research.7449.2>
- Yeoman, K., Nardi, E., Bowater, L., & Nguyen, H. (2017). 'Just Google It?': Pupils' Perceptions and Experience of Research in the Secondary Classroom. *British Journal of Educational Studies*, 65(3), 281-305. <https://doi.org/10.1080/00071005.2017.1310179>

6. APPENDIX

Table A1. Original questionnaire



Male <input type="checkbox"/>	Year 7 <input type="checkbox"/>	Year 10 <input type="checkbox"/>	Year 12 <input type="checkbox"/>	<p><i>We thank you very much for taking the time to help us with our research!</i> <i>Kay Yeoman,</i> <i>Project Director</i></p>
Female <input type="checkbox"/>	Year 8 <input type="checkbox"/>	Year 11 <input type="checkbox"/>	Year 13 <input type="checkbox"/>	
	Year 9 <input type="checkbox"/>	State School <input type="checkbox"/>	Independent School <input type="checkbox"/>	
<p>This short questionnaire aims to explore your views on what is research, who uses it, how it is conducted, whether you see it as something useful and enjoyable, and as something that you are good at and interested in. We expect this to take no longer than 15 minutes to complete.</p>				

Please shade the box 1, 2, 3, 4 or 5, with **1** standing for **Strongly Agree** and **5** for **Strongly Disagree**. Shade 3 if you neither agree nor disagree, or if you are unsure.

	Statement	1	2	3	4	5
1.	Scientists do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2.	Research is a worthwhile activity.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3.	Knowing how to do research will help me in my future career.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4.	People around me would not take me seriously if I said I was interested in a career in research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5.	Research will not be important in my life's work.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6.	I am confident that I can do research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7.	Historians do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8.	Doing research is challenging.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9.	Research can be carried out through collecting data during a fieldtrip.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10.	Artists do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11.	You have to be a genius to do research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12.	Research involves coming up with new theories.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13.	The main purpose of research is to generate new knowledge.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14.	Research involves collecting new data.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15.	Research always involves investigating a question.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16.	Research involves searching through sources, such as libraries.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17.	Philosophers do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18.	Doing research is not useful.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19.	Research can involve collecting data through interviews and questionnaires.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20.	You do research to confirm your own opinion.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
21.	Lawyers do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22.	Research is carried out solely through experiments in a laboratory.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23.	Anybody can do research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
24.	Mathematicians do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
25.	I think I do research in school.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Table A2. Revised version of the questionnaire.

Who does research

	Statement	1	2	3	4	5
1.	Scientists do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7.	Historians do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10.	Artists do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17.	Philosophers do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
21.	Lawyers do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
24.	Mathematicians do a lot of research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

The value of research

2.	Research is a worthwhile activity.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3.	Knowing how to do research will help me in my future career.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5.	Research will not be important in my life's work.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18.	Doing research is not useful.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

The process of research

9.	Research can be carried out through collecting data during a fieldtrip.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13.	The main purpose of research is to generate new knowledge.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14.	Research involves collecting new data.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16.	Research involves searching through sources, such as libraries.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19.	Research can involve collecting data through interviews and questionnaires.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Myself and research

6.	I am confident that I can do research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11.	You have to be a genius to do research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23.	Anybody can do research.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
25.	I think I do research in school.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

MonteCarloSEM: An R Package to Simulate Data for SEM

Fatih Orçan ^{1,*}

¹Trabzon University, Faculty of Education, Department of Educational Sciences, Turkey.

ARTICLE HISTORY

Received: Oct. 02, 2020

Revised: May 23, 2021

Accepted: July 11, 2021

Keywords:

Monte Carlo Simulation,
Structural Equation Modeling,
R package, Data generation

Abstract: Monte Carlo simulation is a useful tool for researchers to estimate the accuracy of a statistical model. It is usually used for investigating parameter estimation procedure or violation of assumption for some given conditions. To run a simulation either the paid software or open source but free program such as R is needed to be used. For that, researchers must have a good knowledge about the theoretical procedures. This paper introduces the R package called MonteCarloSEM. The package helps to simulate and analyze data sets for some simulation conditions such as sample size and normality for a given model. Also, an example is given to show how the functions within the package work.

1. INTRODUCTION

Monte Carlo (MC) simulation studies are used to investigate the accuracy of statistical modeling in educational sciences as well as other social sciences. MC can be used to test such as violations of assumptions (Schumacker & Lomaz, 2010), effect of missing data or sample size on the model-data fit or parameter estimates. In the simplest terms, for example, we can examine how the t-test behaves if the small sample size is small (i.e., de Winter, 2013). Similarly, simulation studies are also used with structural equation modeling techniques. Boomsma (2013) pointed out that 31% of the studies published at the Structural Equation Modeling journal between 1994 and 2012 are MC studies. In order to run a MC study, either paid programs such as Mplus and Lisrel-PRELIS or open source but free program such as R (R Core Team, 2020) is needed to be used. In both cases, it is necessary to know how the programs work.

Even though R is a free program it has a somewhat complex structure, especially for beginners. It is because all the coding must be done by the individual or a ready-made package needed to be used. In addition to the coding difficulties, the individual must also have a good knowledge about the theoretical procedures of simulation studies.

The “MonteCarloSEM” package has been developed to facilitate these operations and to enable researchers who are not familiar with such complex operations to perform MC simulation studies. In short, this package allows to test different conditions for a given Structural Equation Models (SEM) such as confirmatory factor analysis.

To run a MC simulation, a SEM model, the true model, must first be determined: number of factors in the model, the correlation between these factors and number of items loaded on each

*CONTACT: Fatih Orçan ✉ fatihorcan@trabzon.edu.tr 📍 Trabzon University, Faculty of Education, Department of Educational Sciences, Turkey

factor and so on. Then, a model which the data will be tested with is need to be determined. The testing model does not have to be the same with the true model. If it is not the same it is called misspecified model. A simulation study where the true model and misspecified model are used could give the effect of the misspecification on the parameter estimations and the model-data fits.

Different simulation conditions can be used for simulation studies. Number of factors, values of the factor loadings, the sample size and shape of the data (i.e., normal or non-normal) are some of them. This package enables us to use some these conditions. Detailed information about the contents of this package were given below. Each function and its parameters were explained with examples. Then, a simple simulation study is given as an example.

2. MonteCarloSEM PACKAGE

The package can be installed by using `install.packages("MonteCarloSEM")` comment. The documents for the package is available at the Comprehensive R Archive Network (CRAN): <https://CRAN.R-project.org/package=MonteCarloSEM>. The first version of the package was available as of September 22, 2020. The package requires to install some other packages. For example, the “Matrix” package (Maechler & Bates, 2006) is used for the matrix decomposition. The names of the required packages were given at the package’s documents (Orçan, 2020).

There are five functions under this package. In case the arguments required for simulation are given, these functions enable us to a) generate artificial (i.e., simulated) data and b) analyze previously generated data and report the results.

2.1. Generating Artificial Data

The data sets were generated based on the standard normal distribution with mean of zero and standard deviation of one. Based on a given Confirmatory Factor Analysis (CFA) model, first uncorrelated factor scores were generated. Then, Cholesky decomposition method (Higham, 2009) was used to get correlated factor scores. On the other hand, independent error scores for each item were generated. Using correlated factor scores and random error scores, observed item scores were gained by the following formula (Orçan & Yanyun, 2016).

$$X_i = \lambda_i * F + \left(\sqrt{1 - \lambda_i^2} \right) * E_i$$

where X_i is an observed score on item i , F is factor score, λ_i and E_i indicates the factor loading and random error for item i , respectively.

If non-normality is desired, Fleishman’s power transformation method (Fleishman, 1978) was applied to obtain scores with the predefined skewness and kurtosis values:

$$X_{skewed} = A + B * X_n + C * X_n^2 + D * X_n^3$$

where X_n is a previously generated normally distributed variable. The coefficients A, B, C and D are constants corresponding to a specific skewness and kurtosis values given at Fleishman’s (1978) Table 1. For example, for skewness and kurtosis of one the values of B, C and D are 1.0174852, .190995 and -.018577, respectively.

2.2. Analyzing the Data

After the data sets were generated, the data sets can be tested with a given CFA model under “lavaan” package (Rosseel, 2012). To run the simulation under this package, first a lavaan model is needed to be defined. Definition of a lavaan model is not in the scope of this paper. Therefore, please see the lavaan documentation for the modeling strategies. Once the model is

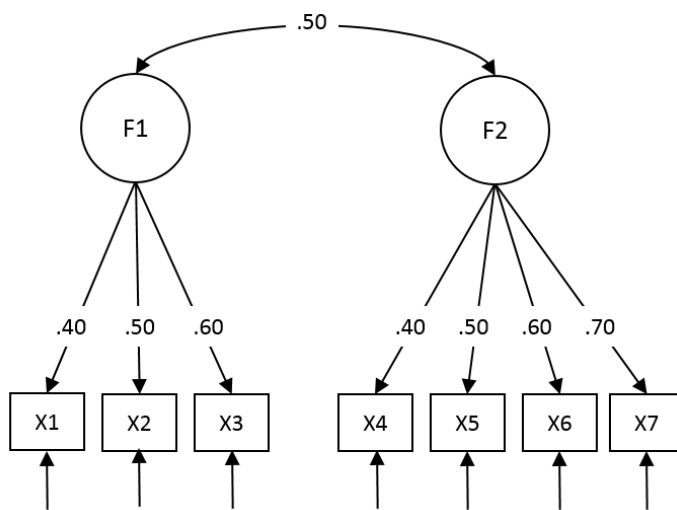
tested with the simulated data sets, the fit indices and the parameters estimated with their standard errors are printed to a Comma Separated Values (CSV) file.

The model given in [Figure 1](#) was used to explain the MonteCarloSEM functions. As seen in [Figure 1](#), there were two factors in the model (F1 and F2). The correlation between these factors was set to .5. Factors were defined with 3 and 4 items, respectively. The factor loadings of these items were indicated in the figure.

2.3. *fcors.value* Function

This function creates a symmetric matrix which specifies the correlation between the factors. If you are familiar with matrix formatting in R there is no need to use the function. Factor correlation matrix can be created by base *matrix* function. *fcors.value* function has two arguments: *nf* and *cors*.

Figure 1. The model which the data simulated based on.



- *nf*: It represents the number of factor/s in the data generation (true) model.
- *cors*: It represents the values of the correlation between the factors. Values of the correlation should be between -1 and +1. In case there is only one factor in the model, following line should be entered "`cors.value(nf = 1, cors = c(1, 1, 1))`" to define factor correlation matrix.

For the model shown in [Figure 1](#), the function should be:

```
fcors.value(nf = 2, cors = c(1, .5, .5, 1))
```

2.4. *loading.value* Function

The function specifies the factor loadings as a matrix. For each factor, the factor loadings values should be given by a column. That is, the columns represent the factors and rows represent the items. If there are *k* factors and *m* different items in the model the function creates a matrix of $k \times m$. This function also has two arguments: *nf* and *fl.loads*.

- *nf*: It represents the number of factor/s in the data generation (true) model. This value should be identical to *nf* argument at *fcors.value* function.
- *fl.loads*: This is a vector where all values of the loadings are given. The values entered should be larger than 0 and smaller than 1.

The values are assigned by columns. Therefore, firstly the values of loading for factor 1, starting from the first item to last, have to be given. For the model shown in [Figure 1](#) the function should be:

```
loading.value(nf = 2, fl.loads = c(.4, .5, .6, 0, 0, 0, 0,
                                0, 0, 0, .4, .5, .6, .7))
```

2.5. *sim.normal* Function

This function generates normally distributed data sets by a given (i.e., true) CFA model. In each data file, the first column shows sample numbers. Starting from the second all other columns show actual simulated data for each item. For the model shown in [Figure 1](#), for example, the column names could be ID, F1X1, F1X2, F1X3, F2X4, F2X5, F2X6, F2X7. Besides the data set files, the function produces two more files. The first of them is "Model_Info.dat". This file shows the factor correlation and the factor loading matrices which were used for the data generation process. The second file is "Data_List.dat". The file includes names of the data files which were generated. *sim.normal* function has five arguments: *nd*, *ss*, *fcors*, *loadings*, *f.loc*.

- *nd*: It is an integer. It shows the number of data sets which will be generated. The default values for the arguments is 10.
- *ss*: It is an integer larger than 10. It indicates sample size for the data generation process. The default values for the arguments is 100.
- *fcors*: A symmetric factor correlation matrix. In order to define this argument *fcors.value* function would be used.
- *loadings*: It is the factor loading matrix. Columns represent factors and non-zero rows represent number of items under each factor. In order to define this argument *loading.value* function would be used.
- *f.loc*: It indicates where the simulated data sets will be saved. In order to run the function successfully, a file path has to be defined.

Let's now assume that we would like to generate 100 data sets by the model given in [Figure 1](#). The sample size is 500 and the data sets will be saved *Documents* folder under C driver. Following codes do this simulation.

```
FCorr <- fcors.value(nf = 2, cors = c(1, .5, .5, 1))
FLoad <- loading.value(nf = 2, fl.loads = c(.4, .5, .6, 0, 0, 0, 0,
                                           0, 0, 0, .4, .5, .6, .7))
FileLoc <- "C:/Users/user/Documents"
sim.normal(nd = 100, ss = 500, fcors = FCorr, loading = FLoad, f.loc = FileLoc)
```

2.6. *sim.skewed* Function

The function is similar to *sim.normal* function except that *sim.skewed* generates non-normally (skewed) distributed data sets by a given CFA model. The function generates data files, model information and data list files as *sim.normal*. However, the model information file has two more information under this function. One of them indicates if the items were non-normal or not. The second one shows Fleishman's B, C and D values which were used for data generation procedure. *sim.skewed* function has seven arguments: *nd*, *ss*, *fcors*, *loadings*, *nonnormal*, *Fleishman* and *f.loc*. The details about new arguments are given here. The others are the same as in *sim.normal* function.

- *nonnormal*: It is a vector of 0 or 1s. Each value in the vector represents an item in the model. If there are seven items in the model, for example, the length of the vector has to be seven. 0 indicates if the item is distributed normally while 1 indicates non-normal distribution for the corresponding item.

- **Fleishman:** It shows B, C and D values from Fleishman’s power method. Since $A = -C$, the value of A is not needed here.

Again, let’s assume that we would like to generated 100 data sets by the model given in [Figure 1](#). Also, assume that only the item X6 and item X7 were non-normally distributed with skewness and kurtosis of 1. The sample size is 500 and the data sets will be saved *Documents* folder under C driver. Following codes do this simulation.

2.7. *fit.simulation* Function

sim.normal and *sim.skewed* functions generate data sets and save them into the specified folder. However, *fit.simulation* function uses these data sets and runs a CFA model, which does not have to be the same with the true model. Therefore, in order to run the function, the data sets have to be generated previously.

```
FCorr <- fcors.value(nf = 2, cors = c(1, .5, .5, 1))
FLoad <- loading.value(nf = 2, fl.loads = c(.4, .5, .6, 0, 0, 0, 0,
                                           0, 0, 0, .4, .5, .6, .7))
IfNon <- c(0, 0, 0, 0, 0, 1, 1)
FleisV <- c(1.0174852, .190995, -.018577)
FileLoc <- "C:/Users/user/Documents"
sim.skewed(nd = 100, ss = 500, fcors = FCorr, loading = FLoad, nonnormal = IfNon,
           Fleishman = FleisV, f.loc = FileLoc)
```

The “lavaan” package (Rosseel, 2012) is used to fit the model to the data sets. Please see the lavaan documentation for more detailed information about how to build a CFA model. Once the model run is done, fit indices and parameters estimated with their standard errors are printed to a Comma Separated Values (CSV) file named as “All_Results.csv”. Each line in the output file represents results of a simulated data set. The columns such as “chisq”, “df”, “cfi” or “rmsea” are self-explanatory but the second column named as “Notes”. This column shows the model-data convergence. If the model is converged without any problem the value will be “CONVERGE”. If it is not converged the value will be “NONCONVERGE” and all the other values in the row will be “NA”. Lastly, if there were some kind of warning such as negative variance during the model run the value will be “WARNING” and based on the types of warnings, some of the values in the row might be “NA”.

To run the simulation, the generated data sets and the list of the data sets’ names (*Data_List.dat*) have to be located at the same folder. *fit.simulation* function has five arguments: *model*, *PEmethod*, *datalist* and *f.loc*.

- **model:** It indicates the lavaan model. The model will be used to test the generated data sets. Therefore, it does not have to be the same with data generation (true) model.
- **PEmethod:** It indicates the parameter estimation method. The default estimation method is Maximum Likelihood (ML). Other estimation methods such as Robust Maximum Likelihood (MLR) or Weighted Least Squares (WLS) are available under lavaan package. Please see the lavaan documentation for more information.
- **dataList:** It shows the name of the file which has list of the data sets’ names. If *sim.normal* and *sim.skewed* functions were used for the data generation, as it was pointed earlier the name of the file is “Data_List.dat”. However, data sets generated with other statistical programs or r-packages can also be used with this function. In that case, name of the file might be different. Therefore, it should be specified here.

- `f.loc`: It indicates where the simulated data sets are located. The `dataList` file and the simulated data sets have to be in this location.

Let's run a simulation based on previously generated data sets from `sim.skewed` function. For this, first a model needed to be defined by `lavaan`. The model given below (LavaanM) defines the model given in Figure 1.

```
LavaanM <- '
# CFA Model
f1 =~ NA*x1 + x2 + x3
f2 =~ NA*x4 + x5 + x6 + x7
# Factor Correlations
f1 ~~ f2
# Factor variance
f1 ~~ 1*f1
f2 ~~ 1*f2
'

DList <- "Data_List.dat"
FileLoc <- "C:/Users/user/Documents"
fit.simulation(model = LavaanM, PEmethod = "MLR", dataList = DList, f.loc =
FileLoc)
```

Once the simulation process is completed the result of the simulation is saved as a CSV file to the location. A part of the CSV file was displayed in Figure 2. As it is seen from the figure, the first column showed replication number. For example, the results which were obtained from the first data set was given at the replication number 1. Based on the results, the first data set was converged. The chi-square value for the model was 6.351 with 13 degrees of freedom. The CSV file also shows fit indices such as the comparative fit indices (CFI) and the standardized root mean square residual (SRMR). Their values for the first data set were 1 and .018, respectively. The column W in the file shows the values of the factor loadings from item X1 to factor F1. The value for the first data was .403. Following two columns (X and Y) shows standard errors of the estimates and the p-values. The standard error of the factor loading for the first data set was .061, for example. On the other hand, column Z (std.est) shows values of the standardized parameter estimates (.411).

Figure 2. Sample view of "All_Results.csv" file.

	A	B	C	D	E	M	O	W	X	Y	Z	AA	AB	AC
1	rep#	Notes	chisq	df	pvalue	cfi	srmr	f1=~x1	se	pvalue	std.est	std.se	pvalue	f1=~x2
2	1	CONVERGE	6.351	13	0.932	1	0.018	0.403	0.061	0	0.411	0.058	0	0.569
3	2	CONVERGE	11.928	13	0.534	1	0.025	0.449	0.057	0	0.459	0.055	0	0.534
4	3	CONVERGE	17.553	13	0.175	0.985	0.03	0.471	0.057	0	0.458	0.052	0	0.441
5	4	CONVERGE	11.047	13	0.607	1	0.025	0.343	0.063	0	0.34	0.059	0	0.493
6	5	CONVERGE	9.534	13	0.731	1	0.02	0.357	0.063	0	0.345	0.059	0	0.566
7	6	CONVERGE	24.218	13	0.029	0.969	0.034	0.413	0.063	0	0.421	0.059	0	0.446
8	7	CONVERGE	12.578	13	0.481	1	0.026	0.364	0.054	0	0.371	0.054	0	0.641
9	8	CONVERGE	13.15	13	0.436	1	0.024	0.417	0.057	0	0.44	0.055	0	0.424
10	9	CONVERGE	8.488	13	0.81	1	0.02	0.424	0.058	0	0.412	0.053	0	0.528
11	10	CONVERGE	13.048	13	0.444	1	0.023	0.421	0.061	0	0.422	0.058	0	0.529
12	11	CONVERGE	17.445	13	0.18	0.988	0.029	0.339	0.066	0	0.332	0.062	0	0.407
13	12	CONVERGE	10.255	13	0.673	1	0.023	0.438	0.063	0	0.438	0.059	0	0.519
14	13	CONVERGE	11.874	13	0.538	1	0.022	0.356	0.053	0	0.375	0.054	0	0.643

3. EXAMPLE SIMULATION STUDY

An example simulation is given here to demonstrated functions in the package. Let's assume that a researcher wanted to see the effects of nonnormality on the parameter estimation under a CFA model. For this purpose, a CFA model was defined which has three factor and each factor was loaded by three items. Factor loadings for the model were all equal to .7. Besides, the correlation among the factors were set to .5.

In order to create a manageable example only three conditions were simulated with sample size of 500 for 1000 times: All items were a) normally distributed (skewness = 0, kurtosis = 0), b) skewed (skewness = 1, kurtosis = 1) and c) skewed (skewness = 1.5, kurtosis = 2.5). The true model then used to get the parameter estimates under each conditions. Following codes simulate normal data sets at the step 1 and analyze the simulated data sets at the step 2.

```
# First, the package has to be installed.
install.packages("MonteCarloSEM")
library("MonteCarloSEM")

## Step 1: Simulate normally distributed data sets
# Define correlation among the factors

Ex.FCorr <- fcors.value(nf = 3, cors = c(1, .5, .5,
                                       .5, 1, .5,
                                       .5, .5, 1))

# Define factor loadings
Ex.FLoad <- loading.value(nf = 3, fl.loads = c(.7, .7, .7, 0, 0, 0, 0, 0, 0,
                                              0, 0, 0, .7, .7, .7, 0, 0, 0,
                                              0, 0, 0, 0, 0, 0, .7, .7, .7))

# Define where the simulated data sets will be saved
Ex.FileLoc <- "C:/Users/user/Documents"
sim.normal(nd = 1000, ss = 500, fcors = Ex.FCorr, loading = Ex.FLoad, f.loc =
  Ex.FileLoc)

# Step 2: Analyze simulated data sets with the true model.

# Define CFA model under lavaan package
Ex.CorrectM <- '
# CFA Model
f1    =~ NA*x1 + x2 + x3
f2    =~ NA*x4 + x5 + x6
f3    =~ NA*x7 + x8 + x9
# Factor Correlations
f1    ~~ f2
f1    ~~ f3
f2    ~~ f3
# Factor variance
f1    ~~ 1*f1
f2    ~~ 1*f2
f3    ~~ 1*f3
'
```

```
# Define the folder which contains names of the simulated data sets
Ex.DList <- "Data_List.dat"

# Define where the data sets are located
Ex.FileLoc <- "C:/Users/user/Documents"

fit.simulation(model = Ex.CorrectM, PEmethod = "ML", dataList = Ex.DList, f.loc =
  Ex.FileLoc)
```

Similarly, following codes simulates skewed data (skewness = 1 and kurtosis = 1) at step 1 and runs simulated data set at step 2. Since the same model was used for the simulations, step 2 is the same as normal data study above. Therefore, the same codes were used for the skewed data at step 2 and codes were not given again. Also, to run the second skewed data study the Ex.FleisV values should be replaced by .9920986, .3452694 and -.0418153 respectively. Later, the step 1 and the step 2 should be run again for the second skewed data simulation.

```
## Step 1: Simulate first skewed (non-normal) data sets.

# Define correlation among the factors
Ex.FCorr <- fcors.value(nf = 3, cors = c(1, .5, .5,
                                         .5, 1, .5,
                                         .5, .5, 1))

# Define factor loadings
Ex.FLoad <- loading.value(nf = 3, fl.loads = c(.7, .7, .7, 0, 0, 0, 0, 0, 0,
                                               0, 0, 0, .7, .7, .7, 0, 0, 0,
                                               0, 0, 0, 0, 0, 0, .7, .7, .7))

# Define which items are non-normal.
Ex.IfNon <- c(1, 1, 1, 1, 1, 1, 1, 1, 1)

# Define Fleishman's values: B, C and D
Ex.FleisV <- c(1.0174852, .190995, -.018577) # Values for Sk = 1, K = 1

# Note: Replace these values with following values for Sk = 1.5, K =
2.5: .9920986, .3452694 and -.0418153

# Define where the simulated data sets will be saved
Ex.FileLoc <- "C:/Users/user/Documents"

sim.skewed(nd = 1000, ss = 500, fcors = Ex.FCorr, loading = Ex.FLoad, nonnormal =
  Ex.IfNon, Fleishman = Ex.FleisV, f.loc = Ex.FileLoc)

# Step 2: Run simulated data sets with the correct model.
## This was the same as above. Therefore, use the same codes at step 2 given
under normal data study.
```

4. RESULTS OF THE EXAMPLE

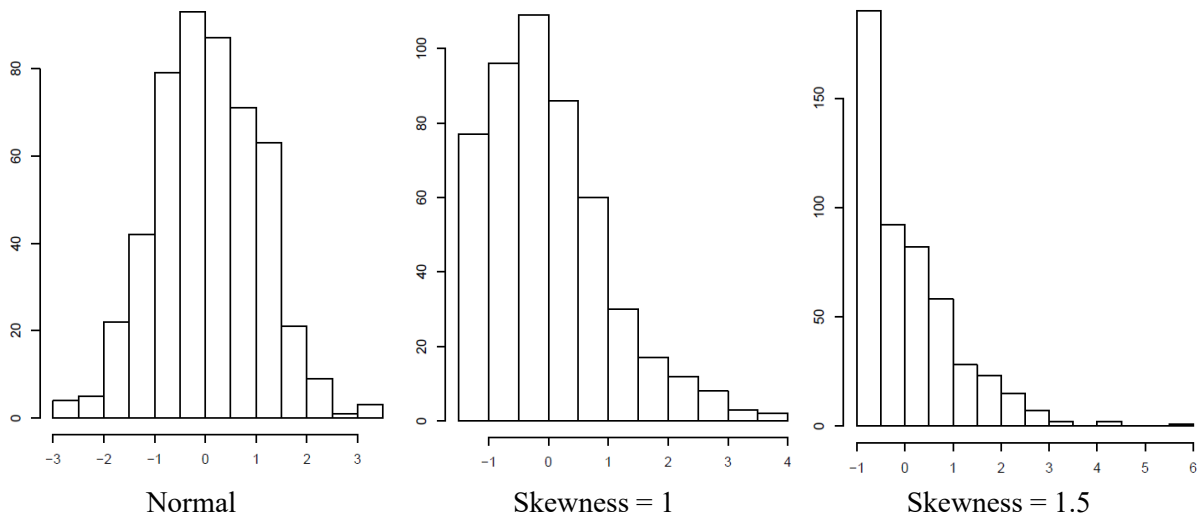
First, sample histograms for item scores generated under each simulation condition were pictured in Figure 3. The left panel shows an item scores under normal data generation while the right panel shows an item scores under skewness = 1.5 conditions. The actual skewness values of the specific items were .07, .99 and 1.46 respectively.

Table 1 summarizes the results of the simulation studies. The first row shows the average Chi-squares values. The second row shows that as the skewness increased the number of chi-square test which had p-values smaller than .05 were also increased. When the data were normal only 4.5% of the models showed no-model-data fit based on the chi-square tests. However, when the skewness increased to 1.5, the ratio increased to 17.4%. Since the factor loading had similar values only four of them were reported in table 1. Based on the results, as it was expected, the values of the parameter estimates were equal to the true values under the normally distributed data. However, as skewness increased the values of the parameter estimates got worse. Similar conclusions can also be made for the values of the correlations among factors. Therefore, as it was expected, as skewness increased the parameter estimated got worsen.

5. CONCLUSION

MonteCarloSEM package was introduced in this study. The package is useful for simulation data sets for a given model and analyzing the simulated data sets. There are five function in the package. This study described the functions and gave examples for the usage of the functions. Also, R codes for an example simulation study were provide in the study.

Figure 3. *The model which the data simulated based on.*



With the help of this package, researchers do not have to pay for expansive software to simulated data sets. Besides, the researchers can do simulation studies with little knowledge about R programing, considering simulating data for a given model is relatively complex and requires some knowledge about the formulas.

Table 1. *The Results of Example Simulation.*

		Normal	Skewness = 1	Skewness = 1.5
Chi-Square		23.84	26.85	29.01
Number of $p < .05$		45	113	174
Factor Loadings	Item 1	.70	.68	.65
	Item 2	.70	.68	.65
	Item 8	.70	.68	.65
	Item 9	.70	.68	.65
Factor Correlations	Factor 1 and 2	.50	.49	.47
	Factor 1 and 3	.50	.49	.46
	Factor 2 and 3	.50	.49	.46

The first version of the package is available now at CRAN (<https://github.com/cran/MonteCarloSEM>). The package has some limitations for now. However, some new functions are under construction now. For example, it is planned to add a function which creates missing data based on missing completely random (MCAR), missing at random (MAR) and missing not at random (MNAR) to the package for the later versions. Besides this, functions which creates item parcels and categorical (i.e., Likert) data sets were also under construction.

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

Fatih Orçan  <https://orcid.org/0000-0003-1727-0456>

6. REFERENCES

- Boomsma, A. (2013) Reporting Monte Carlo Studies in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(3), 518-540. <https://doi.org/10.1080/10705511.2013.797839>
- de Winter, J.C.F. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, 18, 10. <https://doi.org/10.7275/e4r6-dj05>
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532. <https://doi.org/10.1007/BF02293811>
- Higham, N.J. (2009). Cholesky factorization. *WIREs Computational Statistics*, 1, 251-254.
- Maechler, M., & Bates, D. (2006). *2nd Introduction to the Matrix package*. URL: <https://cran.r-project.org/web/packages/Matrix/vignettes/Intro2Matrix.pdf>
- Orçan, F. & Yanyun, Y. (2016). A Note on the Use of Item Parceling in Structural Equation Modeling with Missing Data. *Journal of Measurement and Evaluation in Education and Psychology*, 7 (1), 59-72. <https://doi.org/10.21031/epod.88204>
- Orçan, F. (2020). MonteCarloSEM 0.0.1. <https://CRAN.R-project.org/package=MonteCarloSEM>
- R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved September 10, 2020, from <http://www.R-project.org/>
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48 (2), 1-36. URL: <http://www.jstatsoft.org/v48/i02/>
- Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). Routledge.

Investigating Invariant Item Ordering in Intelligence Tests: Mokken Scale Analysis of KBIT-2

Eren Halil Ozberk^{1,*}, Elif Bengi Unsal Ozberk¹, Sait Uluc², Ferhunde Oktem³

¹Trakya University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, 22100, Edirne, Turkey

²Hacettepe University, Faculty of Letters, Department of Psychology, 06800, Ankara, Turkey

³Hacettepe University, Faculty of Letters, Department of Psychology (Retired), 06800, Ankara, Turkey

ARTICLE HISTORY

Received: Jan. 11, 2021

Revised: June 14, 2021

Accepted: July 11, 2021

Keywords:

Mokken scale analysis,
Intelligence tests,
Invariant item ordering.

Abstract: The Kaufman Brief Intelligence Test – Second Edition (KBIT-2) is designed to measure verbal and nonverbal abilities in a wide range of individuals from 4 years 0 months to 90 years 11 months of age. This study examines both the advantages of using Mokken Scale Analysis (MSA) in intelligence tests and the hierarchical order of the items in the KBIT-2: Turkish form by estimating the parameters of each of the three subtests by testing the dimensionality of the KBIT-2 subtests by using the Invariant Item Ordering (IIO) assumptions. 2850 people participated in the study, including children, adolescents, and adults. Participants' ages varied from 48 months (4 years 0 months) to 539 months (44 years 11 months). Automated Item Selection Procedure (AISP) was applied for the assessment of unidimensionality under three different lower bounds as 0.30, 0.40, and 0.55. The items of all three subtests formed a unidimensional scale. Backward Item Selection (BIS) procedure detected seven items in the Matrices and 17 items in the Verbal Knowledge, while six items in the Riddles subtest violated the IIO criteria. KBIT-2: Reliability values obtained using MSA analysis show that all three subtests have a high degree of internal consistency. However, care should be taken when IIO assumptions do not fit the intelligence scales in the original form.

1. INTRODUCTION

The Kaufman Brief Intelligence Test – Second Edition (KBIT-2) is designed to measure verbal and nonverbal abilities in a wide range of individuals from 4 years 0 months to 90 years 11 months of age (Kaufman & Kaufman, 2004). The first version of the test, KBIT, consisted of only two subtests: Vocabulary and Matrices (MT). Vocabulary subtest aimed to measure crystallized intelligence with questions focusing on expressive language skills and general knowledge gained through school. It is widely accepted that the MT subtest, which includes pictures or abstract patterns, is a good indicator of fluid intelligence (such as non-verbal abilities and instant problem-solving skills) (Cole & Randall, 2003).

KBIT-2, especially the verbal section, was revised within Cattell–Horn–Carroll Theory (CHC) after a comprehensive renovation and norm adjustment study. The number of Vocabulary

*CONTACT: Eren Halil Özberk ✉ erenozberk@trakya.edu.tr 📍 Trakya University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education, 22100, Edirne, Turkey

subtests in the first version was divided into two separate subtests: Verbal Knowledge (VK) and Riddles (RD). Since the test is designed to measure Verbal and Nonverbal intelligence in a wide range of ages, it is essential to start from an item likely to measure the desired latent trait for a given age group and stop the test after a varying number of consecutive incorrect responses.

Starting and discontinue rules are used in various intelligence tests to reduce the burden, shorten the testing time, and minimize error scores, which prevents respondents from answering easy questions far below their abilities (Kaufman & Kaufman, 2004; von Davier et al., 2019). To apply the starting and discontinue rule for each ability group, intelligence test batteries are designed to start with easy to difficult items consecutively for each subtest.

The Turkish Ministry of National Education standardized the test as part of the project called Empowering Special Education (ESE). KBIT-2 has been widely used to identify the children in need of special education in order to decide whether they should have that special education since it was adapted in Turkey. The test has also been of great interest in scientific studies. The validity and reliability of the KBIT-2 studies have been tested many times using item analysis, internal consistency, and split-half consistency, which are all based on the Classical Test Theory (CTT) (Atalay, 2007; Öktem, 2016; Savaşan, 2006; Uluç et al., 2015). Although the studies on CTT provide information about the test's psychometric characteristics, they have several limitations. In CTT, item characteristics such as item difficulty and item discrimination are group dependent (Hambleton et al., 1991), which means the parameter estimations of item difficulty and discrimination change when the group changes. Also, estimated errors are considered to be equal for all individuals irrespective of their intelligence levels.

1.1. Nonparametric Item Response Theory in Psychological Tests

Parametric Item Response Theory (IRT), also called 'latent trait theory,' was developed against the limitations of CTT in the test development, adaptation, and evaluation of measurement tools in education and psychology (Lord & Novick, 1968; Embretson & Reise, 2000; Hambleton et al., 1991). IRT focuses on an individual's responses to each item rather than the total scores obtained from the test.

Numerous studies have been conducted on the advantages of using IRT in developing tests for psychological structures (Embretson & Reise, 2000). Ability measures obtained from the tests designed according to IRT can be obtained independently from the sample of the items applied to the individual. When the model-data fit is achieved, IRT methods reveal more accurate items and ability parameter estimates than CTT does. (Hambleton et al., 1991). Precise parameter estimates are an essential part of intelligence test development; thus, they are so widely used and much research prefers parametric IRT methods to develop psychological structures (Robie et al., 2001; Steinberg, 1994; Waller et al., 2000).

Empirical research has suggested that the nonparametric approach should be preferred over the parametric approach, especially in psychological scales (Meijer et al., 1990; Meijer & Baneke, 2004; Reise & Waller, 2003). In contrast to the large-scale tests used in education, it is not always possible to meet the parametric IRT assumptions in the tests that measure psychological structures. In parametric IRT models, item and ability parameters are estimated with one, two, or three parameters logistic models or normal ogive models. If the unidimensionality and local independence assumption criteria are not met, the item and ability parameter estimates become uncertain. Nonparametric models are less restrictive about the shape item response functions (IRF) (Sijtsma & Van der Ark, 2017). Even though IRFs do not fit logistically as in nonparametric models, they should be in an increasing form.

In nonparametric models, individuals and items ordered according to total scores reflect a latent continuum scale (Meijer & Baneke, 2004). Also, Junker and Sijtsma (2001) state that it is more

advantageous to use the nonparametric IRT method in psychological and sociological studies when the sample size is low. One of the most known nonparametric methods is Mokken Scale Analysis (MSA), proposed by Mokken (1971).

1.2. Mokken Scale Analysis Overview

Mokken (1971), contrary to Guttman's deterministic model, developed a probabilistic nonparametric method. MSA can be used when items are in a hierarchical order to test the relationships between items and the latent ability (Sijtsma & Van der Ark, 2017). The individuals' observed scores are obtained through the sum of the scores on the original scale, while mean item scores are obtained from item scores. Mokken model uses two models to evaluate scales.

The first model is called the Monotone Homogeneity Model (MHM) (Mokken, 1971; Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002; Sijtsma & Molenaar, 2016). MHM is a non-restrictive model that aims to rank individuals (Sijtsma & Van der Ark, 2017). In the MHM, there are unidimensionality, local independence, and monotonicity assumptions. The second model is called the Double Monotonicity Model (DMM). Unlike MHM, the DMM aims to rank individuals and items simultaneously. In the DMM, items are ordered using mean item scores. The equivalent of mean item scores in CTT is the item difficulty. In many intelligence tests, items are ordered from easiest to most difficult, aiming to reduce test anxiety by taking easy questions first and helping practitioners apply the starting point and discontinue rule easily. Item order must be equal for all intelligence score levels to make a fair and unbiased evaluation. At this point, the DMM can provide a practical solution for this situation using invariant item ordering (IIO). The DMM model includes all the assumptions of the MHM model besides nonintersecting IRFs as the fourth assumption. MHM and DMM can be used for dichotomous and polytomous items (Molenaar, 1997; Sijtsma et al., 1990).

There are three different scalability coefficients: MSA item scalability coefficient (H_i), item-pair scalability coefficient (H_{ij}), and total scalability coefficient (H). Also, the H transposed scalability coefficient (H_T) is used in IIO analysis to express the respondents' consistency of invariant item orders (Ligtvoet et al., 2010; Sijtsma & Meijer, 1992). All scalability coefficients can take a range of values from 0 to 1 (Wind, 2017). H_i can also be defined as item discrimination (Sijtsma et al., 2011) that high H_i values are a proof of a highly discriminating item. The H_{ij} coefficient is an indicator of the internal consistency of each item pair. High values indicate that item pairs have high internal consistency. H total scalability coefficient is known as the coefficient indicating the whole scale's quality according to Mokken model (Mokken, 1971; Molenaar & Sijtsma, 2000). The scale can be evaluated according to the H coefficient. Similarly, IIO accuracy is interpreted by the H_T coefficient.

1.2.1. Assumptions of the Mokken Model

There are several assumptions in Mokken models as in parametric models:

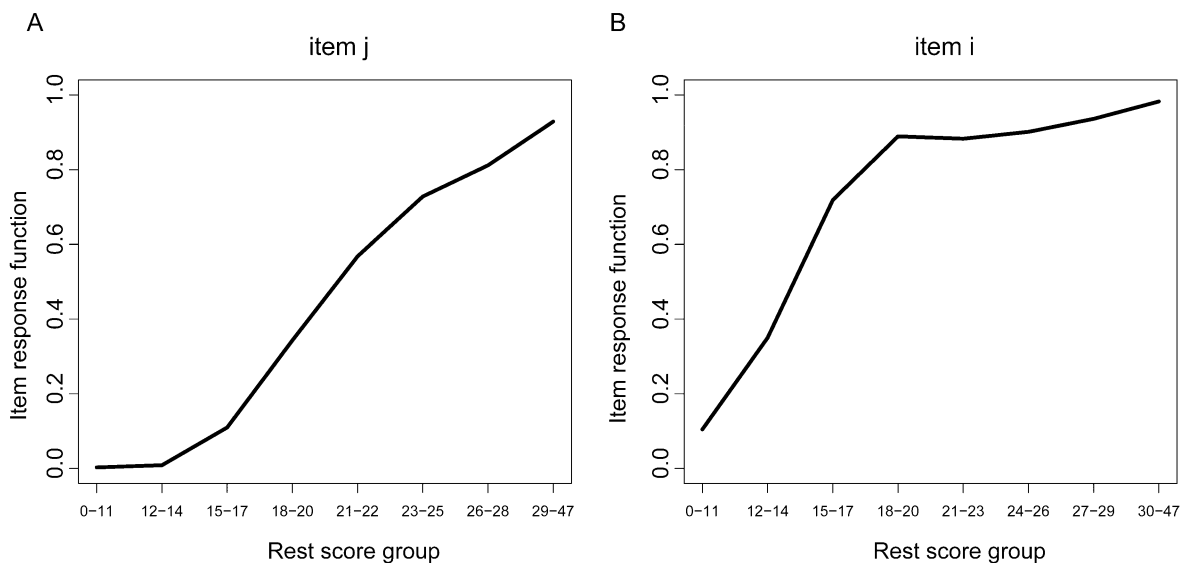
Unidimensionality: Unidimensionality means that a set of items in the scale or test measures only one latent trait (Straat et al., 2013; Sijtsma & Van der Ark, 2017). There are two methods to assess dimensionality. The first method is called the Automated Item Selection Procedure (AISP), which selects the highest H_{ij} item pairs to ensure that they are higher than the minimal lower bound (c) determined by the user (Sijtsma & Van der Ark, 2017). In the next step, a third item having a positive correlation with the selected items and also having a H_i value greater than both zero and c values to produce the highest H coefficient is selected. This process continues until certain conditions are met. If there are any unselected items, AISP follows the same process for another dimension. After creating the dimensions, if there are still unselected items, these items are marked as "non-scaling items," which cannot distinguish high and low ability

individuals and are excluded from the test or scale. The items with low discrimination do not contribute to individual ranking. Another method is called the Genetic Algorithm (GA), which defines random partitioning and evaluates each partitioning according to the specified conditions (crit statistic). This cycle repeats for all partitioning, and the best partitioning is reported when appropriate conditions are met.

Local Independence: Local independence is defined as the responses to one item that does not affect other responses when the latent variable is controlled (Nunnally, 1978; Wind, 2017; Sijtsma & Van der Ark, 2017). The conditional association procedure (CAP), proposed by Straat et al. (2016), is used to assess the local independence. CAP uses W_1 and W_3 indices to determine if the item pairs violate the local independence assumption. Straat et al. (2016) defined W indices to identify locally independent item sets that each index flags suspected item by calculating particular conditional covariances.

Monotonicity: It is also known as the monotonicity of IRFs. As the ability level increases (θ), the probability for a correct response to the item ($P(X_i = 1)$) does not decrease (Wind, 2017). Monotonicity can be shown graphically, as in Figure 1. There was no decrease in probability as the ability level increased in *item j*; whereas in *item i*, when the ability level increased, the probability decreased. Therefore, while *item j* ensured the monotonicity assumption, *item i* did not meet the monotonicity assumption. Besides graphical representation, rest scores and statistical hypothesis tests are used to evaluate monotonicity (Wind, 2017).

Figure 1. Monotonicity plots.

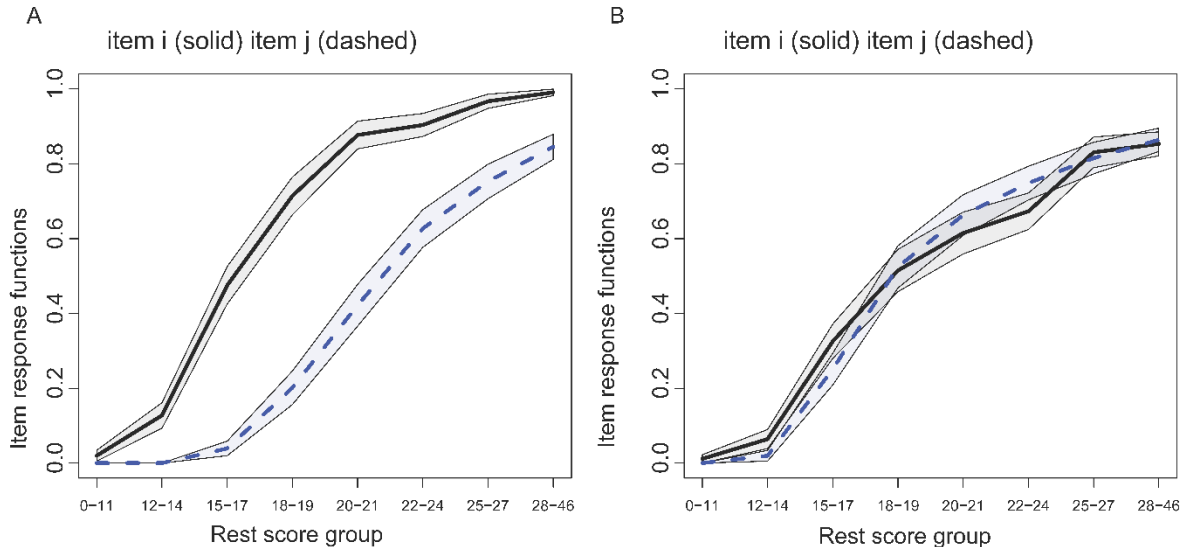


Invariant Item Ordering (IIO): IIO is defined as the IRFs that do not intersect for the specified item set (Sijtsma et al., 2011; Sijtsma & Van der Ark, 2017). This definition explains that the IIO assumption is satisfied, and the items are ordered from easy to difficult hierarchically. IIO can be shown graphically, as in Figure 2. Panel A, IRFs for the two items do not intersect with each other, so the IIO assumption is satisfied. Panel B illustrates intersecting IRFs that violate the IIO assumption.

Several methods evaluate IIO assumptions, including Restscore, P-matrix, and Item Splitting (Sijtsma & Molenaar, 2002; Wind, 2016, 2017). These methods evaluate rest scores and probability for a correct response through graphics. Ligtoet et al. (2010) stated inconsistencies regarding the assumption of nonintersecting IRFs on polytomous data and encouraged researchers to use the manifest IIO (MIIO) method. Sijtsma & Van der Ark (2017) stated the advantages of using the MIIO method over previous methods. In the MIIO method, the

backward item selection (BIS) procedure removes the items which violate the IIO assumption. BIS is an iterative procedure and reestimates H_i scalability coefficients after the items, causing violations from the test. If there are still items in violation, BIS keeps this process continuing until there are no violations.

Figure 2. *Intersecting and Nonintersecting IRFs.*



Ligtvoet et al. (2010) stated the advantages of using IIO in intelligence tests. Intelligence test items are administered in ascending order of item difficulty (Kaufman & Kaufman, 2004; Zhu et al., 2005). There are multiple reasons why intelligence test items are administered in a way from easy to difficult. The first reason for this practice is that respondents will succeed in the first items. Therefore, items will not negatively affect their motivation to proceed with later items to gain confidence and not feel stressed. The second reason for this practice is that since the intelligence tests are applied to various age groups, individuals in the upper age group do not get bored with questions far below their abilities, and item ordering practice shortens the testing time in terms of the usefulness of the test. Therefore, individuals in the upper age group do not take some starting items and start with specific items that better fit their age group and ability. It is assumed that the upper age group will answer the easy items correctly at the beginning. In such a practice pattern, the general assumption relies on that item difficulty orders are equal across each age group. However, since the item parameters cannot be estimated as sample independent in CTT models, the assumption that the "item difficulties are invariant" cannot be tested with distinct ability levels. However, the IIO estimations are sample independent, which provides the opportunity to test the assumption item difficulty invariance across all distinct ability levels.

Many studies apply MSA analyses to psychological scales in the literature, like personality and psychopathology scales (Chernyshenko et al., 2001; Meijer & Baneke, 2004). However, only a few studies have focused on nonparametric methods in intelligence tests (Abdelhamid et al., 2019). There is no discussion of the KBIT-2 subtest properties adapted for the original and Turkish forms. It is essential to test the psychometric properties of items using modern psychometric methods, like MSA, to check whether item orders in each subtest are consistent in the original form and the adapted one. As the item parameters, such as difficulty and discrimination, are sample dependent, person parameters are also dependent on the specific selection of items in the psychological tests. MSA can estimate the psychometric properties of the items independently from the sample, which provides practitioners to create adapted forms of the test using sample independent item parameters.

The main aim of the current study is therefore to examine both the advantages of using MSA in intelligence tests and the hierarchical order of the items in the KBIT-2: Turkish form by estimating the parameters of each of the three subtests by testing the dimensionality of the KBIT-2 subtests by using the IIO assumptions.

2. METHOD

2.1. Participants

2850 people participated in the study, including children, adolescents, and adults. Participants' ages varied from 48 months (4 years 0 months) to 539 months (44 years 11 months). The average age of the participants is $M = 178.72$; the standard deviation is $SD = 103.47$. The Turkish form of the KBIT-2 test was applied to all individuals who participated in the study. All participants were native speakers of the Turkish language. Each test was applied and evaluated by the psychologists, who had KBIT-2 training.

2.2. Instruments

KBIT-2 Turkish form was first adapted in 2014 (Atalay, 2007; Öktem, 2016; Savaşan, 2006, Uluç et al., 2015) and comprised three subtests called MT, VK, and RD that produce Verbal, Nonverbal, and IQ composite scores ($M=100$; $SD= 15$) like the original form developed by Kaufman & Kaufman (2004). VK (60 items) and RD (48 items) subtests comprise the Verbal Standard Score, while MT (46 items) makes up the Nonverbal Standard Score (Kaufman & Kaufman, 2004). All subtests are scored dichotomously.

2.3. Data Analysis

Data analysis was performed with the R package "Mokken version: 3.0.3" (Van der Ark, 2012) in order to investigate the MHM and DMM assumptions. First, the total scalability coefficient (H) was evaluated with the conditions in which $0.30 \leq H < 0.40$ indicates a weak scale, $0.40 \leq H < 0.50$ indicates a medium scale, and $H \geq 0.50$ indicates a strong scale (Wind, 2016). $H < .30$, H indicates that the item does not fit the Mokken scale, which is also called an unscalable item. Also, item scalability coefficient and item-pair scalability coefficient were evaluated with the condition $H_i \geq 0.30$ and $H_{ij} \geq 0$, which indicate items should be selected for Mokken scaling; otherwise, items should be reviewed or excluded from the test, and item pairs should not be negative, respectively.

For unidimensionality assumption AISP, c is set to 0.30, 0.40, and 0.55. Per Element Accuracy (PEA), proposed by Hogarty et al. (2005), is used to evaluate how accurately items were allocated to scales or dimensions with following conditions: $0.80 < PEA \leq 0.90$ mediocre; $0.90 < PEA \leq 0.95$ adequate; $0.95 < PEA \leq 0.99$ good, and $PEA > .99$ excellent.

For the local independence assumption, the W_1 and W_3 indices show that high values indicate item pair positively and negatively locally dependent, respectively (Sijtsma & Van der Ark, 2017). To examine each subtest's monotonicity assumption, IRF graphs, based on nonparametric regression between item scores and total scores, are obtained (Junker & Sijtsma, 2001; Sijtsma & Molenaar, 2002) and significant violations are reported.

IIO assumption is tested with BIS procedures, an iterative method, to detect items that cause violations. Wind (2016) stated that the *Crit* statistic, an impact size measure for item violation (Molenaar & Sijtsma, 2000), is also used in some studies to identify which items violate IIO assumptions. Items indicate no serious violations if $Crit < 40$; minor violation if $40 \leq Crit \leq 80$, and significant violations if $Crit > 80$. However, Crişan et al. (2019) suggested that *Crit* has failed to discriminate fitting and misfitting items for IIO. BIS procedure overcomes this problem using the iterative procedure by removing an item from the scale even though the *Crit* statistic is lower than 40. In this study, items that violate IIO assumptions were determined

using the BIS procedure. Furthermore, H_T coefficients are reported to provide information about the accuracy of IIO based on the following criteria: Item orderings show high accuracy if $H_T \geq 0.50$; medium accuracy if $0.40 \leq H_T < 0.50$; low accuracy if $0.30 \leq H_T < 0.40$, and item orderings are inaccurate if $H_T < 0.30$.

Finally, to assess the reliability of the scale, lambda-2 statistics (Sijtsma, 2009), Molenaar-Sijtsma coefficient (Sijtsma & Molenaar, 2002), and latent class reliability coefficient (LCRC) are reported (Van der Ark et al., 2011).

3. RESULTS

Table 1 provides an overview of the descriptive summaries of the KBIT-2: Turkish form administration. Table 1 shows the minimum and maximum mean score values similar for all subtests except for RD, which has the most challenging item mean score of 0.04. The skewness and kurtosis values are also included in Table 1 in order to interpret the normality assumption, which can be considered acceptable to prove normal univariate distribution. Three reliability coefficients (alpha, split-half, and test-retest) were also estimated and reported. The reliability coefficients of all three subtests were estimated above .90, which shows that the test reliability is high. This finding implies that KBIT-2: Turkish form shows high reliability on each subtest based on CTT.

Table 1. Descriptive Statistics and Reliability Estimates for the Kaufman Brief Intelligence Test-2: Turkish Form Subtests.

	N	Mean	SD	S	K	Reliability		
						Alpha	Split-Half	Test-Retest
Matrices	48	0.031-0.992	0.09-0.50	-0.36	-0.54	0.95	0.96	0.93
Verbal Knowledge	60	0.030-0.995	0.07-0.49	-0.53	-0.45	0.96	0.97	0.94
Riddles	46	0.004-0.993	0.06-0.50	0.20	-0.32	0.93	0.95	0.91

N= number of items; *SD*=standard deviation; *S*=skewness; *K*=kurtosis; *Alpha*= Cronbach's alpha coefficient

3.2. MSA Results

This section summarizes the results from KBIT-2: Turkish form data in which the scalability coefficients from the subtests were estimated according to the MHM and the DMM assumptions. The estimated coefficients were then compared based on the evaluation criteria mentioned earlier to address the research questions in this present study. MHM and DMM results are discussed, respectively:

Table 2 presents the MHM outputs for all three subtests, along with the number of total violations and PEA estimates. The total scalability (H) coefficient was achieved for the criteria $H > 0.5$, which indicates all three subtests formed a strong Mokken scale. Also, item scalability coefficients for each subtest succeeded in satisfying $H_i > 0.30$ criterion, indicating that all items fit for Mokken scaling, and no item was excluded from the test. For MT, VK, and RD subtests, item scalability coefficients ranged between 0.50 to 0.88. Finally, item-pair scalability coefficients (H_{ij}) were all above the minimum value zero, while the lowest H_{ij} value was estimated as 0.59 for the VK subtest.

Table 2 also shows the effect of varying minimal lower bound values (0.30, 0.40, and 0.55) and PEA values for AISP on the assessment of dimensionality. Results indicate that PEA values estimated from various lower bounds provide consistent information about the test dimensionality. For MT and RD (with $c = 0.30$ and 0.40), PEA is excellent; and for the rest, PEA is good for allocating items into the dimensions. Considering the PEA measures for

various conditions, the items can form a single scale in each subtest, which is interpreted as all three subtests that are unidimensional.

For each subtest, the conditional association procedure indices W_1 and W_3 did not flag any items, which indicates all item pairs are locally independent. Thus, it was concluded that all three subtests ensured the local independence assumption.

The probability of a correct response to the question was calculated by creating rest score groups according to their ability levels with the help of IRF graphics to test the monotonicity assumption in MHM analyses. When the analysis results in Table 2 are examined, it can be seen that only the 27th item in the Verbal Knowledge subtest created one violation, however it was not marked as significant. In this respect, it can be said that the monotonicity assumption is ensured for all three subtests. Furthermore, IRF outputs provided strong evidence of monotonicity for all items in all three subtests.

In summary, MHM results indicate that the monotonicity, local independence, and unidimensionality assumptions held for each of the KBIT-2 subtests and PEA values provided consistent estimates on dimensionality assessment.

Table 2. Summary of Scalability Coefficients and Per Element Accuracy Values for the Kaufman Brief Intelligence Test-2: Turkish Form Subtests.

	N	H	H_i	H_{ij}	# $\sum vi$	# $\sum sigvi$	PEA		
							0.3	0.4	0.55
Matrices	48	0.74	0.61-0.88	0.67-0.89	0	0	1.00	1.00	1.00
Verbal Knowledge	60	0.69	0.50-0.86	0.59-0.90	1	0	0.98	0.98	0.97
Riddles	46	0.64	0.54-0.83	0.67-0.90	0	0	1.00	1.00	0.98

$\sum vi$ = total number of violations; # $\sum sigvi$ = total number of significant violations; PEA = per element accuracy

The KBIT-2 data were tested with the MIO method to identify the items that violated the invariant ordering for each subtest. The BIS procedure, which eliminates the lowest scalability item, was used to remove items violating the IIO. Subsequently, the HT coefficient was estimated for selected items in each subtest to check the accuracy of the IIO. The IIO assumption results were solely summarized for the removed items determined by the BIS procedure in Table 3, which shows the number of significant violations for the IIO and crit statistics along with the mean score, item scalability coefficient, and the number of significant violations for monotonicity.

As shown in Table 3, although Molenaar & Sijtsma (2000) suggest that items for which the *Crit* statistic was estimated below 40 can be considered as not seriously violating items and can be safely included in any Mokken scale, the BIS procedure excluded the items regardless of *Crit* statistic. The BIS procedure detected seven items (9, 15, 18, 23, 28, 30, and 36) for the MT, seventeen items (19, 21, 22, 23, 24, 25, 27, 28, 32, 33, 34, 37, 42, 43, 44, 46, and 50) for the VK and six items (14, 15, 16, 19, 21, and 22) for the RD that violated the invariant ordering. Figure 3 demonstrates a graphical illustration of items that violated the IIO and nonintersecting IRF assumptions. As shown in Figure 3, Items 9 and 15 for MT, Items 34 and 37 for VK, and Items 21 and 22 for RD were graphically shown as intersecting IRFs that violated the IIO assumption.

Table 3. Summary of Invariant Item Ordering Results for the Kaufman Brief Intelligence Test-2: Turkish Form Subtests.

Item#	M	SD	Scalability		Monotonicity	IIO		
			H_i	se	#sigvi	MIO #sigvi	crit	
Matrices								
9	0.90	0.30	0.71	0.02	0	2	19	
15	0.90	0.30	0.62	0.02	0	5	43	
18	0.78	0.41	0.73	0.01	0	6	59	
23	0.78	0.42	0.88	0.01	0	10	63	
28	0.49	0.50	0.67	0.01	0	3	33	
30	0.42	0.49	0.67	0.01	0	4	39	
36	0.21	0.40	0.70	0.01	0	2	25	
Verbal Knowledge								
19	0.82	0.39	0.76	0.01	0	5	32	
21	0.80	0.40	0.70	0.02	0	10	53	
22	0.80	0.40	0.77	0.01	0	4	26	
23	0.73	0.44	0.70	0.01	0	14	90	
24	0.74	0.44	0.72	0.01	0	7	74	
25	0.61	0.49	0.50	0.02	0	30	189	
27	0.72	0.45	0.65	0.02	0	13	93	
28	0.73	0.44	0.70	0.01	0	10	71	
32	0.60	0.49	0.68	0.01	0	6	49	
33	0.49	0.50	0.62	0.01	0	3	33	
34	0.56	0.50	0.64	0.01	0	11	79	
37	0.56	0.50	0.62	0.01	0	12	74	
42	0.37	0.48	0.62	0.01	0	2	28	
43	0.28	0.45	0.64	0.01	0	5	36	
44	0.26	0.44	0.56	0.01	0	9	72	
46	0.24	0.43	0.62	0.01	0	3	35	
50	0.16	0.36	0.58	0.01	0	2	24	
Riddles								
14	0.83	0.38	0.63	0.02	0	6	54	
15	0.71	0.45	0.72	0.02	0	1	15	
16	0.55	0.50	0.55	0.02	0	7	55	
19	0.49	0.50	0.60	0.01	0	2	34	
21	0.43	0.50	0.54	0.01	0	8	72	
22	0.45	0.50	0.67	0.01	0	6	58	

Item#= deleted item number; Mean= mean item score; #sigvi = number of significant violations; Crit= critical value for model violations

Figure 3. Example violations of the IIO assumption for Matrices (M), Verbal Knowledge (VK), and Riddles (R) Subtests.

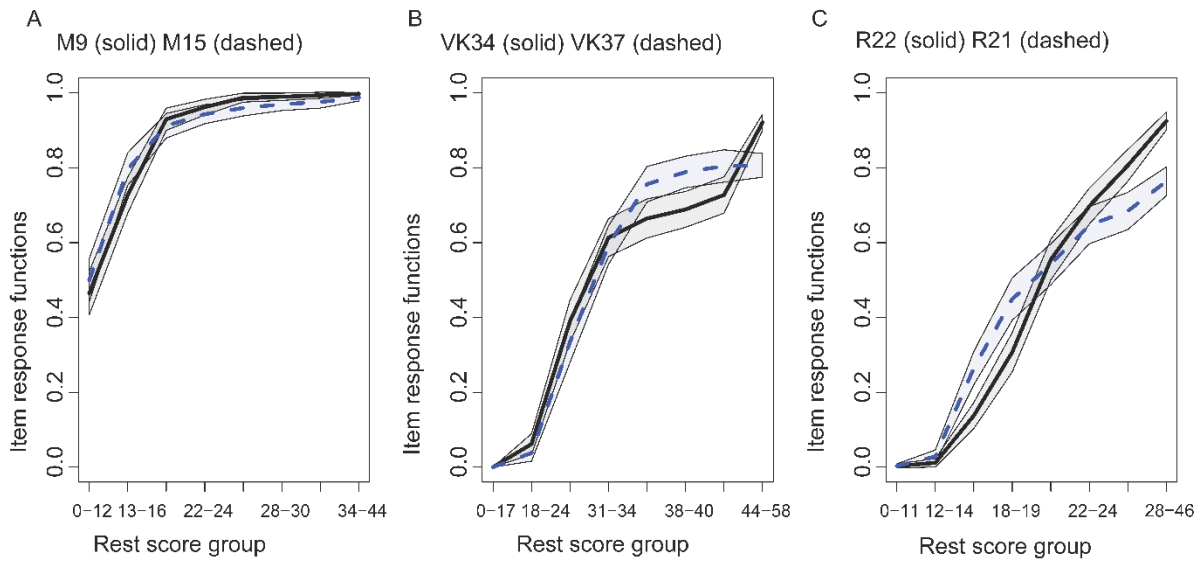


Table 4 summarizes the H_T statistics and reliability estimates for Mokken analysis. H_T values for the MT, the VK, and the RD subtests were found as .88, .91, and .91, respectively, which indicated sufficient item ordering accuracy for the subtests. Finally, Table 4 also provides reliability estimates that MS ranging from .95 to .97; λ_2 ranging from .94 to .96 and LCRC ranging from .96 to .97 that revealed high reliability for each subtest.

According to Wind (2017), items that violate the MHM and DMM assumptions should be removed from the data matrix. If possible, it is recommended to revise items accompanied by content experts and practitioners. After revising or removing items, it is recommended to readminister the updated test items before additional analyses are conducted. Even though updated test items were not readministered in this study, the total scalability coefficient for updated test items is also estimated and reported in Table 4, namely H_{ad} (after deleted). The main reason for reestimating the total scalability coefficient is to predict how the test might behave when the specified items are removed from the test. It is highly recommended to interpret the H coefficient differences after real data application.

Table 4. Summary of Double Monotonicity Model and Reliability Statistics for the Kaufman Brief Intelligence Test-2: Turkish Form Subtests.

	N	H_{ad}	H_T	Reliability		
				MS	λ_2	LCRC
Matrices	48	.75 (.01)	.88	.97	.96	.97
Verbal Knowledge	60	.74 (.05)	.91	.97	.96	.97
Riddles	46	.66 (.02)	.91	.95	.94	.96

H_{ad} = Total scalability coefficient after items deleted (the difference between the previous H coefficient); H_T = transpose H ; MS = Molenaar–Sijtsma coefficient; λ_2 = lambda-2 coefficient; LCRC = latent class reliability coefficient

4. DISCUSSION and CONCLUSION

This study aimed to demonstrate MSA's fundamental principle, including how MHM and DMM can be applied to intelligence tests that aim to rank individuals according to latent ability. It also investigates the psychometric properties of KBIT-2 subtests using modern theoretical approaches rather than CTT, making it possible to spot the differences in ordering items and persons between the KBIT-2: Turkish and the standard version. A detailed assessment of

dimensionality and Invariant Item Ordering (IIO) assumptions were also examined by the KBIT-2 subtests.

Overall, the KBIT-2 test showed robust psychometric specifications on monotonicity, scalability, and local independence. However, IIO results reported items with significant violations. Regarding the IIO results, results lead practitioners to use the KBIT-2 test cautiously.

The findings of the study suggested that MHM fit well to all items of the subtests without creating a significant violation. Item scalability coefficients provided sufficient estimates in which all values range between 0.50 to 0.88. Thus, it can be concluded that the sum score of correct responses for each subtest is a good indicator of the latent ability for ordering individuals. Thus, it can be stated that individuals with a higher level of intelligence would score higher for each subtest. Regarding the subscales, Matrices, Verbal Knowledge, and Riddles showed strong Mokken scalability that the H coefficient was estimated as above 0.74, 0.69, and 0.64, respectively. H coefficients provide support that sum scores for the KBIT-2 subtests are able to order persons based on their intelligence abilities.

AISP is used to evaluate unidimensionality assumptions. While determining the number of Mokken scales in the data, the AISP procedure was replicated separately for 0.30, 0.40, and 0.55 lower bounds. Correct partitioning ratios of the items were interpreted using PEA values that ranged from adequately to excellent for various lower bound conditions. Comparison of the PEA findings with various lower bounds confirms that the total score of each of the KBIT-2 subtests fits the unidimensionality assumption that total scores represent an individual's intelligence level for each subtest. As no significant differences were found in PEA estimations, scalability coefficients for 0.30 lower bound criteria were taken as reference.

W_1 and W_3 indexes flagged no item pairs likely to be positively or negatively local dependent. However, significant violations for the IIO appeared to be tempting to remove items for each three subtests. Abdelhamid et al. (2019) provided an IIO analysis and discussed the importance of testing invariant items in an adult intelligence test, called WAIS, using the BIS procedure. In detecting violating items, the results of the MIIO method for dichotomous items were reported. For KBIT-2 data, the BIS procedure detected seven items for the MT, seventeen for the VK, and six items for the RD that violated the invariant ordering.

In the literature, MSA is applied to evaluate the psychometric quality of tests in psychology, education, and health research (Meijer & Banneke, 2004; Meijer et al., 2011; Watson et al., 2008; Wind, 2017). The MHM and the DMM results demonstrated how an item order affects an intelligence test results even if there was no problem detected in classical analysis. The findings also indicated that IIO provided consistent predictions about item order and item/person order for various ability levels, mainly if the sample ranges from young to adults for KBIT-2 test. It is therefore likely that an item that violates should be removed from the test to better estimate intelligence levels. This finding, while preliminary, suggests that it is essential for the intelligence test that item orders must show the same sequence for each ability level to create accurate norms. As Meijer and Egberink (2012) stated, if the items are not ordered the same way for all ability levels, scores may differ when evaluating the expected symptoms. These findings are in line with the study of Ligtoet et al. (2010), which states that test constructors assume items to be easy for each respondent, but it is not easy to prove this assumption empirically.

4.1. Limitations and Recommendations

The generalizability of these results is subject to certain limitations. The most important limitation lies in the fact that even though some of the psychometric properties of an intelligence test were estimated satisfactory, IIO assumption was not supported. KBIT-2 test was originally conceptualized as an intelligence test that test takers respond to items in an increasing difficulty

order. Empirical support for this assumption is not provided due to the items that violate invariant ordering. As Ligtoet et al. (2010) stated empirical evidence should be tested to make interpretations. Another limitation of this study is that the GA procedure was not applied for dimensionality assumption due to the large sample size and the number of items. Abdelhamid et al. (2019) investigated the differential impact of GA estimation on adult intelligence scales and provided satisfactory results. This study only used the AISP method to investigate unidimensionality (Sijtsma & Van der Ark, 2017).

An additional uncontrolled factor is the possibility that the age range of the sample which might cause peculiarities in IIO assumptions. Current findings must be considered for each age norm with regard to a representative sample size.

For the future adaptations of KBIT-2, MHM and DMM analyses are recommended to examine the psychometric properties of the test. In addition to KBIT data, it is also recommended that MSA can be used for various intelligence tests such as Wechsler Individual Achievement Test (WAIT), Woodcock-Johnson (WJ), and Wechsler Intelligence Scale for Children (WISC). Moreover, MSA can also be used for different psychological tests, consisting of a starting and discontinue rule, such as Vineland Adaptive Behaviour Scales.

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:** Trakya University, 22.07.2020 - 05/10.

Authorship Contribution Statement

Eren Halil Ozberk: Designing the model, the computational framework, analyzing the data. **Elif Bengi Unsal Ozberk:** Literature review, the computational framework, theoretical framework development, interpreting the results. **Sait Uluc:** Data collection, theoretical framework development, interpreting the results. **Ferhunde Oktem:** Data collection, theoretical framework development, interpreting the results.

ORCID

Eren Halil Ozberk  <https://orcid.org/0000-0003-2136-3081>

Elif Bengi Unsal Ozberk  <https://orcid.org/0000-0003-3605-3983>

Sait Uluc  <https://orcid.org/0000-0002-7048-8545>

Ferhunde Oktem  <https://orcid.org/0000-0001-6971-6822>

5. REFERENCES

- Abdelhamid, G. S. M., Gómez-Benito, J., Abdeltawwab, A. T. M., Abu Bakr, M. H. S., & Kazem, A. M. (2020). A Demonstration of Mokken Scale Analysis Methods Applied to Cognitive Test Validation Using the Egyptian WAIS-IV. *Journal of Psychoeducational Assessment*, 38(4), 493–506. <https://doi.org/10.1177/0734282919862144>
- Atalay, Z. Ö. (2007). *Kaufman brief intelligence test the studies of validity, reliability, and pre norm on children who are 13-14 years of age* [Unpublished master's thesis], İstanbul University, İstanbul.
- Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., Williams, B. A. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36, 523–562. https://doi.org/10.1207/S15327906MBR3604_03
- Cole, J. C., & Randall, M. K. (2003). Comparing the cognitive ability models of Spearman, Horn and Cattell, and Carroll. *Journal of Psychoeducational Assessment*, 21, 160-179. <https://doi.org/10.1177/073428290302100204>

- Crişan, D. R., Tendeiro, J., & Meijer, R. (2019). The Crit Value as an Effect Size Measure for Violations of Model Assumptions in Mokken Scale Analysis for Binary Data. <https://doi.org/10.31234/osf.io/8ydmr>
- Embretson, S. E., Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage.
- Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communalities, and overdetermination. *Educational and Psychological Measurement, 65*, 202–226. <https://psycnet.apa.org/doi/10.1177/0013164404267287>
- Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65-81. <https://doi.org/10.1177%2F01466216000241004>
- Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211–220. <https://doi.org/10.1177%2F01466210122032028>
- Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test* (2nd ed.). American Guidance Service.
- Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*, 578–595. <https://doi.org/10.1177%2F0013164409355697>
- Lord, F. M., & Novick, R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354–368. <https://doi.org/10.1037/1082-989x.9.3.354>
- Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*, 283–298. <https://doi.org/10.1177/014662169001400306>
- Meijer, R. R., de Vries, R. M., & van Bruggen, V. (2011). An evaluation of the Brief Symptom Inventory-18 using item response theory: Which items are most strongly related to psychological distress?. *Psychological Assessment, 23*, 193-202. <https://doi.org/10.1037/a0021292>
- Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). Springer.
- Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows. A program for Mokken scale analysis for polytomous items*. Groningen: iecProGAMMA.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Mouton.
- Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430. <https://doi.org/10.1177%2F014662168200600404>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- Öktem, F., (2016). Brief Intelligence Tests and Kaufman Brief Intelligence Test (KBIT-2). *Türkiye Klinikleri J Psychol-Special Topics, 1*(1), 10-6.
- Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items?. *Psychological Methods, 8*(2), 164-184. <https://doi.org/10.1037/1082-989x.8.2.164>

- Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance*, 14, 187–207. https://doi.org/10.1207/S15327043HUP1402_04
- Savaşan, G. (2006). *Kaufman Brief Intelligence Test the studies of validity, reliability and pre norm (age 9-10)* [Unpublished master's thesis], İstanbul University, İstanbul.
- Sijtsma, K. (2009). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing*, 9, 167-194. <https://doi.org/10.1080/15305050903106883>
- Sijtsma, K., Debets, P., & Molenaar, I. W. (1990). Mokken scale analysis for polychotomous items: Theory, a computer program and an empirical application. *Quality and Quantity*, 24, 173-188. <https://doi.org/10.1007/BF00209550>
- Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, 16, 149–157. <https://doi.org/10.1177/014662169201600204>
- Sijtsma, K., Meijer, R. R., & Van der Ark, L. A. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, 50(1), 31-37. <https://psycnet.apa.org/doi/10.1016/j.paid.2010.08.016>
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. SAGE.
- Sijtsma, K., & Molenaar, I. W. (2016). Mokken models. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume One: Models* (pp. 303–321). Chapman & Hall/CRC.
- Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, 70(1), 137–158. <https://doi.org/10.1111/bmsp.12078>
- Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, 66(2), 341–349. <https://psycnet.apa.org/doi/10.1037/0022-3514.66.2.341>
- Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, 30, 72–99. <https://doi:10.1007/s00357-013-9122-y>
- Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement*, 74(5), 809–822. <https://doi:10.1177/0013164414529793>
- Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 12(4), 117–123. <https://doi.org/10.1027/1614-2241/a000115>
- Uluç, S., Öktem, F., Korkmaz, B. (2015). *Brief screening tests: Kaufman Brief Intelligence Test-2 standardization for the Turkish version*. VII. Işık Savaşır Clinical Psychology Symposium, Ankara.
- Van der Ark LA (2012). "New Developments in Mokken Scale Analysis in R." *Journal of Statistical Software*, 48(5), 1–27. <https://doi.org/10.18637/jss.v048.i05>
- Van der Ark, L. A., Van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test score reliability. *Applied Psychological Measurement*, 35(5), 380-392. <https://doi.org/10.1177%2F0146621610392911>
- Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods*, 5(1), 125–146. <https://doi.org/10.1037/1082-989X.5.1.125>

- Watson, R., Deary, L., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine*, 38(4), 575-579. <https://doi.org/10.1017/S003329170800281X>
- Wind, S. (2016). Examining the psychometric quality of multiple-choice assessment items using Mokken scale analysis. *Journal of Applied Measurement*, 17(2), 142–165.
- Wind, S. (2017). An instructional module on Mokken scale analysis. *Educational Measurement: Issues and Practice*, 36(2), 50–66. <https://doi.org/10.1111/emip.12153>
- Zhu, J., Weiss, L. G., Prifitera, A., & Coalson, D. (2004). The Wechsler Intelligence Scales for Children and Adults. In G. Goldstein, S. R. Beers, & M. Hersen (Eds.), *Comprehensive handbook of psychological assessment, Vol. 1. Intellectual and neuropsychological assessment* (p. 51–75). John Wiley & Sons, Inc.