



Volume 8

Issue 4

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 4

2021

Dr. İzzet KARA

Publisher

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

Design & Graphic: 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

Editors

Dr. Eren Can AYBEK, Pamukkale University, Turkey

Editorial Board

Dr. Beyza AKSU DUNYA, Bartın University, Turkey

Dr. Stanislav AVSEC, University of Ljubljana, Slovenia

Dr. Murat BALKIS, Pamukkale University, Turkey

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

Dr. Okan BULUT, University of Alberta, Canada

Dr. Javier Fombona CADAVIECO, University of Oviedo, Spain

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

Dr. R. Nukhet CIKRIKCI, İstanbul Aydın University, Turkey

Dr. Safiye Bilican DEMİR, Kocaeli University, Turkey

Dr. Nuri DOGAN, Hacettepe University, Turkey

Dr. Selahattin GELBAL, Hacettepe 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 University, 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. Halil İbrahim SARI, Kilis 7 Aralık University, Turkey

Dr. Ragıp TERZİ, Harran University, Turkey

Dr. Ozen YILDIRIM, Pamukkale University, Turkey

English Language Editors

Dr. R. Sahin ARSLAN, Pamukkale University, Turkey

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

[Equality of admission tests using kernel equating under the non-equivalent groups with covariates design](#)

Page: 729-743, [PDF](#)

Ozge ALTINTAS, Gabriel WALLIN

[PPSE P121 and P10 calculation method and related issues](#)

Page: 744-763, [PDF](#)

Umit CELEN

[The Validity and Reliability of the Turkish Version of the Attitudes to Fertility and Childbearing Scale \(AFCS\)](#)

Page: 764-774, [PDF](#)

Sinem GORAL, Sevgi OZKAN, Pinar SERCEKUS, Erkan ALATAS

[Development and Evaluation of a Turkish Language Version of the Relational Health Indices](#)

Page: 775-784, [PDF](#)

Nesime CAN, Abdulkadir HAKTANIR, A. Stephen LENZ, Joshua C. WATSON

[Performance and Differences in Grading Practices Among Undergraduates at Business Schools](#)

Page: 785-800, [PDF](#)

Leiv OPSTAD

[The Effect of Formative Assessment Practices on Student Learning: A Meta-Analysis Study](#)

Page: 801-817, [PDF](#)

Pınar KARAMAN

[Investigation of a Middle School Preservice Teacher's Knowledge of Content and Students](#)

Page: 818-841, [PDF](#)

Ebru ERSARI

[Assessing Measurement Invariance of Achievement Emotions Questionnaire for Teachers in Prospective Teacher Sample](#)

Page: 842-854, [PDF](#)

Sevilay KILMEN

[Factor structure and measurement invariance of the TIMSS 2015 mathematics attitude questionnaire: Exploratory structural equation modelling approach](#)

Page: 855-871, [PDF](#)

Seyma UYAR

[Validation of a new State Test Anxiety Scale \(STAS\)](#)

Page: 872-887, [PDF](#)

Alper SAHIN

[The Study of Developing and Validating the Union Bias Scale](#)

Page: 888-913, [PDF](#)

Ender KAZAK

[Adaptation of the Adlerian Personality Priority Assessment into Turkish](#)

Page: 914-927, [PDF](#)

Abdi GUNGOR, Dalena DILLMAN TAYLOR

[Classification of Scale Items with Exploratory Graph Analysis and Machine Learning Methods](#)

Page: 928-947, [PDF](#)

Ilhan KOYUNCU, Abdullah Faruk KILIC

[Examining the Discrimination of Binary Scored Test Items with ROC Analysis](#)

Page: 948-958, [PDF](#)

Sait CUM

[A Comparison of Latent Class Analysis and the Mixture Rasch Model Using 8th Grade Mathematics Data in the Fourth International Mathematics and Science Study \(TIMSS-2011\)](#)

Page: 959-974, [PDF](#)

Turker TOKER, Kathy GREEN

[The Opinions of Field Experts on Online Test Applications and Test Security During the COVID-19 Pandemic](#)

Page: 975-990, [PDF](#)

Hakan KILINC, Muhammet Recep OKUR, Ilker USTA

[Fuzzy logic expert system for evaluating the activity of university teachers](#)

Page: 991-1008, [PDF](#)

V. Florin POPESCU, M. Sorin PISTOL

Equality of admission tests using kernel equating under the non-equivalent groups with covariates design

Ozge Altintas^{1,*}, Gabriel Wallin²

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

²Université Côte d'Azur, Inria, CNRS, Laboratoire J. A. Dieudonné, team Maasai, Sophia-Antipolis, France

ARTICLE HISTORY

Received: Mar. 03, 2021

Revised: July 03, 2021

Accepted: July 30, 2021

Keywords:

Kernel equating,
Non-equivalent groups
design,
NEC design,
Background variables,
Admission tests.

Abstract: Educational assessment tests are designed to measure the same psychological constructs over extended periods. This feature is important considering that test results are often used for admittance to university programs. To ensure fair assessments, especially for those whose results weigh heavily in selection decisions, it is necessary to collect evidence demonstrating that the assessments are not biased and to confirm that the scores obtained from different test forms have statistical equality. Therefore, test equating has important functions as it prevents bias generated by differences in the difficulty levels of different test forms, allows the scores obtained from different test forms to be reported on the same scale, and ensures that the reported scores communicate the same meaning. In this study, these important functions were evaluated using real college admission test data from different test administrations. The kernel equating method under the non-equivalent groups with covariates design was applied to determine whether the scores that were obtained from different periods and measured the same psychological constructs were statistically equivalent. The non-equivalent groups with covariates design was specifically used because the test groups of the admission test are non-equivalent and there are no anchor items. Results from the analyses showed that the test forms had different score distributions and that the relationship was non-linear. Thus, the equating procedure was adjusted to eliminate these differences and thereby allowing the tests to be used interchangeably.

1. INTRODUCTION

Throughout much of human history, tests have been figured prominently as measurement tools in all areas of life. They are used for many purposes including monitoring the development process of individuals, determining the level of readiness for school, identifying the learning achievements of students, issuing diplomas or certificates, and deciding on proper treatment methods for psychological problems. This widespread use reveals the importance of tests in human life. Cronbach (1990) states that tests provide evidence for understanding individuals and gaining knowledge about human behavior. Anastasi (1988) defines psychological tests as

*CONTACT: Özge ALTINTAŞ ✉ oaltintas@ankara.edu.tr 📍 Ankara University, Faculty of Education, Department of Educational Sciences, Educational Measurement and Evaluation, Ankara, Turkey

an objective, standardized measure of a psychological variable such as intelligence, ability, aptitude, interest, attitude, and behavior.

One of the most common uses of tests is in schools. In pre-school, primary, and secondary school, basic life skills are taught, while in high school, the focus shifts to developing basic mental skills and orienting students to a future profession. Higher education programs, on the other hand, aim to equip individuals with the requisite set of skills and competencies associated with their profession of choice and at the same time, to enrich their intellectual, factual, and scientific knowledge. With the growing competitiveness in securing admittance to prestigious universities, it is common for students to take multiple admission tests to improve their chances of being accepted (Altıntaş & Kutlu, 2020).

Different forms of tests are used for entrance exams to universities and other educational institutions, for personnel selection, and for exams administered in different years or periods to ensure the security and integrity of the assessment process. In some cases, parallel versions of a test are used to allow the students more than one chance to be evaluated in certain periods. However, the use of different test forms on different dates raises concerns over whether the difficulty level of these forms differs (Kolen & Brennan, 2014). If no adjustment for difficulty differences is made, it is not possible to fairly compare test-takers who have been issued different test forms.

Similar questions asked in different formats, such as graphically, verbally, or symbolically, can be used multiple times in exams that measure the same construct, which is usually the case in exams administered for selection purposes. Although the use of parallel test forms that measure the same characteristics seems to be a reasonable way to ensure fairness (Kan, 2010) and exam security, the issue regarding the comparability of the scores obtained from these different tests is a source of concern.

The construction of parallel forms depends in equal measure on expert judgment and empirical data. The judgment comes into play in determining whether the items on these parallel forms measure the same function, a decision that sometimes is quite difficult to make (Levine, 1955, p.4). Proving that two tests, which are supposed to measure the same construct, are psychometrically equal (equivalent) to one another is essential in terms of preventing possible sources of bias.

Lord (1950) describes “comparability” in the sense that scores from two different tests each represent an equivalent amount of training or promises an equivalent degree of future success in a particular activity or other fields of knowledge. The comparability of scores obtained on different forms of a test depends on the accurate equating of these scores (Holland & Thayer, 1985, p.109). In selection processes, the comparability of the scores acts as an important indicator that the selection procedures are fair. As emphasized by Dorans and Holland (2000, p.281), the comparability of measurements made by different methods and researchers under different conditions is an essential component of the scientific method. Psychological and educational measurements are no exception to this rule.

Equating is a statistical process that is applied to confirm that scores on different test forms are comparable. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content (Kolen & Brennan, 2014). Equality/equivalence of test scores, or test equating was defined by Angoff (1971, 1982) as the conversion of the unit system of one form to the unit system of another form. Test equating is a numerical arrangement made to ensure that scores obtained from forms at different difficulty levels can be used interchangeably (Braun & Holland, 1982). Similarly, Felan (2002) stated that test equating is often used in situations where multiple forms of a test exist, where exams consisting of different forms are compared to each other, or when researchers want to overcome problems of practice effects.

A study by von Davier et al. (2004) argued that while there is no unified perspective on test equating, all equating approaches feature at least the five following “requirements”: (1) equal construct, (2) equal reliability, (3) symmetry, (4) equity, and (5) population invariance. Here, equality is expressed in terms of the persons taking the exam, equal reliability and population invariance are related to the size of the population, symmetry is a mathematical property, and equal structure is related to the nature and use of the tests.

Since standardized tests are typically given at different times and with different test forms, the test that is administered by the test-takers must not unfairly affect the results capable of being attained (Andersson et al., 2013b). In effect, this means that the comparability of the scores obtained from a test and the interchangeability of the scores obtained in different years are important, insofar as they allow test-takers to compare their current scores with past and future scores. As is the case throughout the world, some tests are used in Turkey regularly (every year, twice a year, etc.) for the same purposes (selection, placement, etc.). The institutions responsible for developing and applying these tests accept that the different forms of the tests make equivalent measurements to realize the same purpose. The Ankara University Examination for Foreign Students (AYOS), which has been applied since 2011 for admission of international students to Turkish universities, especially Ankara University, is considered equivalent to each other. Research on the AYOS Basic Learning Skills Test (BLST) scores, such as measurement invariance and differential item functioning studies (Altıntaş & Kutlu, 2019, 2020), has revealed that AYOS has equivalence in terms of individuals in different groups (i.e., country and gender) who took the test the same year.

Although the psychological constructs measured by the test do not change, AYOS tests are developed for the same purpose and applied once every year. Hence, the groups taking the test are different (Kutlu & Bal, 2011). The gold standard is to use common items, also known as anchor items, to adjust for this kind of imbalance in ability between the test groups. However, AYOS does not include any common items. This study, therefore, follows the suggestion by Wiberg and Bränberg (2015) and uses background information about the test-takers. The idea behind this study is to investigate the equality of test forms that had no anchors, were assumed to measure the same construct and were applied to different groups in different years. This design is known as the non-equivalent groups with covariates (NEC) design (Wiberg & Bränberg, 2015). In the non-equivalent groups with anchor test (NEAT) design, the anchor test score is used as a proxy for the latent variable of ability, while in the NEC design, covariates instead act as proxies of ability. The latter can therefore be viewed as a generalization of the NEAT design since the anchor test score can be seen as a covariate. The NEC design allows for the inclusion of more than one covariate.

Accordingly, the purpose of this research is to identify the statistical equality of the different test forms of the AYOS BLST using the kernel equating method under the NEC design.

2. METHOD

2.1. Research Model

The basic research approach was used as the aim of this research was to equate AYOS tests that were administered in 2017 and 2018, and were assumed to measure the same psychological construct by testing existing techniques on real data. As part of this aim, we utilize covariates gathered at the time of the test administration within the NEC design to equate the test forms. Evaluation of the results is conducted by calculating the standard error of equating (SEE) and the standard error of equating difference (SEED).

In basic research, which is a type of scientific research concerned with clarifying the underlying processes and better understanding the phenomena, the hypothesis is usually expressed as a theory (Fraenkel & Wallen, 2009). Basic research can be exploratory, descriptive, or

explanatory. Given that descriptive research is used to describe the characteristics of a population or phenomenon, which was part of the aim of this study, this specific type of basic research was applied.

2.2. The Study Group of the Research

The study group of this research consisted of 5,223 individuals who took the AYOS BLST – 2,460* took it in 2017, and 2,763* took it in 2018. In the 2017 group, there were slightly more men (52.2%), while there were slightly more women (52.19%) in the 2018 group. Regarding the age groups, about half of the individuals from the 2017 and 2018 groups were below 19 years of age (50.33% and 49.69%, respectively).

2.3. Data Set and the Test Equating Design

The research data included the test-takers' responses to the AYOS tests applied in 2017 and 2018. The AYOS is an assessment to determine international students' qualifications for admission to Ankara University and other universities (those accepting the AYOS score) in Turkey. The test is simultaneously implemented in different countries (exam centers) in a single session once a year. In brief, the AYOS dataset consists of the test-takers' scores (AYOS 2017 and AYOS 2018) and two covariates, gender (with values of 1 if man and 0 if woman), and age (with values of 1 if 18 years of age or younger and 0 if age 19 years of age or older). This means that there are $2 \times 2 = 4$ possible combinations of covariates and that the frequency vector has a length of $81 \times 4 = 324$. The data were first sorted by age followed by gender and the test scores on AYOS 2017.

The AYOS BLST is a non-verbal aptitude test with two sections and a total of 100 binary-scored multiple-choice items. The first section tests letter, number, and shape relations as a measure of cognitive skills, such as analytical thinking, reasoning, and abstract and spatial thinking (with 60 items). The second section measures numerical thinking skills that require the use of mathematics and geometry knowledge (with 40 items). The scores obtained from the test are valid for two years. The test is newly developed every year following the psychometric properties of the test applied in the previous year.

Table 1. Descriptive statistics of AYOS tests.

AYOS BLST	n	\bar{X}	S ²	S	KR-20**	Ave. Dif.	Skew.	Kurt.
2017	2.460	54.22	413.10	20.32	0.96	0.54	0.03	-0.86
2018	2.763	57.14	387.15	19.68	0.96	0.57	-0.00	-0.77

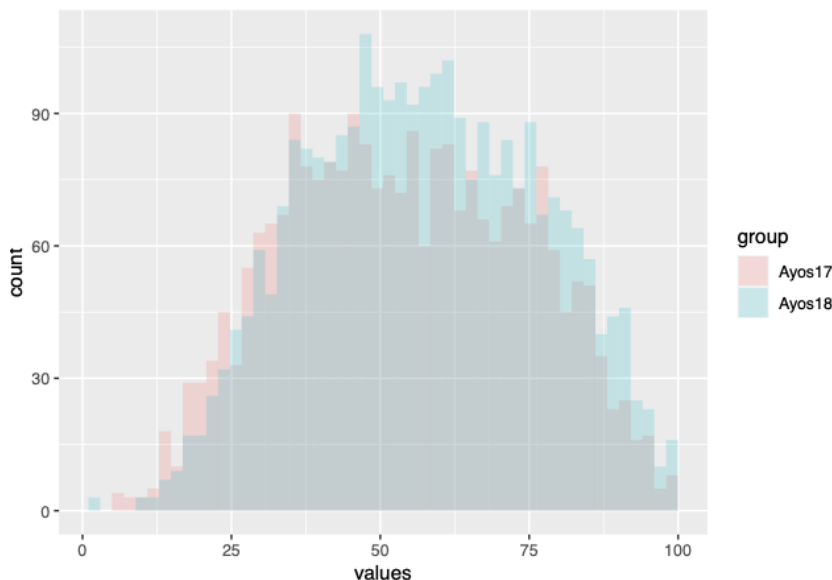
**The KR-20 formula was applied in cases where the items varied greatly in difficulty (Kuder & Richardson, 1937, p.160).

Table 1 shows the mean, standard deviation, variance, KR-20 reliability, average difficulty, skewness, and kurtosis coefficient values for the AYOS 2017 and 2018 tests. The first noteworthy finding was that the KR-20 reliabilities of both tests were equal and quite high (0.96). The KR-20 value is an overall measure of internal consistency (Cronbach, 1951, p.300) and provides information about the purity of random errors. Therefore, the fact that the values obtained from both tests were quite high is evidence that the tests involving the identification of number, shape, and letter relationships do measure the cognitive skills they aim to measure as a whole. Although the mean score on the 2018 test was higher than that on the 2017 test, both tests have values close to the average difficulty value of 0.50, which indicates that the

* Individuals from whom data on gender and age variables were collected were included in the study group.

students' scores generally hover around 50 points, which is the average score of the tests. This also shows that students can answer about 50% of the items on the test. Moreover, since the skewness coefficient of the score distribution in both tests is positive, the distribution is skewed to the right of what is considered normal. The kurtosis coefficients were negative for both tests, meaning that the score distributions, when compared to the normal distribution, were slightly flattened. The two score distributions are also presented in a histogram, which is given in [Figure 1](#).

Figure 1. *AYOS BLST Score Distributions.*



[Figure 1](#) shows that the score distributions are slightly skewed, with relatively few test-takers having low scores and many having high scores. This is reflected in the SEE plot ([Figure 3](#)). Considering the test design, a NEAT design is typically preferred in the test score equating, but some tests do not have common items. If the groups are non-equivalent, an equivalent groups (EG) design cannot be applied (Sansivieri & Wiberg, 2017). When the test groups are non-equivalent and no anchor items are available, Bränberg and Wiberg (2011), Andersson et al. (2013a), and Wiberg and Bränberg (2015) recommend that background information about the test-takers be used to adjust for the ability difference, a design referred to as the non-equivalent groups with covariates (Wallin, 2019).

The idea of test linking using variables is not new, as demonstrated by Kolen (1990) and Livingston et al. (1990), who suggested that linking can be used in cases of groups matching on variables other than ability (as cited in Wiberg & Bränberg, 2015). The NEC design is an important alternative to the NEAT design when there is no anchor test available for equating (Wiberg & Bränberg, 2015). In the NEC design, background information on the individuals taking the tests is used instead of using an anchor test to facilitate the equating of two tests when the groups taking the test are not equivalent (Andersson et al., 2013a). As is the case in the NEAT design, two groups are independently sampled from different populations, P and Q, and each is administered either of the test forms, X and Y. In the absence of an anchor test form, the NEC design uses relevant covariates, denoted by C, that can account for differences in the groups of test-takers (González & Wiberg, 2017).

According to Wallin and Wiberg (2019), equating non-equivalent test groups requires adjusting for two sources of bias: differences in the difficulty of the forms and differences in the abilities of the test groups. A proper equating conversion should address both of these, but when the

second is observed, some substitutes are required in place of ability. The most common substitute is an anchor test. However, since not all test programs can include an anchor, the background information of test-takers can be used. This is the scenario for the NEC design, where the fundamental assumption is that if the test groups are conditionally equivalent concerning the background information, they will differ only randomly from one another in terms of ability. The NEC design was applied in this study due to the non-equivalent test groups of AYOS and the absence of anchor items. According to Bränberg and Wiberg (2011), one important consideration when using background information is the choice of variables, which should be correlated with the test scores. On the other hand, the variables should “explain” the differences between the groups in the non-equivalent groups design. Accordingly, the covariates used in this study were age and gender, denoted as A and G, respectively based on the availability of the AYOS data.

2.4. Data Analysis

The scores on the AYOS BLST 2017 and 2018 tests were equated using the kernel equating method under the NEC design in this study (Wiberg & Bränberg, 2015). The R package “kequate” was used for kernel equating analyses (Andersson et al., 2013a, 2013b; R Core Team, 2018).

The analysis of the data was carried out in two stages. In the first stage, pre-smoothing, continuization, equating, and evaluation of the equating function (computing the SEE) processes were carried out. In the second stage, a linear equating function was used to determine the degree of difference in the results of the 2017 and 2018 tests, and the SEED was calculated.

2.4.1. The Kernel Equating Framework

The kernel method of test equating includes the following five steps (von Davier et al., 2004; Andersson et al., 2013a, 2013b; Wiberg & Bränberg, 2015; González & Wiberg, 2017; González & von Davier, 2017; Wallin & Wiberg, 2017, 2019): Pre-smoothing, Estimation of the Score Probabilities, Continuization, Equating, and Evaluation of the Equating Function (Calculating the SEE and SEED).

The goal of test equating – if we let X and Y denote the test score from test form X and the test score from test form Y respectively – is to equate X to Y (or vice versa). The test group that was administered the test form X is a sample from population P , while the group that was administered the test form Y is a sample from population Q . To define the kernel equating estimator used in this study, let $r_j = P(X = x_j)$ and $s_k = P(Y = y_k)$ denote the score probabilities for scores $x_j, j = 1, \dots, J$ and $y_k, k = 1, \dots, K$. Furthermore, let μ_X and σ_X^2 denote the mean and variance of the X scores, respectively, let V denote a continuous random variable with mean 0 and variance σ_V^2 , and let $a_X^2 = \sigma_X^2 / (\sigma_X^2 + \sigma_V^2 h_X^2)$, where h_X is a smoothing parameter called the bandwidth. Using these defined quantities, a continuous version of the random variable X was introduced:

$$X(h_X) = a_X(X + h_X V) + (1 - a_X)\mu_X.$$

The random variable $X(h_X)$ is defined as such that its mean and variance are the same as for X , and its cumulative distribution function (CDF) is given by

$$F_{h_X}(x) = P(X(h_X) \leq x) = \sum_j r_j K(R_{jX}(x)),$$

where $K(\cdot)$ is the kernel function following from the distribution of V (which is commonly set to the Gaussian distribution) and $R_{jX} = (x - a_X x_j - (1 - a_X)\mu_X) / a_X h_X$. Corresponding

quantities can be defined to introduce the continuized CDF G_{h_Y} . Replacing the terms in F_{h_X} and G_{h_Y} with estimated quantities, the kernel equating estimator used in this study was defined as

$$\hat{\varphi}_Y(x) = \hat{G}_{h_Y}^{-1}(\hat{F}_{h_X}(x)).$$

The SEE, which was used as part of the evaluation of $\hat{\varphi}_Y(x)$ in this study, equals

$$SEE_Y(x) = \|\hat{\mathbf{J}}_{\varphi_Y} \hat{\mathbf{J}}_{DF} \mathbf{C}\|,$$

where $\hat{\mathbf{J}}_{\varphi_Y}$ equals the Jacobian of the equating function, $\hat{\mathbf{J}}_{DF}$ equals the Jacobian of the design function that is set according to the data collection design, and \mathbf{C} is defined such that

$$\text{Cov}\left(\begin{matrix} \hat{\mathbf{R}} \\ \hat{\mathbf{S}} \end{matrix}\right) = \mathbf{C}\mathbf{C}^T,$$

with $\hat{\mathbf{R}}$ and $\hat{\mathbf{S}}$ denoting vectors of pre-smoothed score distributions. Lastly, we defined the SEED as

$$SEED_Y(x) = \|\hat{\mathbf{J}}_{\varphi_Y} \hat{\mathbf{J}}_{DF} \mathbf{C} - \hat{\mathbf{J}}_{\varphi_L} \hat{\mathbf{J}}_{DF} \mathbf{C}\|,$$

where φ_L equals the linear equating function

$$\varphi_L = \mu_Y + \left(\frac{\sigma_Y}{\sigma_X}\right)(x - \mu_X).$$

3. RESULTS

In the first stage of the equating process, pre-smoothing of the observed score distributions using the log-linear pre-smoothing was performed. A statistical model was fitted to the empirical distribution obtained from the sampled data in the pre-smoothing step. It is assumed that many of the irregularities observed in the empirical distributions are due to sampling error; thus, the pre-smoothing aims to reduce this error (Wiberg & Bränberg, 2015). Several log-linear models should be fitted and compared in the pre-smoothing step to decide which model fits the data the best (González & Wiberg, 2017).

3.1. Log-linear Pre-smoothing

González and Wiberg (2017) emphasize that several log-linear models should be fitted and compared in the pre-smoothing step regardless of the chosen data collection design. Here, the R function `glm()` was used to obtain a log-linear model in the pre-smoothing step to be used in the conjunction. The models were evaluated using the Bayesian Information Criterion (BIC; Schwarz, 1978), as it was shown to be an appropriate choice for bivariate smoothing (Moses & Holland, 2010). This led to log-linear models that preserved the first four moments of the X/Y score, the first two moments of the covariates, and the first cross-moment of the score variable and each covariate.

3.2. Estimation of the Score Probabilities

In the second step, the estimated score probabilities were generated by mapping the pre-smoothed score distributions into the score probability vectors for X and Y using a design function. This function, known as the design function, depends on the data collection design (see Wallin and Wiberg (2019) for the explicit expression of the design function for the NEC design).

3.3. Continuization

The Gaussian kernel was used in kernel equating to continuize the two estimated discrete cumulative distribution functions. The Gaussian kernel function is used to smooth the discrete

score distributions, and the full penalty function is applied to select the smoothing parameter (von Davier et al., 2004). The estimated distributions \hat{r}_j and \hat{s}_k , the bandwidths h_X and h_Y , and estimates of the means and variances of X and Y in population T were used in the application of the Gaussian kernel smoothing.

According to von Davier et al. (2004, pp.61-64), there is a variety of ways to select the bandwidth (h_X), which refers to controlling the degree of smoothness in the continuization, but the most common way was used in this study to minimize the penalty function.

The bandwidth for each continuized score distribution was selected by minimizing the sum of the squared distances between the observed score probabilities and the estimated density. To ensure smoothness in the estimated, continuized score distributions, the minimization operation included a term that penalized a density that had more than a few modes along with an added penalty term that penalized large fluctuations in the estimated density. Specifically, the bandwidth that minimized the following function was selected:

$$\sum_j (\hat{r}_j - F'_{h_X}(x_j))^2 + \sum_j A_j,$$

where $F'_{h_X}(x_j)$ denotes the derivative of $F_{h_X}(x_j)$, $A_j = 1$ if $f'_{h_X}(x_j - v) > 0$ and $f'_{h_X}(x_j + v) < 0$, or $f'_{h_X}(x_j - v) < 0$ and $f'_{h_X}(x_j + v) > 0$, and $A_j = 0$ otherwise.

3.4. Equating

In the last step, the results were graphically examined by plotting the equated scores (Figure 2) and SEE (Figure 3). The table presenting the equated scores can also be found in the appendix (Annex 1).

Figure 2. Equating results.

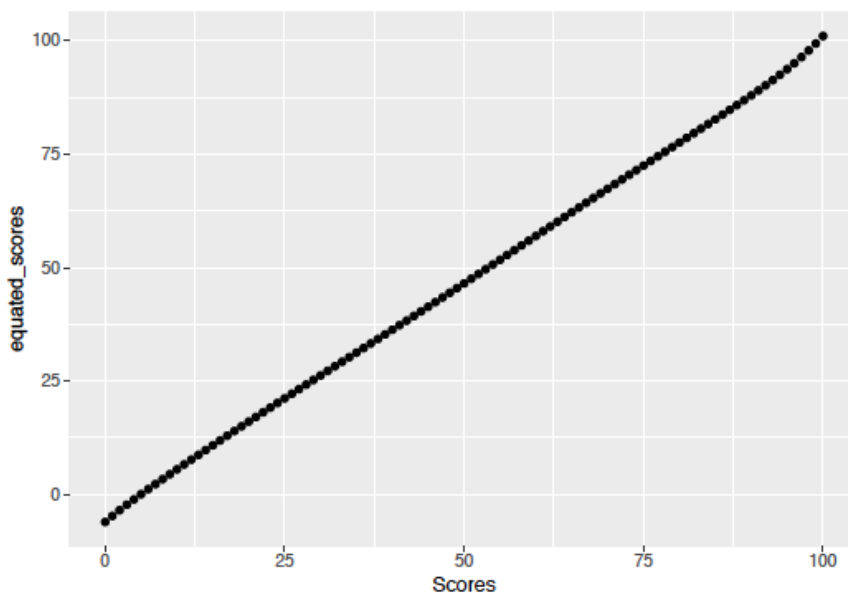
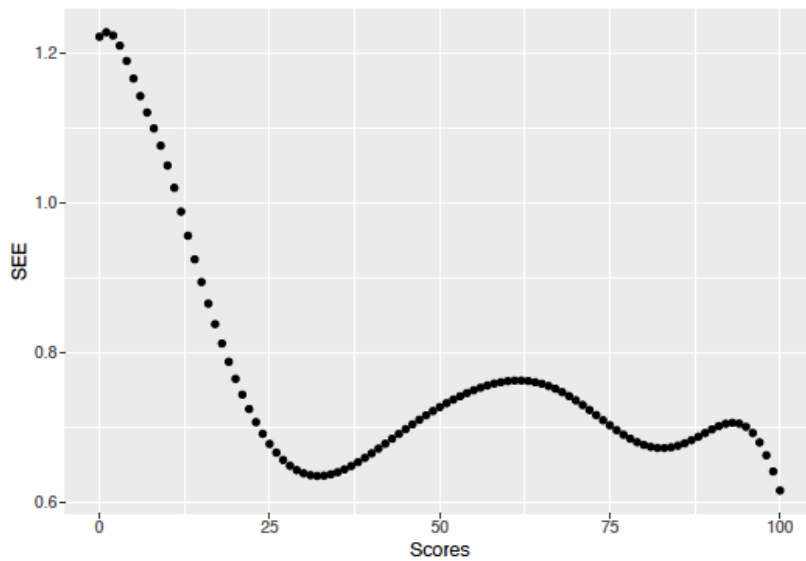


Figure 2 shows that there was a linear relationship between the raw scores and equated scores. Although the equating function was linear, there were non-linearities in the tails of the score distribution. It is also clear that the Y test form (AYOS 2018 test) was easier than the X test form (AYOS 2017 test), a difficulty difference that the equating function helped to adjust for.

3.5. Standard Error of Equating

Figure 3 shows the values of the SEE obtained for raw scores from the equating function.

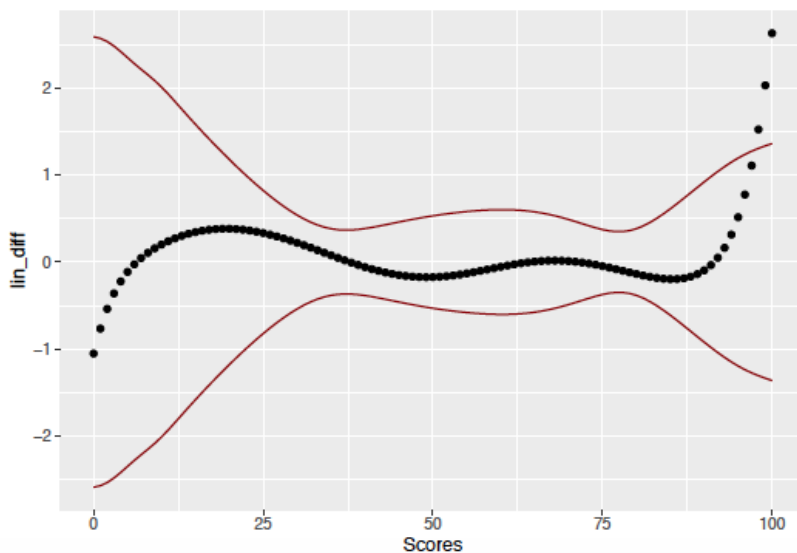
Figure 3. Standard error of equating.



As illustrated in [Figure 3](#), the SEE was larger at the lower end of the score scale. This is quite natural though, as there were very few test-takers with a total score below 10 (See [Figure 1](#)). Moreover, the SEE was relatively lower in the range of 25 - 100 scores.

Furthermore, an examination was performed to determine how the results obtained from a linear equating function would differ from the results already obtained. Therefore, the test forms were equated using a linear equation function, and then the difference between the previous equation function and the linear equating function was calculated. The results of this calculation indicated that the relationship between AYOS 2017 and 2018 tests was non-linear. In addition, the SEED was also calculated and added to [Figure 4](#).

Figure 4. Standard error of equating difference between current and linear equating function.



[Figure 4](#) shows that the linear equating function deviated from the non-linear equating function. The line indicated by black dots shows the difference, while the red lines ($\pm 2SEED$) represent twice the standard error of the difference between equating functions. The black line, however, only breaks through the SEED barrier once. This indicates that a non-linear equating method should be used instead of linear equating. Moreover, the SEED is relatively higher at the lower

and higher ends of the score scale, which means that the linear equation methods give higher standard errors in extreme scores than in middle scores.

4. DISCUSSION and CONCLUSION

The aim of this research was to investigate the equality of test forms that had no anchors, were assumed to measure the same construct and were applied to different groups in different years. To fulfill this aim, variables correlated with AYOS BLST scores were used as a substitute for common items in non-equivalent groups with covariates design. This method introduced in studies by Bränberg and Wiberg (2011), Andersson et al. (2013a), and Wiberg and Bränberg (2015).

The NEC design was specifically used because the test groups of the admission test were non-equivalent and there were no anchor items. Results from the analyses showed that the test forms had different score distributions and that the relationship was non-linear. The equating procedure was thus adjusted to eliminate these differences and thereby allow the tests to be used interchangeably. Real data from a non-verbal aptitude college admissions test were used.

In a similar study, Akin-Arkan (2020) used real data from the Monitoring and Evaluation of Academic Skills Project in Turkey to examine the NEAT and NEC designs comparatively. In this context, she equated the scores obtained from Mathematics subtests according to the kernel chained equipercentile, kernel post-stratification equipercentile, kernel chained linear, and kernel post-stratification linear methods. Furthermore, she sought to determine the affection status of the covariates (gender variable and socioeconomic index) used in the NEC design. From her research, it was determined that test forms can be equated using covariates when there are no anchor items. This is a noteworthy finding in terms of contributing valuable information for future studies to be carried out using the NEC design. When the findings obtained using the methods under the NEC design were specifically examined, the lowest error value was found in the design involving the socioeconomic index as a covariate, while the highest error value was found in the design involving the gender variable as a covariate. Akin-Arkan reported that the reason for this was the relationship between the covariates and the test.

In this research, the point-biserial correlations were very low, and for the values between the covariates, none of the correlations were statistically significant ($p > 0.05$). However weak correlation values between the covariates and the test scores do not mean that they are not good proxies of the latent ability. As we controlled for covariates that were confounders of the relationship between the test form assignment and the test score, we argue that as a rule the subject-matter knowledge of such covariates could be included to achieve a strong correlation. Similarly, Bränberg et al. (1990), in their research, found that there was a correlation between gender, education, and age in the test scores obtained from the Swedish Scholastic Aptitude Test (SweSAT).

In her research on real data, Yurtçu (2018) used gender, mathematics self-efficacy scores, and common item scores as covariates to obtain equated scores with the Bayesian nonparametric model. She concluded that covariates can be used in place of common items, and in some cases, perform even better, and that equated scores obtained with the said model can generate results closer to the target test.

The use of real-life data is important insofar as it reveals the psychometric properties of the tests used in real life. However, Wiberg and Bränberg (2015) warned that using real data is limiting because the true equating is not known. Therefore, simulation studies are recommended as they allow defining the true value of the equating (parameter) function, and they should be conducted using an NEC design within the kernel equating framework.

The evidence from the simulation study performed by Bränberg and Wiberg (2011) indicates that using covariates in the equating process can increase the accuracy of equating. In the present study, gender and age variables were used as covariates in the equating model. A review of the literature showed that background variables, such as gender, age, educational status, socioeconomic index, mathematics self-efficacy scores, etc., are being used as covariates (Bränberg & Wiberg, 2011; González et al., 2015; Wiberg & Bränberg, 2015; Wiberg & von Davier, 2017; Yurtçu, 2018; Akın-Arkan, 2020). There are additional factors that may affect the student's success. These include student background variables, as used in PISA, such as the number of books at home, time allocated to studying, etc., or high school grades and performance test scores of the students. These variables can be taken as covariates in test equation studies using the NEC design. González et al. (2015) stated that an additional advantage of including covariates in the modeling of the equating function is the possibility of a customized transformation between any pair of subpopulations as long as they are characterized by covariates.

Since the test groups were non-equivalent and the AYOS tests do not contain any common items, this analysis used background information about the test-takers to equate the test forms. Although common items are the gold standard for adjusting for ability imbalance between test groups, previous studies have shown that equating under the NEC design produces smaller standard errors (Sansivieri & Wiberg, 2017) and lower MSE (Bränberg & Wiberg, 2011). While the model specification and the kernel equating framework are somewhat more complicated (Andersson et al., 2013a), they have advantages in terms of modeling flexibility.

In this research, since there were no anchor items, the two covariates, gender and age, were used to equate the different test forms of AYOS. Using covariates to obtain equated scores in the Bayesian nonparametric model, Yurtçu (2018) emphasized that the use of two covariates was more effective than the use of anchor items. Similarly, in another study, it was stated that a large number of covariates would cause a decrease in the number of individuals who fall into common categories and thereby result in errors in score estimation (Wallin & Wiberg, 2017).

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Ozge ALTINTAS: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing of the original draft. **Gabriel WALLIN:** Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing of the original draft.

ORCID

Ozge Altintas  <https://orcid.org/0000-0001-5779-855X>

Gabriel Wallin  <https://orcid.org/0000-0002-7930-6701>

5. REFERENCES

- Akın-Arkan, Ç. (2020). The impact of covariate variables on kernel equating under the non-equivalent group design. *Journal of Measurement and Evaluation in Education and Psychology*, 11(4), 362-373. <http://dx.doi.org/10.21031/epod.706835>
- Altıntaş, Ö., & Kutlu, Ö. (2019). Investigating differential item functioning of Ankara University Examination for Foreign Students by Rasch model. *International Journal of Assessment Tools in Education*, 6(4), 602-616. <http://dx.doi.org/10.21449/ijate.554212>

- Altıntaş, Ö., & Kutlu, Ö. (2020). Investigating the measurement invariance of Ankara University Foreign Student Selection Test by latent class and Rasch model. *Education & Science, 45*(203), 287-308. <http://dx.doi.org/10.15390/EB.2020.8685>
- Anastasi, A. (1988). *Psychological testing* (6th ed.). Macmillan.
- Andersson B., Bränberg, K., & Wiberg, M. (2013a). kequate: The Kernel Method of Test Equating. *R package version 1.6.3*. <https://CRAN.R-project.org/package=kequate>
- Andersson, B., Bränberg, K., & Wiberg, M. (2013b). Performing the Kernel Method of Test Equating with the Package kequate. *Journal of Statistical Software, 55*(6), 1-25. <https://www.jstatsoft.org/v55/i06/>
- Angoff, W. H. (1971). Scale, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 509-600). American Council of Education.
- Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55-69). Academic.
- Bränberg, K., Henriksson, W., Nyquist, H., & Wedman, I. (1990). The influence of sex, education, and age on test scores on the Swedish Scholastic Aptitude Test. *Scandinavian Journal of Educational Research, 34*(3), 189-203. <https://www.tandfonline.com/doi/abs/10.1080/0031383900340302>
- Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48*(4), 419-440. <https://www.jstor.org/stable/41427533>
- Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). Academic.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). Harper Collins.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281-306. <https://doi.org/10.1111/j.1745-3984.2000.tb01088.x>
- Felan, G. D. (2002, February, 14-16). *Test equating: Mean, linear, equipercenile, and item response theory*. [Paper presentation]. The Annual Meeting of the Southwest Educational Research Associations, Austin, TX, United States. <https://files.eric.ed.gov/fulltext/ED462436.pdf>
- Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). McGraw-Hill.
- González, J., Barrientos, A. F., & Quintana, F. A. (2015). Bayesian nonparametric estimation of test equating functions with covariates. *Computational Statistics & Data Analysis, 89*, 222-244. <https://doi.org/10.1016/j.csda.2015.03.012>
- González, J., & von Davier, A. A. (2017). *An illustration of the Epanechnikov and adaptive continuization methods in kernel equating*. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology* (pp. 253-262). IMPS 2016. Springer Proceedings in Mathematics & Statistics, vol 196. Springer. https://doi.org/10.1007/978-3-319-56294-0_23
- González, J., & Wiberg, M. (2017). *Applying test equating methods using R*. Springer.
- Holland, P. W., & Thayer, D. T. (1985). Section pre-equating in the presence of practice effects. *Journal of Educational Statistics, 10*(2), 109-120. <https://www.jstor.org/stable/1164838>
- Kan, A. (2010). Test eşitleme: Aynı davranışları ölçen, farklı madde formlarına sahip testlerin istatistiksel eşitliğinin sınanması [Test equating: Testing the statistical equality of tests that measure the same behavior, and have different item forms]. *Journal of Measurement and Evaluation in Education and Psychology, 1*(1), 16-21. <https://dergipark.org.tr/en/download/article-file/65994>

- Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, 3(1), 97-104. https://doi.org/10.1207/s15324818ame0301_7
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151-160. <https://doi.org/10.1007/BF02288391>
- Kutlu, Ö., & Bal, Ö. (2011). *Ankara Üniversitesi Yabancı Uyruklu Öğrenci Seçme ve Yerleştirme Sınavı (AYÖS) projesi kesin raporu* [Ankara University Student Selection and Placement Exam for Foreign Students (AYOS) project final report]. (Project No. 11Y5250001). Ankara University Scientific Research Project Office.
- Levine, R. (1955). Equating the score scales of alternate forms administered to samples of different ability. *ETS Research Bulletin Series*, 55(2), i-118. Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.1955.tb00266.x>
- Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3(1), 73-95. https://doi.org/10.1207/s15324818ame0301_6
- Lord, F. M. (1950). Notes on comparable scales for test scores. *ETS Research Bulletin Series*, 50(48), 1-20. Educational Testing Service. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1950.tb00673.x>
- Moses, T., & Holland, P. W. (2010). A comparison of statistical selection strategies for univariate and bivariate log-linear models. *British Journal of Mathematical and Statistical Psychology*, 63(3), 557-574. <https://doi.org/10.1348/000711009X478580>
- R Core Team (2018). *R: A language and environment for statistical computing*. [Computer software]. R Foundation for Statistical Computing. <http://www.R-project.org/>
- Sansivieri, V., & Wiberg, M. (2017). IRT observed-score equating with the nonequivalent groups with covariates design. In L. A. van der Ark, M. Wiberg, S. S. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology* (pp. 275-285). IMPS 2016. Springer Proceedings in Mathematics & Statistics, vol. 196. Springer. https://doi.org/10.1007/978-3-319-56294-0_25
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464. https://projecteuclid.org/download/pdf_1/euclid.aos/1176344136
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer.
- Wallin, G. (2019). *Extensions of the kernel method of test score equating*. [Doctoral dissertation, Umeå University]. Umeå University Libraries. <http://umu.diva-portal.org/smash/get/diva2:1378833/FULLTEXT01.pdf>
- Wallin G., & Wiberg, M. (2017) Nonequivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. S. Culpepper, J. A. Douglas, & W. C. Wang (Eds.), *Quantitative psychology* (pp. 309-319). IMPS 2016. Springer Proceedings in Mathematics & Statistics, vol. 196. Springer. https://doi.org/10.1007/978-3-319-56294-0_27
- Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics*, 44(4), 390-414. <https://doi.org/10.3102/1076998619838226>
- Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement*, 39(5), 349-361. <https://doi.org/10.1177/0146621614567939>
- Wiberg, M., & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test. *International Journal of Testing*, 17(2), 105-126. <https://doi.org/10.1080/15305058.2016.1277357>

Yurtçu, M. (2018). *Parametrik olmayan Bayes yöntemiyle ortak değişkenlere göre yapılan test eşitlemelerinin karşılaştırılması* [The comparison of test equating with covariates using Bayesian nonparametric method]. [Doctoral dissertation, Hacettepe University]. Hacettepe University Libraries. <http://hdl.handle.net/11655/5295>

6. APPENDIX

Annex 1. Equating results.

Scores	Equated Scores	Scores	Equated Scores
0	-6.04	51	47.55
1	-4.72	52	48.59
2	-3.46	53	49.63
3	-2.25	54	50.67
4	-1.08	55	51.72
5	0.07	56	52.77
6	1.19	57	53.82
7	2.29	58	54.87
8	3.39	59	55.91
9	4.47	60	56.96
10	5.55	61	58.01
11	6.62	62	59.06
12	7.69	63	60.11
13	8.75	64	61.15
14	9.81	65	62.19
15	10.86	66	63.23
16	11.91	67	64.27
17	12.95	68	65.30
18	13.99	69	66.33
19	15.03	70	67.36
20	16.07	71	68.39
21	17.10	72	69.41
22	18.12	73	70.44
23	19.15	74	71.46
24	20.17	75	72.47
25	21.18	76	73.49
26	22.20	77	74.50
27	23.21	78	75.52
28	24.22	79	76.53
29	25.23	80	77.55
30	26.24	81	78.57
31	27.24	82	79.58
32	28.25	83	80.60
33	29.25	84	81.63
34	30.26	85	82.66
35	31.26	86	83.69
36	32.27	87	84.73
37	33.27	88	85.79
38	34.28	89	86.85
39	35.28	90	87.93
40	36.29	91	89.02
41	37.30	92	90.14
42	38.32	93	91.29
43	39.33	94	92.47
44	40.35	95	93.70
45	41.37	96	95.00
46	42.39	97	96.37
47	43.42	98	97.82
48	44.44	99	99.36
49	45.48	100	100.99
50	46.51		

PPSE P121 and P10 calculation method and related issues

Umit Celen ^{1,*}

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

ARTICLE HISTORY

Received: Feb. 05, 2021

Revised: May 30, 2021

Accepted: June 29, 2021

Keywords:

Score calculation,
PPSE P10,
PPSE P121,
Teacher appointments.

Abstract: This study examined the calculation methods of P121 and P10 scores used in teacher appointments. The statistics regarding the Public Personnel Selection Examination (PPSE) subtests used by Measurement, Selection and Placement Center (MSPC) in 2018, 2019 and 2020 were accessed from the website of the institution. The parameters not published on this webpage were calculated by using the candidates' results. The public openly debates the allegations made by the candidates who took the exam in 2019 that their scores had been miscalculated for various reasons and the examinee scores, in fact, had to be higher. The study was conducted (i) to determine whether such disparity actually existed, (ii) and if so, the reason behind it, (iii) how the differences arising from the parameters in the formula being used to calculate the scores would affect exam takers' scores. In particular, the study identified the issues caused by converting the scores obtained by using different subtests in the same manner in calculating P121 without considering an equating method. Based on the examined exam scores for the last three-years, it was concluded that 2019 candidates were disadvantaged in most teaching fields. Based on the findings, it is suggested that (i) the use weighted standard scores instead of P121 and P110, to calculate separate scores for each teaching field is better and (ii) the validity period of such exam scores should be limited to one year.

1. INTRODUCTION

Since 2002, the Ministry of National Education (MoNE) has been utilizing the scores obtained in the Public Personnel Selection Exam (PPSE) held annually by the Measurement Selection and Placement Center (MSPC), to appoint new teachers to its affiliated institutions. Before this exam, the Selection Exam for Civil Servants who would be appointed for the first time (SECS) had been used starting in 1999. Initially, the validity period of the exam scores was determined to be 2 years and it was implemented in this manner until 2013. Between 2013 and 2016, the validity period for the scores was 1 year, but with the change in the regulation published in the Official Gazette dated 15 August 2018, the validity period, which was increased to 2 years again starting with the 2017 exam, was reduced to 1 year again with the change on 7 November 2019.

*CONTACT: Ümit Çelen ✉ umitcelen@yahoo.com 📍 Amasya University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation, Amasya, Turkey

e-ISSN: 2148-7456 /© IJATE 2021

Initially, teacher candidates had to take the General Culture (GC), General Ability (GA) and Educational Sciences (ES) tests from the PPSE tests, however, MSPC announced on 27 February 2013 that Teaching Content Knowledge Test (TCKT) would be given in the fields of Turkish, primary school mathematics, science/science and technology, social studies, Turkish language and literature, geography, mathematics (high school), physics, chemistry, biology, religious culture and ethics, and foreign language (German, French, English). In 2014, counseling and classroom teaching, in 2017 pre-school teaching and in 2019 physical education and religious vocational school vocational teaching were added to these fields. In 2019, the number of questions in TCKT was increased from 50 to 75.

When the scores from two different exams are to be used in the same application, the equivalency of these scores becomes important. It is very difficult for tests consisting of different questions to be completely parallel and to produce the same or similar results for each individual who takes the test and therefore, this cannot be expected (Kan, 2010). This difficulty arises from the limitation of the Classical Test Theory which states that “all test and item statistics obtained are affected by the group to which the test is applied”. The statistics of test items depend on the sample and are interpreted depending on the group to which the test is applied. (Embreston & Reise, 2000).

In addition, it can be argued that individuals’ abilities may vary depending on the items they respond to and that they can perform differently in tests with different difficulty levels (Hambleton & Swaminathan, 1985). In specific, it is possible to equate the scores of tests applied to different people with different questions by using Item Response Theory framework (Crocker & Algina, 1986) but it is publicly known that such an equating process is not used in the exams conducted by MSPC. Instead, “the method of standardizing the scores by relieving them from the effects of mean and standard deviation” which is the most common method used for comparison (Tekin, 1996; Turgut & Baykul, 2010) is preferred.

In each of the PPSE tests, MSPC takes the difference between candidates’ correct numbers of questions, corrected for their lucky guessing, the average of the test. Then divides it by the standard deviation of the test to obtain the z-score and multiplies this score by 10 and adds 50 to it to obtain the t-score. Thus, within the limitations of the Classical Test Theory, the performances of the candidates in all tests become relatively comparable and collectable even though they have been taken from different tests. In the calculation of P121, t-scores are calculated by multiplying the sum of General Culture (GC) and General Ability (GA) test scores by coefficient 0.15, the Educational Sciences (ES) Test score by coefficient 0.2, and the Teaching Content Knowledge Test (TCKT) score by coefficient 0.5 and the Weighted Standard Score (WSS) is calculated. When P10 is calculated, WSS is found by multiplying and adding the scores of the GC, GA and ES tests with the coefficients 0.3, 0.3 and 0.4, respectively. Following this conversion, MSPC uses the following formula to get a score out of 100 for each PPSE score type (Measurement, Selection and Placement Center, 2019).

$$PPSE\ Score = 70 + \frac{30 [2 (WSS - X) - S]}{[2 (B - X)] - S}$$

Abbreviations

PPSE : Public Personnel Selection Examination

WSS : Weighted Standard Score

X : Average of the WSS distribution

S : Standard deviation of the WSS distribution

B : The highest score in the WSS distribution

Examination of PPSE guidelines shows that this formula was first used in 1999 in the Civil Servants Exam (CSE) manual, which was an exam given before the PPSE (Measurement, Selection and Placement Center, 1999). The formula includes the arithmetic mean and standard deviation from all candidates, as well as the B. Çelen's (2013) study based on 2010 PPSE data presents the effect of this transformation, applied to obtain the highest score 100 in each score type, on candidates' scores points out the possible validity issues in PPSE scoring methods.

Candidates who took the PPSE in 2019 claimed that they encountered an unfair situation as they were appointed together with the candidates who took the exam in 2018 because of the increase in the number of items in the teaching field knowledge test in 2019, the change in the duration of the exam, and the higher level of difficulty in the items. They brought it to the attention of the public that although they were at the top of the rankings, they were not appointed but the candidates who ranked lower in the 2018 exam were appointed. This study aimed to determine whether the professed unfairness in 2018 and 2019 scores really occurred and if there was indeed unfairness in the exam scores, to identify the reason behind it. For this purpose, answers to the following research questions were sought.

1. What are the X, S, and B values used in the P-121 account of the PPSE 2018, 2019 and 2020 exams?
2. According to the calculated parameters, does the P-121 corresponding to the same ASP change in 2018, 2019 and 2020?
3. Does the ASP required to have the score of the last appointed person in an appointment period vary in 2018, 2019 and 2020?
4. Are the assignment percentages of teachers appointed with 2018 and 2019 scores different?
5. What are the X, S, and B values used in the P-110 account of the PPSE 2018, 2019 and 2020 exams?

2. METHOD

In this research in the descriptive survey model, the population is the 2018, 2019 and 2020 PPSE scores. 70 result documents were achieved for each year. Thus, the sample consists of the result documents of 210 candidates.

The arithmetic means and standard deviation values of the PPSE subtests in 2018, 2019 and 2020 were used in this study. These values were taken from the MSPC web page. The formulas in the test manual were used for calculating the WSS of the candidates who took the exams. Based on candidates' correct and incorrect number of answers, corrections for lucky guessing were calculated and z-scores were obtained by taking the difference of these values from the mean and dividing them by the standard deviation. Then, these z-scores were multiplied by 10 and converted into t-scores by adding 50. The WSS were calculated by multiplying the t-scores with the coefficients of the PPSE subtests.

Unlike the 2018 manual, page 46 of the guide published by MSPC for 2019 PPSE included 18 TCKT score types (Measurement, Selection and Placement Centre, 2019). Again, page 23 of the same document specified that "PPSE score distribution will be obtained for each PPSE score type out of 100." When this specification and the PPSE P121 score types added to the guide in 2019 were taken together, there was a perception that P121 would be calculated by using the highest scoring WSS for each teaching field that year. Therefore, while trying to estimate the parameters of the exam, different calculations were made for each field, but it was understood that a single calculation method was used in all PPSE P121 score types, since the same B and the same X and S values were obtained in all fields.

The X, S and B values in the formula used in PPSE P121 and P10 score calculation are not published by MSPC. These values were found in the following manner: Since the formula is a

first order equation, there is a linear relationship between WSS and P121. A line can be created by using the data of 2 exam results related to the exam for a specific year. When the slope of this line and the point where it intersects the y axis is known, P121 corresponding to each WSS can be calculated. Based on the correct and incorrect number of items in the exam result, the candidate's WSS can be calculated but the error in the estimation of the correct items may be large due to the test mean and standard deviations and the rounding made by MSPC in the 5th digit after the comma in the calculated P121. In order to minimize this error, the error was reduced to 3 out of ten thousand by using 70 exam results for each exam year.

The following method was followed to determine how many people were appointed in the January 2020 Contractual Teacher Appointments with the results of 2018 and 2019 exam scores and the exam scores that provided the basis for these appointments. The field-based and institution-based base scores lists, created after these appointments and published by the Ministry of Education, were used to determine the rank of the last appointed person from a specific teaching field with 2019 scores. It was assumed that before the last candidate assigned with 2019 points, all candidates from a field were assigned with 2019 scores as well. It can be argued that this number was an upper limit for 2019 and the actual number would actually be lower because some candidates did not select a post, took both 2018 and 2019 exams or were actually assigned with a 2018 score. For example, the highest-ranking candidate in 2019 exam who scored 100 points was regarded to be appointed in this calculation, although he/she was not included in the appointment lists.

The lowest field-based and institution-based scores obtained as a result of appointments made by the Ministry of Education using 2018 and 2019 exam scores were accessed on the Ministry's website. The consent of the candidates whose PPSE results and scores were used in the study was obtained to be used in this study.

3. RESULT / FINDINGS

To answer the first research question, the parameters used in P-121 calculation were estimated with the operations described in the method section. [Table 1](#) presents the obtained values.

Table 1. *Parameters Used in 2018, 2019 and 2020 PPSE P121 Calculation.*

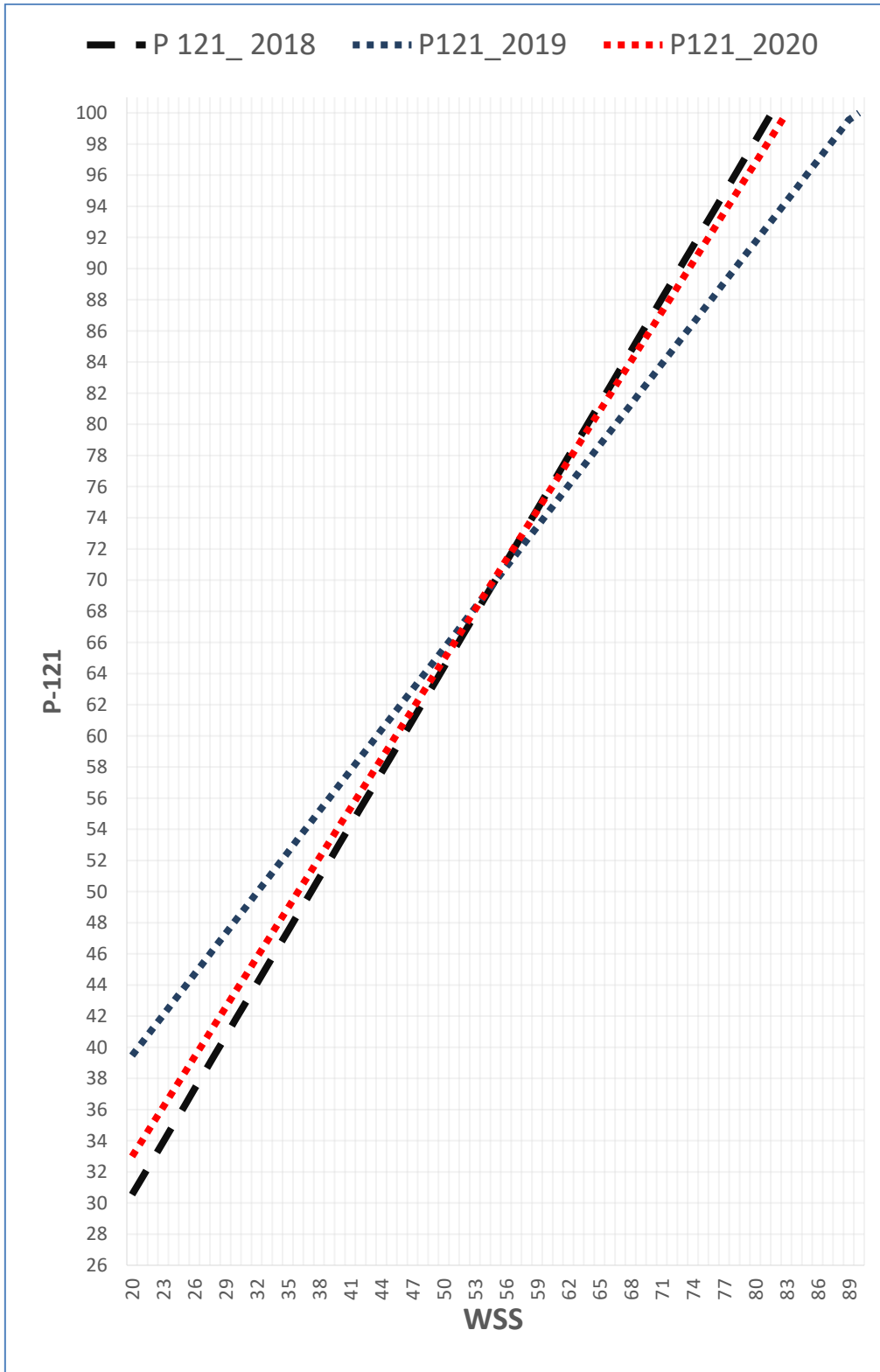
Year	B	$X + 0.5 \cdot S$
2018	81.857283	54.987993
2019	89.525328	55.070568
2020	83.029900	54.812020

The calculation was used to find the slope of the line produced by the P121 formula and the point where this line intersected the y-axis. Since the formula aimed to increase the score of the candidate with the highest WSS to 100, thanks to this relation, the WSS of the person who scored 100 could be found. Again, with the formula, the score of the candidate with a WSS over 0.5 standard deviation was found to be 70. Although X and S values could not be found separately, it was possible to calculate $X + 0.5 \cdot S$ value in this way.

Among the parameters used in the formula, X, which is the average of WSS, and S, its standard deviation, can be expected to be 50.0 and 10.0, respectively, since they are the values obtained from T scores. However, since candidates other than teacher candidates also participated in GC and GA tests, it was found that they differed from these values with small deviations. In this formula, B value was the most important parameter that would affect P121 in response to a WSS. The drawn line passed the point of score 70 around 55 WSS in all the 3 years, but there was a significant difference between years in the B value, which was the main value that would

determine the slope. While this value was 81.857283 in 2018, it increased to 89.525328 in 2019 and dropped to 83.0299 in 2020. In the form of a chart, Figure 1 presents the differences in score caused by these differences. The chart includes weighted standard scores on the horizontal axis and P121 scores corresponding to these WSS scores on the vertical axis.

Figure 1. The Change in P121 Scores Received by the Same WSS Values by Year.



The answer to the second research problem can be seen in [Figure 1](#), B value affected the slope of the line. High B value generated lower P 121 values for the WSS which were bigger than half standard deviation of the mean and generated higher P121 values for smaller WSS. In other words, those with a P121 score above 70 received lower scores than they should have while those below 70 scored higher. This situation was reversed when the B value was low. In that case, those who scored over 70 had high scores and those below 70 had low scores. If a candidate performed very high in any of these 3 years and achieved a very high WSS, other individuals with high WSS scores in the same year would receive lower P121 scores than they would get in another year's exam. This is beyond standardizing individuals' scores using mean and standard deviation. Mean and standard deviation values were obtained from the scores of all candidates taking the exam. While the number of candidates was approximately 400,000, the score of a single person with the highest score in the P121 calculation affected all other candidates' scores. The fact that B value affected the scores that much caused a significant difference on the 2019 exam as can be seen in [Figure 1](#). For example, the P121 of the candidates with a WSS of 80, that is, approximately 3 standard deviations above the candidates who took the exam with them, were 98.20978 in 2018, 91.70623 in 2019 and 96.77874 in 2020. If these three candidates selected teaching posts in the same appointment period, the 2019 candidate would be far behind the others since the appointments were score based and would even fall behind the candidates with 75 WSS in 2018 and 76 in 2020, despite having a higher level of achievement.

The situation was not different when the same chart was prepared according to the Z scores, which expressed in standard deviation the distance of the candidates to the WSS average so that the changes that may occur due to the differences of X and S values of each examination year could be included in the calculation ([Figure 2](#)). The horizontal axis included the Z scores calculated using the WSS received by the candidates and the mean and standard deviation of the whole WSS distribution showing their position within the WSS distribution while the vertical axis included the P121 scores corresponding to these Z scores. Analysis of the chart shows that the most disadvantaged group was the candidates of 2019, with scores above 70, the score that set the basis for higher number of appointments. An individual who performed 2.5 standard deviations above the average could obtain a score of 92.50245 in 2018 and a score of 90.46371 in 2020, while the equivalent of the same performance was a score of 87.65991 in 2019.

Apart from the effect described above, there is another aspect of the effect of B value on scores that causes unfairness. All candidates answered 4 tests. 3 of these tests were common, but TCKT tests were different tests consisting of different questions for each teaching field. For example, 18 different TCKT score types were defined in the 2019 PPSE (Measurement, Selection and Placement Center, 2019). The means and standard deviations of these tests were naturally different from each other. Despite the fact that these 18 score types were defined in the 2019 guide, unlike in 2018, and it was stated in the same guide that "PPSE score distribution will be obtained for each PPSE score type out of 100", it was understood that a single B was used in the calculations for all teaching fields. Therefore, this B value belongs to only one of the fields. The scores of the candidates in the other 17 teaching fields were determined according to the performance of the person with the highest score in a test they may or may not have taken. This situation is unacceptable in terms of principles in assessment and measurement. [Table 2](#) displays the Z and T scores that the candidates would get from the TCKT test in 2019 if all questions were answered correctly.

Figure 2. Change in P121 Scores Obtained by Candidates according to Distance to the WSS Average by Year.

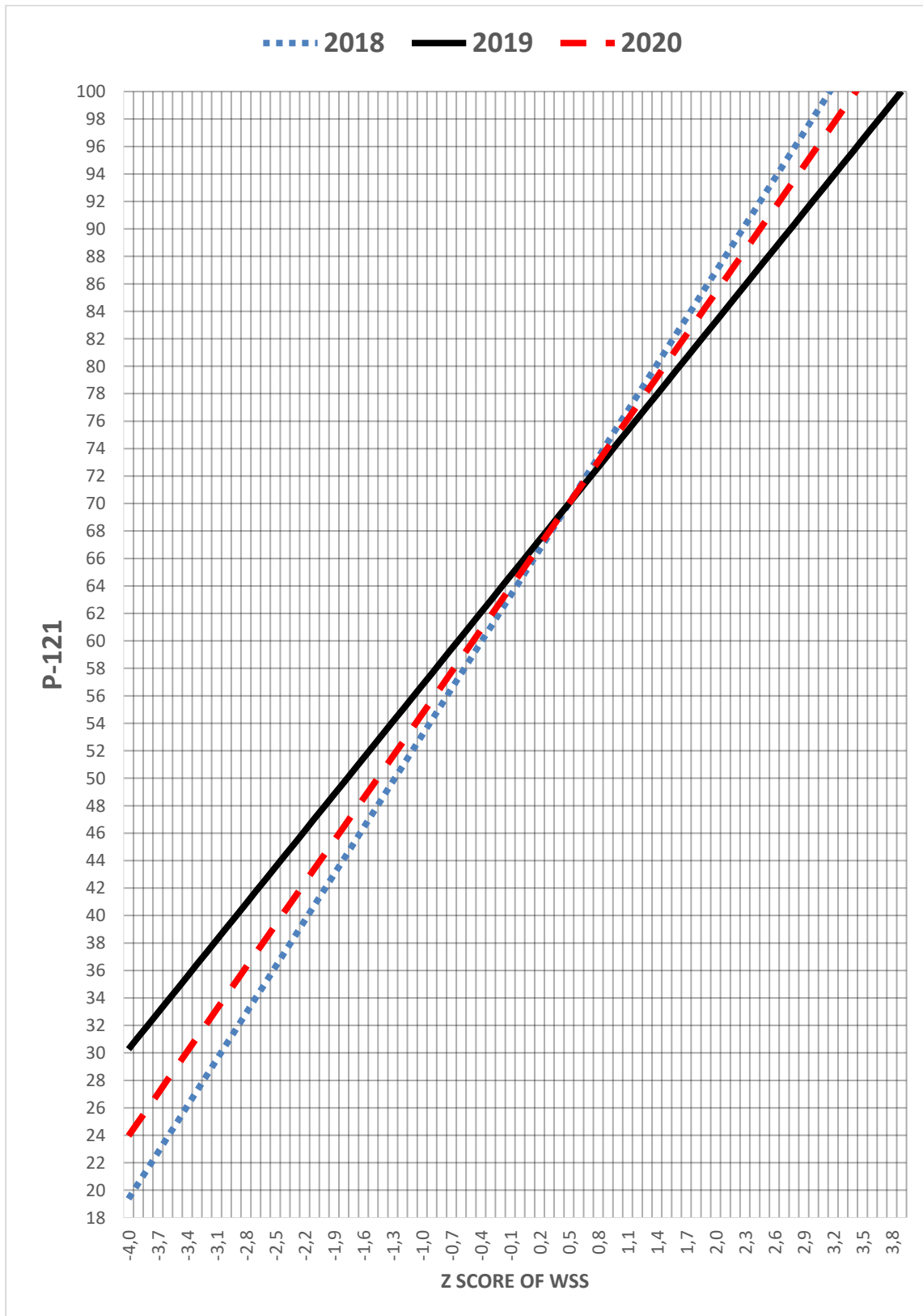


Table 2. Z and T Scores Corresponding to 75 Correct Answers in TCKT Tests in 2019.

Tests	Mean	S	Score that can be obtained with 75 correct answers	
			Z	T
Turkish Language Teaching	48.424	10.196	2.606512	76.06512
Elementary Mathematics Education	30.693	10.425	4.250072	92.50072
Science / Science and Technology	24.496	8.373	6.031769	110.3177
Social Studies	37.551	11.579	3.234217	82.34217
Turkish Language and Literature	27.951	13.867	3.392875	83.92875
History	32.482	13.628	3.119900	81.19900
Geography	35.347	12.424	3.191645	81.91645
Mathematics (High School)	24.268	12.425	4.083058	90.83058
Physics	32.032	15.711	2.734899	77.34899
Chemistry	28.278	14.364	3.252715	82.52715
Biology	25.229	10.802	4.607573	96.07573
Religious Culture and Ethics	42.599	11.592	2.795117	77.95117
Foreign Language (English)	33.863	14.283	2.880137	78.80137
Counselor	50.568	11.481	2.128038	71.28038
Classroom Teaching	32.434	8.823	4.824436	98.24436
Pre-school Teaching	37.382	10.578	3.556249	85.56249
Physical Education Teaching	26.514	8.256	5.872820	108.7282
Religious Vocational School Vocational Classes	38.082	10.353	3.565923	85.65923

As Table 2 shows, the arithmetic means of TCKT tests in the 2019 PPSE varied between 24,268 and 50,568. Candidates who achieved high success in difficult tests could get very high Z and T scores because they were far away from the average of the distribution. In easy tests, on the other hand, Z and T were lower because there was not much difference between the average and 75, the number of questions in the test. For example, a candidate taking the counseling teacher TCKT can have a score of 71.28038 T with 75 corrected score, while the candidate who answers all questions correctly in the science TCKT test can get 110.3177 points. The difference in B will cause unfair decisions from year to year even if the candidates take the same tests. Although it is clear enough, some people would still argue that “if someone can perform very high in an exam, the other candidates after him/her would of course score lower”. There are two arguments against this supposition:

1. While this high B value decreases the scores up to 70, it increases the scores below 70, 2. It is not likely to get high scores in some tests anyway. It is clear that calculating the scores of the individuals with a parameter of a test whose average and standard deviation are different from the test they have taken (since candidates can only take one TCKT and no chance of taking the test whose results are being used to calculate results) and comparing these scores with the scores from another year that are valid for 2 years can lead to serious unfairness among candidates. Even if the candidate answered all the questions in the GC, GA, ES and TCKT tests correctly in 2019, a candidate could not have exceeded the B value of 89.525328, used in the score calculation of this exam in 14 of the 18 fields where teachers were appointed with this score type. Since other tests were common, it was necessary to have a high net in a TCKT area with a low average and standard deviation in order to be the first in the exam, others would not have such a chance.

The above section discussed the reasons for the unfair scoring in the exams held in different years due to the calculation method in the P121 formula (research problem 3). Below, the

consequences of these inequalities were presented through practical examples. Table 3 presents the field-based base score and number of appointments for some teaching fields in January 2020 (Ministry of National Education, 2020). The table also includes the WSS required to get a base score in the exams of the last 3 years.

Table 3. Field Based Base Scores used in January 2020 Contractual Teacher Appointments and the Required WSS to Get These Scores*

Field	Number of Appointments	Base Oral Exam Score	Base PPSE Score	Required WSS to Get the Base Scores		
				2018	2019	2020
Biology	93	81	80.54589	64.33843	67.18244	64.73144
Geography	147	81	81.03726	64.77410	67.74677	65.19362
Science	1002	76	76.01989	60.32549	61.98436	60.47430
Physics	156	75	74.96992	59.39454	60.77848	59.48671
Primary School Mathematics	1699	74	74.02994	58.56111	59.69892	58.60257
English	1731	71	70.82538	55.71981	56.01851	55.58837
Chemistry/ Chemical Technology	151	77	76.54140	60.78788	62.58331	60.96483
Mathematics (High School)	501	79	79.26013	63.19842	65.70575	63.52206
Pre-school	1513	77	77.44372	61.58792	63.61962	61.81355
Counseling	1257	79	79.05386	63.01553	65.46885	63,32804
Classroom Teaching	3007	75	74.55471	59.02639	60.30162	59.09616
Social Studies	684	80	79.68232	63.57275	66.19064	63.91917
History	197	81	81.16877	64.89070	67.89781	65.31732
Turkish Language and Literature	384	83	82.92439	66.44731	69.91413	66.96865
Turkish Language Teaching	1293	77	77.22951	61.39799	63.37360	61.61207

*The fields for which there are no TCKT Exam or where the Foreign Language Exam test score is used while calculating the P121 are not included. Physical Education and Religious Culture and Ethics, whose calculation of WSS-P-121 score conversion can be predicted with relatively more errors due to cancelation of items, are also not included in the table.

Examination of the scores in Table 3 points to the need for a higher WSS in 2019 in order to have the score of the candidate who was appointed the last in all fields. Since this difference was caused by the B value in the calculation and the base scores were all higher than 70, 2019 was the most disadvantaged year in which the largest value B was used. This disadvantage was the greatest in the field of Turkish language and literature, where the score required for appointment was the highest. For candidates who took the 2019 exam to be appointed, their performance in the exam should have been 3.47 standard scores higher than that of 2018 and 2.95 standard scores higher than that of 2020. This difference was lower in the field of English language teaching, where base score for appointment was close to 70. There is another point to take into consideration here: Although the most important decision in teacher appointments is related to whether teachers would or would not be appointed in the first place, there is also the issue about where they would be appointed. Since this decision is also taken according to the level of their scores, the appointed candidates may not be assigned to their first choices if they do not have a very high score.

Table 4 presents the contractual teacher appointments of June 2020 for some teaching fields with the base score and the number of appointments together with the WSS required to get the base score in the exams of the last 3 years (Ministry of National Education, 2020).

Table 4. Field Based Base Scores used in June 2020 Contractual Teacher Appointments and the Required WSS to Get These Scores *

Field	Number of Appointments	Base Oral Exam Score	Base PPSE Score	Required WSS to Get the Base Scores		
				2018	2019	2020
Biology	94	80	79.81466	63.60090	66.34263	64.04365
Geography	138	80	80.25618	64.08156	66.84971	64.45894
Science	1026	74	74.16045	58.67683	59.84881	58.72532
Physics	151	74	74.18967	58.70273	59.88237	58.75281
Primary School Mathematics	1701	71	70.51645	55.44590	55.66371	55.29779
English	1739	68	67.90726	53.13248	52.66707	52.84360
Chemistry/ Chemical Technology	154	76	75.64861	59.99629	61.55795	60.12508
Mathematics (High School)	498	78	78.10515	62.17437	64.37969	62.43569
Pre-school	1518	76	75.97089	60.28204	61.92809	60.42822
Counseling	1373	78	77.87273	61.96829	64.11234	62.21708
Classroom Teaching	2831	72	71.76993	56.55729	57.10332	56.47681
Social Studies	665	79	78.65655	62.66326	65.01255	62.95434
History	201	81	80.56881	64.35875	67.20876	64.75300
Turkish Language and Literature	344	82	82.17302	65.78112	69.05118	66.26191
Turkish Language Teaching	1300	76	75.99971	60.30759	61.96119	60.45532

*The fields for which there are no TCKT Exam or where the Foreign Language Exam test score is used while calculating the P121 are not included. Physical Education and Religious Culture and Ethics, whose calculation of WSS-P-121 score conversion can be predicted with relatively more errors due to cancelation of items, are also not included in the table.

Table 4 shows that all fields of teaching except English were disadvantageous in 2019 in terms of the standard score required to get the base score in that field and to be assigned to a teaching post. This disadvantage was more prominent in fields such as Turkish language and literature, history, geography, biology, where higher scores were required to be appointed. Since the base score in the English field fell below 70 points, the critical threshold produced by the P121 score calculation formula, it worked the other way and candidates with 2018 scores in this field were disadvantaged during this appointment period.

Since the validity of the exam scores is 2 years, candidates who took the 2018 and 2020 exam with a relatively equivalent B value will not make a choice in the same appointment period. However, 2019 candidates will continue to experience the victimization they have experienced with the 2018 candidates in the future appointment periods in which they will apply with the 2020 candidates. In these appointments, 2020 candidates with lower standard scores will be ahead of the 2019 candidates and will be appointed before them.

Below are the scores of some candidates along with their exam results and the scores they would receive in other years' exams if they performed similarly. The aim here is to reveal the unfairness described above with concrete examples.

The social studies teacher candidate, whose score reports given in [Figure 3](#), obtained a score of 78.64483 and could not be appointed to a teaching post in January and June 2020 appointment periods with this score. This candidate's WSS was 64.99914554. If the candidate succeeded in obtaining this standard score in 2018 and 2020, the candidate's scores would be 81.291107 and 80.8305, respectively.

In the June appointment period, the base score remained at 78.65655. The candidate who received the same WSS in 2018 was assigned a teaching post, but this candidate could not be appointed. While this candidate fell behind the candidates who received lower scores in 2018 in the appointment periods, he/she will still be behind the candidates with lower standard scores in the 2020 exam in the appointment periods for 2021, and probably will not be appointed to any teaching post with this score.

The history teacher candidate whose score report is presented in [Figure 4](#) had a P121 score of 80.17244. The candidate could not be appointed because he/she fell approximately 0.4 scores behind the last appointed person in the June 2019 appointment period. The candidate's WSS was 66.75334794. If the candidate with the highest WSS in this exam from another TCKT field had received the score that could be received in the 2018 exam or in the 2020 exam, the P121 that the history teacher candidate would get with this WSS would be 83.26955 and 82.695549, respectively. Then, the candidate would be above the base score with both of these two scores, but he/she fell behind those with lower standard scores and could not be appointed.

The Turkish Language teacher candidate with the above score report ([Figure 5](#)) could not be appointed to a teaching post due to obtaining only 0.003 points lower than the base score in June 2020 appointment period. However, the equivalent of this performance was WSS 61.95714931. Despite having a WSS score of about 1.2 standard deviations higher than those who took the exam with him/her, the candidate was even behind the 60.30759 WSS candidates in the 2018 exam. If he/she had obtained the same WSS in the 2018 exam, the candidate would have been appointed as a teacher with a score of 77.86016.

The score report presented in [Figure 6](#) is from the field of biology in 2019. The candidate scored 79.80467 and ranked 84th among the candidates who took the exam in this field but was not appointed to a teaching post with 187 open positions during the 2 appointments periods in 2020. The fact that the candidates with 2018 exam scores were placed ahead of the candidate who ranked 84th in the 2019 exam after participating in 2 previous appointment periods can only be explained by the unfair scoring between the exams held in these two years. Since the B value in the 2018 exam was lower than that of the 2019 exam, the 2018 exam candidates who received a lower standard score were placed ahead of this candidate's score. However, if the same candidate had received the same standard score in the 2018 exam, his P121 score would be 82.79361 and would be within the limits that would make it possible for him/her to be appointed. The candidate could only be appointed to a position he/she did not choose in the additional appointment periods available in 2020.

It is possible to present hundreds of examples in this regard. This outcome was experienced because the PPSE of 2019 had an extremely high B score when compared to other years and this B value in the score calculation had no function in formula other than ensuring that the highest score was 100.

Figure 3. A Score report from Social Studies Field in 2019.

Education Level		Bachelor's Degree									
Date of Exam		14th, 20th, 21st, 28th July 2019									
NUMBER OF TRUE AND FALSE ANSWERS FOR EACH TEST											
General Aptitude		General Culture		Educational Science		Teaching Field Knowledge (Morning)		Teaching Field Knowledge (Noon)			
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
28	8	47	12	57	22	61	12	-	-		
Public Administration		International Relations		Labor Econ. & Industrial Rel.		Law					
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
-	-	-	-	-	-	-	-	-	-	-	-
Economy		Finance		Business		Accounting		Statistics			
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
-	-	-	-	-	-	-	-	-	-	-	-
SCORES AND RANKS											
SCORE NAME	SCORE	RANK	No of Candidates	SCORE NAME	SCORE	RANK	No of Candidates	SCORE NAME	SCORE	RANK	No of Candidates
P1	73.03193	112103	602945	P2	75.12638	78923	602945	P3	77.39235	54883	602945
P10	78.44819	33551	356471	P121-4	78.64483	867	17547				

Figure 4. A Score report from History Field in 2019.

Education Level		Bachelor's Degree									
Date of Exam		14th, 20th, 21st, 28th July 2019									
NUMBER OF TRUE AND FALSE ANSWERS FOR EACH TEST											
General Aptitude		General Culture		Educational Science		Teaching Field Knowledge (Morning)		Teaching Field Knowledge (Noon)			
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
29	6	52	6	66	11	56	17	-	-	-	-
Public Administration		International Relations		Labor Econ. & Industrial Rel.		Law					
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
-	-	-	-	-	-	-	-	-	-	-	-
Economy		Finance		Business		Accounting		Statistics			
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
-	-	-	-	-	-	-	-	-	-	-	-
SCORES AND RANKS											
SCORE NAME	SCORE	RANK	No of Candidates	SCORE NAME	SCORE	RANK	No of Candidates	SCORE NAME	SCORE	RANK	No of Candidates
P1	77.94967	42027	602945	P2	80.39786	23624	602945	P3	83.07494	13059	602945
P10	86.70711	3730	356471	P121-4	80.17244	269	19936				

Figure 5. A Score report from Turkish Language Teaching Field in 2019.

Education Level		Bachelor's Degree									
Date of Exam		14th, 20th, 21st, 28th July 2019									
NUMBER OF TRUE AND FALSE ANSWERS FOR EACH TEST											
General Aptitude		General Culture		Educational Science		Teaching Field Knowledge (Morning)		Teaching Field Knowledge (Noon)			
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE		
27	4	43	9	55	19	63	10	-	-		
Public Administration		International Relations		Labor Econ. & Industrial Rel.		Law					
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE		
-	-	-	-	-	-	-	-	-	-		
Economy		Finance		Business		Accounting		Statistics			
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE		
-	-	-	-	-	-	-	-	-	-		
SCORES AND RANKS											
SCORE NAME	SCORE	RANK	No of Candidates	SCORE NAME	SCORE	RANK	No of Candidates	SCORE NAME	SCORE	RANK	No of Candidates
P1	74.38420	87622	602945	P2	75.93372	67280	602945	P3	77.59316	52624	602945
P10	78.13399	35699	356471	P121-4	75.99621	2251	16481				

Figure 6. A Score report from Biology Field in 2019.

Education Level		Bachelor's Degree									
Date of Exam		14th, 20th, 21st, 28th July 2019									
NUMBER OF TRUE AND FALSE ANSWERS FOR EACH TEST											
General Aptitude		General Culture		Educational Science		Teaching Field Knowledge (Morning)		Teaching Field Knowledge (Noon)			
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
47	6	40	13	52	26	49	24	-	-	-	-
Public Administration		International Relations		Labor Econ. & Industrial Rel.		Law					
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
-	-	-	-	-	-	-	-	-	-	-	-
Economy		Finance		Business		Accounting		Statistics			
TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
-	-	-	-	-	-	-	-	-	-	-	-
SCORES AND RANKS											
SCORE NAME	SCORE	RANK	No of Candidates	SCORE NAME	SCORE	RANK	No of Candidates	SCORE NAME	SCORE	RANK	No of Candidates
P1	87.52383	2452	602945	P2	87.29447	2757	602945	P3	87.00006	3728	602945
P10	82.99804	11601	356471	P121-4	79.80467	84	5568				

Another investigation performed in the framework of this study included the comparison of 2018 and 2019 exam scores with the number of appointed teachers (research problem 4). However, since the necessary data for this analysis were not published, a specific method was followed which was mentioned in regard to the limitations in the method section. The number that was attained as the number of candidates appointed with 2019 exam scores was only an upper limit, the actual result is likely to be lower than that figure. For some fields, the calculation could not be completed because the candidate last assigned to a teaching post could not be reached. [Table 5](#) displays the data on the fields for which calculations can be performed.

Table 5. 2019 and 2020 Base Scores and Appointment Rates in Some of the Teaching Fields.

Field	Total Quota for 2020 Appointments	2019 Base Score	2020 Base Score	Appointed with 2018 Score		Appointed with 2019 Score	
				Number	%	Number	%
Biology	187	81.57524	79.81466	104	55.62	83	44.38
Geography	285	81.62987	80.25618	94	32.92	191	67.08
Science	2028	75.22787	74.16045	141	6.95	1887	93.05
Physics	307	75.45017	74.18967	72	23.45	235	76.55
Chemistry/ Chemical Technology.	305	77.88505	75.64861	107	35.08	198	64.92
Mathematics (High School)	999	80.89463	78.10515	328	32.83	671	67.17
Counseling	2630	79.42048	77.87273	400	15.21	2230	84.79
Social Studies	1349	80.28933	78.65655	486	36.03	863	63.97
History	398	81.48067	80.56881	194	48.74	204	51.26
Turkish Language and Literature	728	83.60853	82.17302	392	53.85	336	46.15
Turkish Language Teaching	2593	77.17681	75.99971	345	13.31	2248	86.69

[Table 5](#) shows that the base scores decreased in all teaching fields included in the table for 2020 appointments. Except for the field of teaching English, which is the only area not included in this table and previously determined to be more advantageous in 2019 exam scores, all base scores dropped in 2020 appointments. For instance, if we take the field of biology teaching as an example, no candidates who participated in the appointment periods with their 2018 exam scores above 81.57524 points should have been left for future periods, (if they had selected their preferred teaching positions and if they had selected enough number of positions to be appointed). In 2020 appointment periods, the candidates with 2019 exam scores which were higher than this base score should have been assigned to posts first, and if there was still a quota, the two groups should have been assigned to these quotas in a mixed manner. However, since the B value of the highest scoring WSS in the 2019 exam, was approximately 7.7 points higher than that of 2018, the scores of the applicants from this field were not very high. Even if all questions were answered correctly in some field tests, it was not possible to score as high as one could in the field of science, after all. After a small number of candidates with 2019 points were appointed, the candidates with both years' scores were placed in their preferred posts and the quota was filled with candidates who scored 79.81466. The ratio of appointments shows

that 44.38% of those appointed from this field were the candidates from 2019 the most. The percentage of 2018 candidates already included in two prior appointment periods and appointed with 2019 candidates was found to be much higher than expected.

It is possible to explain the fact that 2019 candidates were appointed relatively in higher numbers in fields such as science, counseling, and Turkish, where both the number of candidates taking the exam and appointment quota was high: Since the number of applicants was 17,460 in science, 16,916 in counseling and 16,548 in Turkish, it can be expected that the number of people who scored above the base score of the previous year would be higher than a field where less candidates were available such as the field of biology with 5662 candidates. Another reason was related to the fact that the base scores in these fields were closer to 70 where the difference between years was zero. In the case of Turkish language and literature where the number of applicants was high and which has the highest base score, more than half of those appointed were 2018 candidates, despite having lower standard scores just as the case in the field of biology. In addition, since the standard deviation of the Turkish language and literature field test was as high as 13,867, candidates who took the exam in this field did not have the opportunity to get a higher standard score even if they answered almost all of the questions correctly.

Since the B value used when calculating the 2020 P121 scores was approximately 6.5 points lower than that of 2019, 2020 candidates will be advantageous in both appointment periods in 2021 in which the 2019 and 2020 candidates will apply and 2019 candidates will not be appointed although they obtained high standard scores.

To answer the research question 5, the 2018 and 2019 exams were also examined in terms of fields where appointments were made with P10 scores in addition to the P121 scores in the TCKT test. P10 is calculated over the weighted standard scores found by converting the scores of general culture, general ability and educational sciences tests into T scores and multiplying them by the coefficients 0.3, 0.3 and 0.4. Since the same formula is used, the scores are affected by the B value, which is the score of the candidate with the highest WSS, as in the P121 calculation.

Table 6. *B Values Used in 2018, 2019 and 2020 PPSE P10 Calculation.*

Year	B
2018	77.142393
2019	80.683712
2020	79.483411

When the B values were examined by years presented in [Table 6](#), it was found that the highest value of B was obtained in the 2019 exam while the lowest value of B was in the 2018 exam. 2018 candidates experienced a very disadvantageous situation over P121 points previously described in a detailed manner in areas with a base score over 70. 2018 candidates were advantageous in 2018-2019 mixed appointments in these fields. 2020 candidates will be more advantageous in 2019-2020 mixed appointments as well. The advantageous group in a small number of fields such as teaching music with a base score below 70 is the 2019 candidates.

Another issue that may create unfairness in the 2019 PPSE was experienced in the fields of physical education teaching, and in religious vocational high school vocational courses. In 2018, there was no TCKT in these fields and the basic score for appointment for a post was the P10 score. In 2019, TCKT became mandatory in these two fields and the base score for appointment for a post changed to P121. It is considered to be a problematic practice to treat

the P10 scores calculated with the standard scores obtained from 3 tests as the equivalent of P121 scores obtained from 4 tests without applying any equalization procedure in the mixed assignments of 2018 and 2019 candidates.

4. DISCUSSION and CONCLUSION

The following results are obtained based on the investigation described in a detailed manner in the findings and interpretations sections of the PPSE P121 and P10 scores which constitute the basis for teachers' appointment to teaching posts and specific teaching institutions.

The teaching fields where appointments were done with the P121 score, unfair decisions were observed in the 2018-2019 mixed appointment process due to the existence of different tests for each field, the differences in the test statistics for these tests and the effect of the weighted standard score of a candidate over the scores of all candidates. The reason for the unfairness observed here is not only because the statistics of the test items, which is the limitation of the classical test theory stated by Embreston and Reise (2000), depend on the sample. The main reason is to make an extra point conversion after choosing the way of "standardizing the scores by freeing them from the effects of the mean and standard deviation" (Tekin, 1996; Turgut & Baykul, 2010) suggested by the Classical Test Theory.

Some TCKTs were difficult for the candidates and the average was low. In some others, the average was higher. Hambleton & Swaminathan's (1985) criticism that individuals' abilities may vary depending on the items they answer and that they may perform differently in tests of different difficulty is a criticism of using the scores of two different tests that are claimed to measure the same thing together. Here, not only are the difficulties of the tests different, but also the features they measure. In high-average tests, even answering all the questions correctly was not enough to get a high standard score. When calculating P121, using the B value obtained in one of the 18 fields to calculate the scores of the other fields have resulted in lower scores in the TCKT fields with a large mean and standard deviation. This causes field-based unfairness while using the scores from different years in the same appointment period.

Since the B value calculated for 2019 was much higher than that of 2018 and 2020, this created an unfairness against the 2019 candidates in fields with a base score over 70. This situation, also noted in Çelen's (2013) study which presented how the differentiation of the B value would cause a problem in comparing the scores obtained in different years, resulted in a very high level of unfair decisions due to the difference of approximately 8 points in the B value.

The inequalities were also reflected in the number of appointments in related fields. Although 2018 candidates in some fields were included in the appointment for the 4th time, they were appointed at a higher rate than 2019 candidates who were only included in the first two appointments.

Similar investigations show that the scores obtained in the 2019 exam with the highest B value were disadvantageous compared to other fields in areas with a base score of more than 70 even when the P10 was used for appointments in the fields. 2018 candidates were also disadvantaged in a small number of fields with base scores lower than 70 such as teaching music. In the mixed appointment period where 2019-2020 candidates will participate, 2020 candidates will be advantageous in high scoring fields.

TCKT was added to the fields of physical education and religious vocational high school vocational courses in 2019 creating another unfair application by matching the 2018 P10 scores obtained from 3 tests with the P121 scores obtained from 4 completely different tests.

The cases examined in this study which believed to create unfairness should not be considered only in relation to being appointed in a specific period or not. It is an undeniable fact that

unfairness in the scores will affect not only being appointed, but also being appointed at an earlier appointment period and being appointed to a higher-ranking institution.

The following suggestions can be made to the Ministry of National Education and MSPC based on the results of this study which investigated the unfair decisions resulting from the PPSE P121 and P10 formulas and from using a single score for all the candidates who took different field knowledge tests. After calculating the weighted standard scores of the candidates, announcing these scores as an exam result without using a conversion formula can eliminate the inequality between years within the limitations of the classical test theory. If this cannot be done, calculating a separate score for each field can mitigate the inequality, even if it does not fully eliminate it in fields that require teaching field knowledge tests.

Using an exam score in the same application with another exam score without using any equating procedure will cause problems in any case. For this reason, the practice that allows for the validity period of the exams to be longer than one year should be abandoned and this practice should never be utilized again. New calculations can be made using the actual exam data held by MSPC and MoNE, the number of candidates who suffered from the errors in the calculation method can be identified and these grievances can be eliminated by giving them the chance to apply for additional appointments. Similar investigations should also be conducted in regards to university entrance exam where the same or similar formulas are used and for exams such as academic personnel and postgraduate education entrance exam (ALES) which has a validity period of 5 years.

Acknowledgments

A part of this study was presented as an oral presentation at the "National Congress on Measurement and Evaluation Practices in Education" held on 29-30 May 2021 in cooperation with Boğaziçi University and Istanbul Assessment and Evaluation Center.

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. **Ethics Committee Number:** Amasya University Social Sciences Ethics Committee, E-30640013-108.01-1195

ORCID

Ümit Çelen  <https://orcid.org/0000-0001-7014-2221>

5. REFERENCES

- Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich College Publisher.
- Çelen, Ü. (2013). KPSS P10 Hesaplama yöntemine ilişkin sorunlar [Calculation method problems of PPSE P10]. *Ankara University, Journal of Faculty of Educational Sciences*, 46 (1), 127-142. <https://dergipark.org.tr/tr/download/article-file/508720>
- Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Erlbaum Publishers.
- Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory principles and applications*. Kluwer.
- Kan, A. (2010). Test eşitleme: aynı davranışları ölçen, farklı madde formlarına sahip testlerin istatistiksel eşitliğinin sınanması [Test equation: testing statistical equality of tests measuring same behaviors and having different item forms]. *Journal of Measurement and Evaluation in Education and Psychology*, 1(1), 16-21. <https://dergipark.org.tr/tr/download/article-file/65994>

- Ministry of National Education. (2020). https://personel.meb.gov.tr/meb_iys_dosyalar/2020_03/19115547_taban_puan_yeni.pdf
- Measurement, Selection and Placement Center. (1999). *İlk Defa Devlet Memuru Olarak Atanacaklar İçin Seçme Sınavı (DMS) Kılavuzu* [Selection Examination Guide for First Time Civil Servants], MSPC. <https://www.osym.gov.tr/TR,3564/dms-kilavuzu.html>
- Measurement, Selection and Placement Center (2019). *Kamu Personeli Seçme Sınavı (KPSS) Kılavuzu A Grubu ve Öğretmenlik* [Public Personnel Selection Examination (PPSE) Guide Group A and Teaching]: MSPC. <https://dokuman.osym.gov.tr/pdfdokuman/2019/KPSS/kilavuz22072019.pdf>
- Tekin, H. (1996). *Eğitimde Ölçme ve Değerlendirme* (9. Baskı) [Measurement and Evaluation in Education], Yargı Yayınları.
- Turgut, M.F., & Baykul, Y. (2010). *Eğitimde Ölçme ve Değerlendirme* [Measurement and Evaluation in Education], Pegem Akademi.

The Validity and Reliability of the Turkish Version of the Attitudes to Fertility and Childbearing Scale (AFCS)

Sinem Goral Turkcu¹, Sevgi Ozkan², Pinar Sercekus^{1,*}, Erkan Alatas²

¹Pamukkale University, Faculty of Health Sciences, Department of Obstetrics and Gynecology Nursing, Denizli, Turkey

²Pamukkale University, Faculty of Medicine, Denizli, Turkey

ARTICLE HISTORY

Received: July 24, 2020

Revised: Aug. 25, 2021

Accepted: Sep. 22, 2021

Keywords:

Attitude,
Fertility,
Scale,
Validity,
Reliability.

Abstract: This study, a descriptive and methodological type of research, was conducted to evaluate the validity and reliability of the Turkish adaptation of the Attitudes to Fertility and Childbearing Scale (AFCS), developed by Söderberg et al. (2015). The sample of this study consisted of 224 women who had not given birth yet and who were between the ages of 20-30. The scale is a Likert-type measuring instrument consisting of 21 items, in three dimensions. Internal consistency analyses were conducted to determine its reliability. After confirming the linguistic validity, expert opinions were obtained for the content validity. Furthermore, the Item Content Validity Index (I-CVI) and the Scale Content Validity Index (S-CVI) were used to assess its content validity. The construct validity was performed using confirmatory factor analysis. As a result of the confirmatory factor analysis carried out for the construct validity, a three-factor structure of the scale was found to have a good level of model fitness indices (RMSEA=.067, SRMR=.075, CFI=.96). As a result of the scale reliability analysis, the internal consistency coefficient was found to be .82 for the total scale and internal consistency reliability coefficients of the sub-scales were found to be .93 for the "importance of fertility for the future" sub-scale, .87 for the "childbearing as a hindrance at present" sub-scale, and .81 for the "social identity" sub-scale. AFCS is a valid and reliable measurement tool that can be used to measure the fertility and childbearing attitudes of women in a fertile age.

1. INTRODUCTION

As their age increases, women's number of follicles and egg quality decreases, which is called a decrease in fertility, i.e. a decrease in ovarian reserve (number of eggs in the ovaries) (Fitzgerald et al., 1998; Coccia & Rizzello, 2008; Alviggi et al., 2009).

Today, circumstances such as women's desire to improve their level of education, their desire to pursue their career, their desire to reach a certain maturity before having children, their inability to find the right partner, and their thought that their independence will be limited may cause them to delay their first pregnancy (Sleebos, 2003; Tydén et al., 2006; Benzies et al., 2006; Proudfoot et al., 2009; Cooke et al., 2012). The increased maternal age, however, poses

*CONTACT: Pinar Sercekus ✉ pinarsercekus@gmail.com 📧 Pamukkale University, Faculty of Health Sciences, Department of Obstetrics and Gynecology Nursing, Denizli, Turkey

a risk for the health of the mother and her baby and may lead women not to have children at all (Sleeboos, 2003; Tydén et al., 2006; Cooke et al., 2012). In addition, success rates of assisted reproductive techniques used to conceive decrease as the age of the women increases (Yoldemir, 2016).

Although having children may seem to be an obstacle in women's current life, motherhood is important for them in the future (Söderberg et al., 2015). There are very few studies examining the fertility and attitudes towards having children (Söderberg et al., 2013; Söderberg et al., 2015). Söderberg et al. (2013) developed the Attitudes to Fertility and Childbearing Scale (AFCS) using a sample of Swedish women. The AFCS was later revised in a larger sample and reduced from 27 items to 21 items (Söderberg et al., 2015). Söderberg et al. (2015) conducted this scale in young women with a high education level. Similarly, it was thought that it would be appropriate to use the AFCS scale in Turkish young women with a high level of education to determine fertility and childbearing attitudes. Moreover, there is no other tool to measure women's attitudes towards childbearing and fertility in Turkish. The aim of this study is to adapt the AFCS to Turkish and examine the reliability and validity of the Turkish version.

2. METHOD

This study is a descriptive and methodological type of research.

2.1. Study Group

The study population consisted of healthy women who were studying at Pamukkale University, Denizli, Turkey. The sample size in a scale development study is expected to be at least 5-10 times the number of items in the scale (Çapık et al., 2003; Özkan & Sevil, 2007). Since the Attitudes to Fertility and Childbearing Scale consists of 21 items, it was determined that it would be appropriate to include at least 210 women in the sample. The sample of the study consisted of 224 women and these women were in the 20-30 age group, who could read and understand Turkish and who had not yet had children. Women who were not in the 20-30 age group, who had children, who could not read or understand Turkish, and who had a health problem that prevented them from giving birth were excluded from the scope of the research.

2.2. Ethical Aspects of the Study

Permission was obtained from Söderberg to study the Turkish validity and reliability of the AFCS. Ethics committee approval was obtained from Pamukkale University non-interventional clinical research ethics committee. Then, permission from the institution was obtained to be able to carry out the research.

2.3. Data Collection Instruments

A "Personal Information Form" and "AFCS" were used to collect the data for the study.

2.3.1. Personal information form

This form includes questions on age, educational status, marital status, place of residence, use of a method of birth control, working status, and the age range that they plan to become pregnant.

2.3.2. Attitudes to fertility and childbearing scale

AFCS is used to measure attitudes towards having children and fertility in individuals who have not yet had children (Söderberg et al., 2015). In the validity and reliability study of the original scale, the Cronbach alpha coefficients of the subscales were found between .95 and .86. The scale has 3 sub-scales and include the importance of fertility for the future (items no 1, 2, 3, 4, 5, 6, 7), childbearing as a hindrance at present (items no 8, 9, 10, 11, 12, 13, 14, 15, 16), and social identity (items no 17, 18, 19, 20, 21) (Söderberg et al., 2015). The scale is a Likert-type

scale consisting of 21 items and each item is scored over 5 points. On this scale, point 5 shows the optimal, and point 1 shows the weakest attitude. The lowest and highest scores of the scale are 21 and 105, respectively. Low scores reflect low levels of fertility and attitudes to childbearing. The scale development process is given in the title of validity analyses.

2.4. Data Collection Method

The researcher introduced herself before starting the data collection. An introductory information form and a draft scale form were given to the participating women. The participants filled in the scale themselves and the application time of the scale was approximately 3-5 minutes.

2.5. Data Analysis

For validity and reliability analyses, IBM SPSS Statistics v20 was used and for confirmatory factor analysis, Lisrel version 8.8 program was used. Hotelling T2 analysis was conducted to determine whether the mean item scores of all items in the scale and response bias were equal to each other. The floor and ceiling effects were calculated for the whole scale.

2.5.1. Item and reliability analyses

Internal consistency analyses were conducted to determine the reliability of the scale. Item-Total Score Analysis and Pearson Correlation Coefficient were calculated to explain the relationship between the scores obtained from the items in the scale and the total scale score (Table 2). The internal consistency of the scale was calculated using the Composite reliability coefficient, and Cronbach alpha coefficient (Table 3).

2.5.2. Validity analyses

The structure, language, and content validity of the scale were evaluated. The scale was translated into Turkish by two linguists who had good command of both English and Turkish. Independently, the researcher also compared the Turkish versions of the scale. The final version of the scale was translated back into English by two different experts in their fields. The scale translated into English and the original scale were compared. Consequently, it was decided that the translation of the scale was appropriate. Then, Turkish linguists reviewed the conformity of the statements and made the necessary recommendations and redactions. In the final stage, eight experts in the field assessed each item on the scale for theoretical suitability.

Expert opinions were obtained for the content validity of the scale. In addition, the content validity index of the scale was calculated. After the Scale Content Validity Index (S-CVI) and Item Content Validity Index (I-CVI) analyses, which were performed in accordance with expert opinions, a draft scale with 21 items was created. The construct validity, however, was performed using confirmatory factor analysis. Principal axis analysis and varimax rotation were performed for CFA (Table 4).

3. RESULTS / FINDINGS

3.1. Study Sample and Sample Properties

Table 1 shows the sociodemographic characteristics of 224 women in the 20-30 age group. The average age of the women was 21.93 ± 1.74 . Of the women, 96.9% was single, 96% was a high school graduate, 92% was unemployed, 57.1% was living in the city centre, 78.1% was student, 86.2% had social security, and 72.8% had moderate level of income. Of the women, 98.7% was not using a method of birth control. Of the women, 4.9% was of foreign nationality. Of the women, 95.1% was planning to have children in the future and 86.2% was planning to have children between the ages of 25-29.

Table 1. Socio-demographic characteristics of the women (n=224).

Variables	n	%	Variables	n	%
Education			Situation of wanting to have children		
Literate	9	4.0	Yes	213	95.1
High school	215	96.0	No	11	4.9
Marital status			The age she wants to have a child		
Single	7	3.1	20–24 years old	10	4.5
Married	217	96.9	25–29 years old	193	86.2
Job			30–34 years old	19	8.4
Student	175	78.1	35–39 years old	2	.9
Officer	49	21.9			
Income					
Bad	45	20.1			Mean ± Sd
Middle	163	72.8	Age*		21.93 ± 1.74
Good	16	7.1	Number of children she wants *		2.06 ± .40

*Mean ± standard deviations are given.

3.2. Reliability Analysis

3.2.1. Item total score analysis

In the reliability study, item-total score correlations were calculated for the 21-item draft scale. The correlation coefficients of the items varied between .34 and .57 ($p < .000$) (Table 2).

Table 2. Item total score analysis of the scale.

No	Items	Item-Total Correlation	
		<i>r</i>	<i>p</i>
1	I look forward to one day become a mother	.49	<.05
2	I can imagine being pregnant and giving birth	.39	<.05
3	Becoming a mother is important to me	.52	<.05
4	I look forward to being pregnant in the future	.49	<.05
5	Having a child is an essential part of life	.55	<.05
6	It is important for me to be able to get pregnant in the future	.56	<.05
7	Being fertile is an important part of my future life	.53	<.05
8	Having children would limit my leisure time activities	.34	<.05
9	Childbearing does not fit into my life right now	.41	<.05
10	I do not want to take the responsibility as a mother now	.43	<.05
11	An unplanned pregnancy would hinder me in my current life	.50	<.05
12	Having children would limit socializing with my friends	.53	<.05
13	Being a mother would take too much of my own time	.48	<.05
14	Having children would limit my study opportunities	.53	<.05
15	I want to take advantage of my freedom before I have children	.45	<.05
16	Having children would limit my career	.40	<.05
17	Being fertile is important to my feeling of femininity	.57	<.05
18	My fertility makes me feel communion with other women	.51	<.05
19	Becoming a mother is important for my identity as a woman	.52	<.05
20	Being fertile is an important part of my present life	.37	<.05
21	It is important for me to be able to get pregnant any time	.39	<.05

3.2.2. Item total score analysis of the sub-scales

The correlation coefficients between the sub-scale item scores and the sub-scale total scores of the scale were in the range of .76-.90 in the "Factor 1" sub-scale, .56-.79 in the "Factor 2" sub-scale, and .49-.87 in the "Factor 3" sub-scale, respectively and were found to be statistically significant ($p=.000$).

3.2.3. The Scale sub-scales and total score analysis

In order to examine the alignment of each sub-scale with the scale, correlations of the sub-scale scores and the total score of the scale were calculated. The correlation coefficients of the sub-scales were between .60 and .64 and were statistically significant ($p=.000$) (Table 3).

Table 3. Reliability analysis results of the AFCS.

Factors	Sub-Dimension		Cronbach's alpha	Composite Reliability Coefficient	Two half reliability	Guttman Split-half	Spearman Brown	Floor Effect	Ceiling Effect
	Total Score Correlation								
	<i>r</i>	<i>p</i>							
1.Factor (Importance for future)	.60	<.05	.93	.94				.40	20.50
2.Factor (Hindrance at present)	.64	<.05	.87	.80				1.3	4.50
3.Factor (Female identity)	.63	<.05	.81	.83				.40	9.40
Total AFCS			.82		.88	.88	.88	.00	1.30
	Pre-test	Post-test		<i>p</i>					
Total AFCS (Test retest)	76.55±13.18	76.68 ± 11.34		.973					
Hotelling T ²	T ² =570.2, <i>p</i> =.000								

3.2.4. Reliability coefficients

The total Cronbach alpha coefficient of the scale was determined to be .82. Cronbach alpha coefficients for subscales were .93 for "importance of fertility for the future", .87 for "childbearing as a hindrance at present", and .81 for "social identity" (Table 3).

3.2.5. Stability coefficient

To measure the invariance of the scale over time, the test was repeated with 22 women for 15 days after the first application. In the test-retest results, performed to test the relationship between the measurements obtained with a certain time interval and under similar conditions, no significant differences were found between the scores ($p=.973$) (Table 3).

3.2.6. Hotelling's T² test

Hotelling T² analysis was conducted to determine whether the mean item scores of all items in the scale and response bias were equal to each other. It was found that the item averages were different and there was no response bias (Hotelling T²=570.2, $p=.000$) (Table 3).

3.2.7. Ceiling and floor effect of scale

The floor and ceiling effects were calculated for the whole scale. The floor effect of the scale was .00 and the ceiling effect was 1.3. The floor effect of Factor 1 was .40, the ceiling effect

was 20.5, the floor effect of Factor 2 was 1.3, and the ceiling effect was 4.5, while the floor effect of Factor 3 was .40 and the ceiling effect was 9.4 (Table 3).

3.3. Validity Analyses

3.3.1. Linguistic and content validity

After the linguistic validity of the draft scale was ensured, expert opinions of eight experts in their fields were obtained. Eight experts rated each item as '1= not relevant', '2= slightly relevant', '3= highly relevant', and '4= highly relevant'. Then, the experts were asked to give suggestions for responses other than 'highly relevant'. In the expert opinion assessment, all the items were above .78 (I-CVI=.88-1) and the scale validity index was found to be .99. In accordance with the analysis results, no item was removed or changed from the scale. Content validity of the scale was provided by 21 items.

Items not answered by women were identified in the pilot study. After the pilot application, it was decided that data were to be collected through a 21-item draft scale.

3.3.2. Construct validity

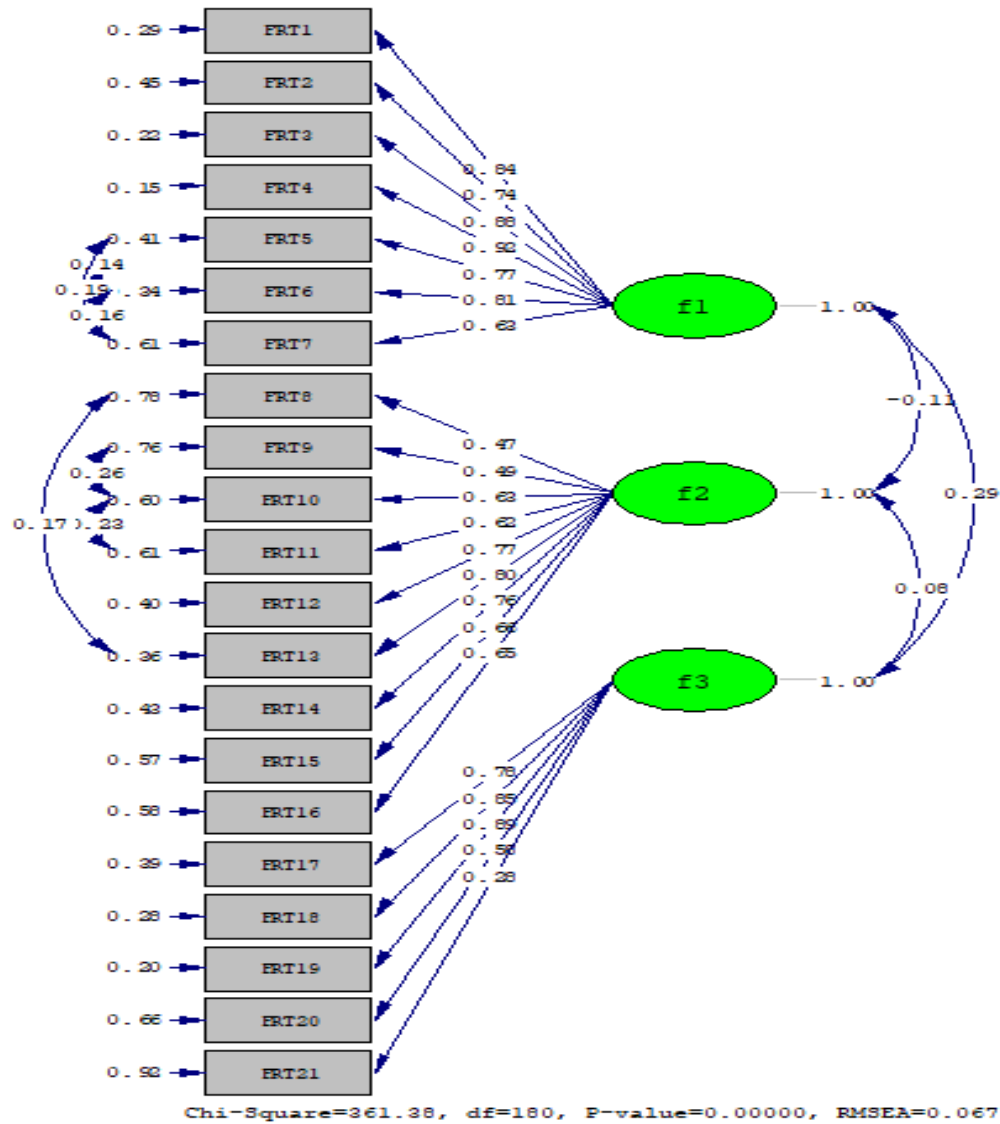
3.3.2.1. Confirmatory Factor Analysis. Model fitness of the AFCS in Turkish culture was investigated by the first level CFA. Scale factor loads were found to be between .28 and .92 as a result of the analysis. Factor loading values of the Attitudes to Fertility and Childbearing Scale were in the range of .63-.92 in the "importance of fertility for the future" sub-scale, .47-.80 in the "childbearing as a hindrance at present" sub-scale, and .28-.89 in the "social identity" sub-scale. Modifications were made among some items in the same sub-dimension. These items are located under the same structure and measure the same value (Figure 1). According to the findings of the confirmatory factor analysis, it was determined that the scale was compatible with the model (Table 4).

Table 4. Findings about first level confirmatory factor analysis.

Fit indices	Values obtained from the scale	Results
χ^2/df	2.00	Good fit
RMSEA	.067	Acceptable fit
SRMR	.075	Acceptable fit
CFI	.96	Good fit
GFI	.87	Acceptable fit
NFI	.93	Acceptable fit
NNFI	.95	Good fit
IFI	.96	Good fit
RFI	.91	Acceptable fit

RMSEA: Root Mean Square Error of Approximation, SRMR: Standardized Root-Mean-Square Residual, CFI: Comparative Fit Index, GFI: Goodness of Fit Index, NFI: Normed Fit Index, NNFI: Non-Normed Fit Index, IFI: Incremental Fit Index, RFI: Relative Fit Index

Figure 1. Confirmatory factor analysis related to AFCS.



4. DISCUSSION and CONCLUSION

This study was conducted to evaluate the validity and reliability of the Turkish adaptation of the AFCS in order to determine the attitudes of women who did not have children about childbearing and fertility.

4.1. Validity Analyses

During the Turkish adaptation of the AFCS, developed by Söderberg et al. (2015), expert opinions were first taken to ensure its linguistic validity. Content validity of the scale was evaluated after linguistic validity was performed. Eight experts were consulted for content validity. The content validity analysis was performed by expert evaluations. If there are six or more experts in the content validity analysis, it is recommended that the I-CVI should not be lower than .78 and the S-CVI should be .90 or higher (Polit & Beck, 2006). As a result of the analysis, I-CVI was above .78 and S-CVI was found to be .99. Thus, content validity of the items in the scale was accepted. According to this result, it was concluded that the scale had sufficient content to identify the attitudes to fertility and childbearing of young women who had no children yet.

As a result of the confirmatory factor analysis, it was revealed that the factor loads of the scale

varied between .28 and .92. In order to look at whether an item is related to the conceptual structure, one needs to look at the factor load of that item. It was stated by Tavşancıl (2010) that factor loads ranging from .30 to .40 can be taken as the lower threshold point. All factor loadings (except for the 21st item) were above .30 (Figure 1). The 21st item is believed to be important for the scale. It was therefore decided to keep it on the scale. Moreover, model fitness indicators, RMSEA=.067, $\chi^2/df=2.00$, SRMR=.075, CFI=.96, GFI=.93, NFI=.95, NNFI=.96, IFI=.87, and RFI=.91 show that the model has a good fit.

In accordance with the statistically significant results, it has been concluded that the scale has the content and construct validity. The reason for the high content and construct validity of the scale is thought to be sufficient language validity and high social adaptation. In addition, the experts whose opinions were obtained included nursing faculty members with many years of experience on the subject. It is believed that obtaining the opinions of appropriate experts on the subject also affected the results.

4.2. Reliability Analyses

The relationship between the total score of the test and the scores of the scale items was determined by the item total-score analysis. Item total score correlations should not be negative and should be above .25 (Kalaycı, 2010). Pearson correlation coefficients of all items in the scale were determined between .34 and .57 by item analysis. The fact that all the items in the scale were greater than .25 correlation value and the analysis results showed that the items were understandable and clear.

The total Cronbach alpha internal consistency reliability coefficient of the scale is .82. As a result of this value, it can be said that the scale has a high reliability (Eser & Baydur, 2007). In the study by Söderberg et al. (2015), the internal consistency coefficients of the sub-scales ranged from .862 to .945. In this study, the Cronbach's alpha internal consistency reliability coefficients of the sub-scales were found to be .93 for the "importance of fertility for the future" sub-scale, .87 for the "childbearing as a hindrance at present" sub-scale, and .81 for the "social identity" sub-scale. The reliability of the scale was also assessed using the two split-half method. According to the split-half test result, the correlation value between the two halves of the scale was found as .88. Based on these results, a strong and significant relationship was found between the two halves. The Guttman split-half and the Spearman-Brown coefficients were $> .88$. The obtained analysis results proved the reliability of the scale as high (Şencan, 2005; Rattray & Jones, 2007; Nunnally & Bernstein, 2010; Çam & Baysan-Arabacı, 2010). As a result, it is seen that internal consistency of the sub-scales and the scale was confirmed.

With the Hotelling T^2 test, bias in responses to the scale items was evaluated. In the Hotelling T^2 test (Hotelling $T^2=570.2$, $p=.000$) item score averages were found to be different. This result shows that the participating women who responded to the scale items were not biased and perceived the items in the same way, which is an important issue that has an impact on the reliability of the scale (Özdamar, 2002; Şencan, 2005). According to these results, it was concluded that women were not biased when filling in the scale.

Determining the floor and ceiling effect of the scale is important in validity and reliability studies, while these values should not exceed 20% (Rattray & Jones, 2007; Şencan, 2005). In this study, it can be said that it is a reliable scale since the floor and ceiling effect of the scale does not exceed 20%.

The AFCS consists of three sub-scales; namely, the "social identity", "childbearing as a hindrance at present" and "importance of fertility for the future" sub-scale. The 21-item Likert-type scale is scored over 5 points. On this scale, point 5 shows the optimal and point 1 shows the weakest attitude. The lowest and highest scores of the scale are 21 and 105, respectively. Low scores reflect low levels of fertility and attitudes to childbearing. As a result, the Turkish

version of AFCS is a valid and reliable measurement tool that can measure the attitudes of young women in the 18-30 age group with no children yet towards childbirth and fertility.

Acknowledgments

We would like to express our gratitude to all participants who participated in this study.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Ethics Committee for Non-Interventional Investigations of the University of Pamukkale, 60116787-020/68045.

Authorship Contribution Statement

Sinem Goral Turkcu: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing the Original Draft. **Sevgi Ozkan:** Methodology, Supervision, and Validation. **Pinar Sercekus:** Methodology, Supervision, and Validation. **Erkan Alatas:** Supervision, and Validation.

ORCID

Sinem Göral Türkcü  <https://orcid.org/0000-0003-1574-0186>

Sevgi Özkan  <https://orcid.org/0000-0001-8385-210X>

Pinar Serçekuş  <https://orcid.org/0000-0002-9326-3453>

Erkan Alataş  <https://orcid.org/0000-0001-6423-5106>

5. REFERENCES

- Alvigi, C., Humaidan, P., Howles, C.M., Tredway D., & Hillier, S.G. (2009). Biological versus chronological ovarian age: implications for assisted reproductive technology. *Reproductive Biology and Endocrinology*, 2009, 7(101), 1-13. <https://doi.org/10.1186/1477-7827-7-101>
- Benzies, K., Tough, S., Tofflemire, K., Frick, C., Faber, A., & Newburn-Cook, C. (2006). Factors influencing women's decisions about timing of motherhood. *Journal of Obstetric, Gynecologic & Neonatal Nursing*, 35(5), 625-633. <https://doi.org/10.1111/j.1552-6909.2006.00079.x>
- Coccia, M.E., & Rizzello, F. (2008). Ovarian reserve. *Annals of the New York Academy of Sciences*, 1127, 27-30. <https://doi.org/10.1196/annals.1434.011>
- Cooke, A., Mills, T.A., & Lavender, T. (2012). Advanced maternal age: delayed childbearing is rarely a conscious choice: a qualitative study of women's views and experiences. *International Journal of Nursing Studies*, 49(2012), 30-39. <https://doi.org/10.1016/j.ijnurstu.2011.07.013>
- Çam, M.O., & Baysan-Arabacı, L. (2010). Qualitative and quantitative steps on attitude scale construction. *Hemşirelikte Araştırma Geliştirme Dergisi*, 12(2), 59-71.
- Çapık, C., Gözüm, S., & Aksayan, S. (2018). Intercultural scale adaptation stages, language and culture adaptation: Updated guideline. *FNJN Florence Nightingale Journal of Nursing*, 26(3), 199-210.
- Eser, E., & Baydur, H. (2007). *Sağlıkla ilgili yaşam kalitesi ölçeklerinin kültürel uyarlaması* [Cultural adaptation of health-related quality of life scales]. 2. Sağlıkta yaşam kalitesi kongresi kongre öncesi kurslar kitabı. <https://www.saykad.net/p/2ulusal-saglikta-yasam-kalitesi-kongresi.html>
- Fitzgerald, C., Zimon, A.E., & Jones, E.E. (1998). Aging and reproductive potential in women. *Yale Journal of Biology and Medicine*, 71(1998), 367-381.

- Kalaycı, Ş. (2010). *Factor analizi: SPSS uygulamalı çok değişkenli istatistik teknikleri* [Factor analysis: SPSS applied multivariate statistical techniques]. Asil Yayın Dağıtım Ltd. Şti.
- Nunnally, J.C., & Bernstein, I.H. (2010). *Psychometric theory*. Mc Graw Hill India.
- Özdamar, K. (2002). *Paket programlar ile istatistiksel veri analizi* [Statistical data analysis with package programs]. Kaan Kitabevi.
- Özkan, S., & Sevil, Ü. (2007) The study of validity and reliability of inventory of functional status after childbirth. *TSK Halk Sağlığı Bülteni*, 6(3), 199-208.
- Polit, D.F., & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489-497. <https://doi.org/10.1002/nur.20147>
- Proudfoot, S., Wellings, K., & Glasier, A. (2009). Analysis why nulliparous women over age 33 wish to use contraception. *Contraception*, 79(2), 98-104. <https://doi.org/10.1016/j.contraception.2008.09.005>
- Rattray, J., & Jones, M.C. (2007). Essential elements of questionnaire design and development. *Journal of Clinical Nursing*, 16, 234-243. <https://doi.org/10.1111/j.1365-2702.2006.01573.x>
- Sleebos, J.E. (2003). *Low fertility rates in OECD countries: facts and policy responses*. In OECD Labour Market and Social Policy Occasional Papers. 2003. <http://ideas.repec.org/p/oec/elsaaa/15-en.html>
- Söderberg, M., Lundgren, I., Christensson, K., & Hildingsson, I. (2013). Attitudes toward fertility and childbearing scale: an assessment of a new instrument for women who are not yet mothers in Sweden. *BMC Pregnancy and Childbirth*, 13, 197. <http://www.biomedcentral.com/1471-2393/13/197>
- Söderberg, M., Christensson, K., Lundgren, I., & Hildingsson, I. (2015). Women's attitudes towards fertility and childbearing - A study based on a national sample of Swedish women validating the Attitudes to Fertility and Childbearing Scale (AFCS). *Sexual & Reproductive Healthcare*, 6(2), 54-58. <https://doi.org/10.1016/j.srhc.2015.01.002>
- Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik* [Reliability and validity in social and behavioral measures]. Seçkin Yayıncılık.
- Tavşancıl, E. (2010). *Tutumların ölçülmesi ve SPSS ile veri analizi* [Measuring attitudes and data analysis with SPSS]. Nobel Yayın Dağıtım.
- Tydén, T., Svanberg, A.S., & Karlström, P.O. (2006). Female university students' attitudes to future motherhood and their understanding about fertility. *The European Journal of Contraception & Reproductive Health Care*, 11(3), 181-189. <https://www.tandfonline.com/loi/iejc20>
- Yoldemir, T. (2016). Fertility in midlife women. *Climacteric*, 19(3), 240-246. <https://doi.org/10.3109/13697137.2016.1164133>

6. APPENDIX

Attitudes to Fertility and Childbearing Scale (AFCS)

Chapter 1 . Fertilite ve Çocuk Doğurmaya Yönelik Tutumlar Ölçeği (FÇDYTÖ)

Sayın Katılımcı,

Bu ölçek, fertilite ve çocuk doğurmaya yönelik tutumları belirlemeye yönelik ifadeleri içeren 21 maddeden oluşmaktadır. Lütfen her maddeyi dikkatlice okuyup 1 ile 5 arası derecelerden birini işaretleyiniz. Katkılarınızdan dolayı teşekkür ederiz.

(1 = tamamen katılmıyorum.....5 =tamamen katılıyorum).

Chapter 2	1	2	3	4	5
GELECEK İÇİN DOĞURGANLIĞIN ÖNEMİ					
1. Bir gün anne olmayı çok istiyorum.					
2. Hamile olduğumu ve çocuk doğurduğumu hayal edebiliyorum.					
3. Anne olmak benim için önemlidir.					
4. Gelecekte hamile kalmayı çok istiyorum.					
5. Çocuk sahibi olmak hayatın önemli bir parçasıdır.					
6. Gelecekte hamile kalabilmek benim için önemlidir.					
7. Doğurgan olmak gelecekteki yaşamımın önemli bir parçasıdır.					
ÇOCUK SAHİBİ OLMANIN GETİREBİLECEĞİ SINIRLAMALAR					
8. Çocuk sahibi olmak boş zaman aktivitelerimi sınırlar.					
9. Çocuk doğurmak şu anki yaşam şeklime uygun değil.					
10. Şu anda anne olmanın sorumluluklarını üstlenmek istemiyorum.					
11. Planlanmamış bir gebelik şu anki yaşamımı zorlaştırır.					
12. Çocuk sahibi olmak arkadaşlarımla olan sosyal yaşamımı sınırlar.					
13. Anne olmak kendime ayıracağım zamanı sınırlar.					
14. Anne olmak öğrenim görme fırsatımı sınırlar.					
15. Çocuk sahibi olmadan önce özgürlüğümün tadını çıkarmak istiyorum.					
16. Çocuk sahibi olmak kariyerimi engeller.					
KADINSAL KİMLİK					
17. Doğurgan olmak kadın olduğumu hissetmem için önemlidir.					
18. Doğurgan olabilmem diğer kadınlar gibi hissetmemi sağlıyor.					
19. Anne olmak, kadınlık kimliğim için önemlidir.					
20. Doğurgan olmak şu anki yaşamımın önemli bir parçasıdır.					
21. İstedğim herhangi bir zamanda hamile kalabilmek benim için önemlidir.					

Development and Evaluation of a Turkish Language Version of the Relational Health Indices

Nesime Can ^{1,*}, Abdulkadir Haktanir ², A. Stephen Lenz ³, Joshua C. Watson ⁴

¹Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Guidance and Psychological Counseling

²Necmettin Erbakan University, Ereğli Faculty of Education, Department of Educational Sciences, Guidance and Psychological Counseling

³Texas A&M University-San Antonio, College of Education and Human Development, Department of Counseling, Health and Kinesiology

⁴Texas A&M University-Corpus Christi, College of Education and Human Development, Department of Counseling and Educational Psychology, Counselor Education

ARTICLE HISTORY

Received: Apr. 19, 2020

Revised: Mar. 23, 2021

Accepted: Aug. 24, 2021

Keywords:

Relational-cultural theory,
Relational health indices,
Confirmatory factor
analysis,
Relationship,
Counseling.

Abstract: Counseling scholars have increasingly utilized the relational-cultural theory (RCT) to promote growth fostering connections as a healthy way of managing various life problems. The Relational Health Indices (RHI) was developed to understand relational interactions among women. In an attempt to broaden the utility of the RHI, the purpose of this study was to develop and validate a Turkish language version of the RHI for research and clinical use. In translating the RHI from English to Turkish, we followed a seven-step process. Data were collected from 213 Turkish-speaking college students enrolled in two Turkish public universities with the mean age of 22.29 (SD= 3.41). The findings revealed that the RHI-T proved to be a two-factor structure (the Peer and Mentor subscales) among Turkish students and that the Community subscale was not an acceptable fit even after removing several items. Potential explanations, implications, and recommendations for clinical use and future research are provided.

1. INTRODUCTION

There is a growing body of literature focusing on the significance of multiculturalism in the field of counseling (Karairmak, 2008; Lam & Yeung, 2017; Zaker & Boostanipoor, 2016). Although multiculturalism is a broad concept and can refer to any particular culture or subculture, scholars view cultures as either individualistic or collectivistic. An individualistic culture refers to a worldview that prioritizes the individual including the individual's goals, uniqueness, and self-control over a group. A collectivistic culture emphasizes we and refers to a worldview in which the social context is centralized, and individuals represent products of their social and cultural context (Hofstede, 2001; Oyserman et al., 2002; Sue & Sue, 2013). Hofstede (2001) indicated that persons from individualistic cultures are likely to have a stronger self-concept and are responsible only for themselves and perhaps their nuclear families. In

*CONTACT: Nesime Can ✉ nesime.can@ankara.edu.tr 📍 Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Program of Guidance and Psychological Counseling, Ankara.

collectivistic cultures, however, others (e.g., peers, community, neighbors) play an essential role in the individual's life. Thus, interpersonal relationships are critical (Hofstede, 2001).

Individualistic and collectivistic cultures are not mutually exclusive concepts, instead they are on a continuum. In other words, culture can reflect both concepts, but usually one is more dominant. Kagitcibasi (2017) studied and compared individualistic and collectivistic cultures and concluded that autonomy and relatedness are basic human needs. However, in individualistic cultures, the need for autonomy is well accepted and supported, while in collectivistic cultures relatedness is supported. In both, one of these needs can be neglected to some extent (Kagitcibasi, 2017).

Hofstede's (2001) study results concluded that individuals from Western cultures reported a higher individualistic worldview, while individuals from Eastern cultures as well as Latino cultures reported a higher collectivistic worldview. After collecting data in 40 countries from 116,000 participants around world, Hofstede (2001) concluded that Turkish people reported higher collectivistic scores than individualistic scores among participants. Another study supporting this finding (Sargut, 2001) investigated general worldview tendencies in Turkey and showed that the average score of the individualistic values was 37%. Thus, collectivistic values were more common among the participants. Based on these studies, one can infer that Turkish culture is predominantly collectivist and that interpersonal relationships are pivotal.

Culturally competent counselors consider their clients' cultural backgrounds and are aware of the elements that are important in their lives (Sue & Sue, 2013). As the cultural and social context has significant implications for individuals' definitions of self, family and relationship dynamics, counselors need to be aware of these dynamics to provide quality care for their clients. In both collectivistic and individualistic cultures, it is essential to understand the meaning of individuals' relationships with others.

1.1. Relational Health

Family structure and relationships are affected by socioeconomic factors and cultural context where individuals live. However, regardless of living in collectivistic or individualistic cultures, increases in perceived availability of social and emotional support were found to be related to decrease in stress, depressive symptoms and also mortality (L'Abate et al., 2010). Brown (2010) noted that people require connection with other people throughout their lives and Relational-Cultural Theory (RCT) stated the significance of growing through and toward connection with development of a healthy "felt of sense" (Frey, 2013, p.178; Jordan, 2017).

As a feminist therapeutic approach, RCT mainly highlights the importance of meaningful connections with others from a multicultural perspective (Frey, 2013). For instance, an individual from an individualistic culture may accept that seeking support is a sign of weakness; as the inherited message of individualistic culture is that one should stand alone and compete (Jordan, 2017). However, RCT highlights mutual empathy which means that when individuals care about each other's well-being and the relationship between them, a growth-fostering relation occurs, and that leads to happiness and overall well-being (Jordan, 2010). Jordan (2010; 2017) also mentions that mutual empathy contributes to zest, clarity, creativity, worth and a desire for more connections. According to RCT, when individuals practice these "five good things", their interactions and connections become stronger (Jordan, 2017, p. 235; Lenz, 2016). These interactions also enhance mutually empathic relations within communities. Researchers stated that individuals with a high levels of relational health may cope with personal and social problems to greater degrees (Lenz et al., 2015).

According to Jean Baker Miller, healing occurs through real connections, and counselor's empathy and understanding lead to therapeutic change (Jordan, 2017). Based on RCT, using empathy, acceptance, and compassionate understanding in counseling help individuals evaluate

impacts of useful connections and disconnections in their lives (Jordan, 2008; Jordan, 2010). Individuals may develop survival strategies to protect themselves from chronic disconnections and counselors should be aware of the signs of these strategies such as invalidation, shaming, anger, rejections (Frey, 2013; Jordan, 2017; Lenz, 2016). Constant disconnections may lead to hopelessness and isolation, and from a broader perspective, racism, homophobia, class prejudice, and sexism; that also create chronic disconnections for individuals and societies (Jordan, 2010; 2017). Therefore, to overcome such problems, RCT has been used in clinical settings with diverse client populations (Crumb & Haskins, 2017; Joe et al., 2020; Singh & Moss, 2016)

RCT approach also has a social justice focus and discusses issues about privilege in counseling, as disconnections can be based on power differences (Comstock et al., 2008). These issues are related to values and biases which can be embedded in both individualistic and collectivistic cultures. Therefore, counselors need to consider these social and political values and work with their clients from a multicultural standpoint (Frey, 2013). Accordingly, it is necessary to understand the concept of relational health for individuals from both individualistic and collectivistic cultures. Although initially, RCT had a focus on the experiences of marginalized women in individualistic cultures, after the development of a measurement tool based on RCT, the concept of relational health has been studied on various topics with various groups around the world (Frey, 2013; Kress, 2018; Lenz, 2016).

1.2. Purpose of the Study

The Relational Health Indices (RHI) is a measurement tool, developed and based on RCT. The RHI measures qualities of growth-fostering relationships with peers, mentors, and community (Liang et al., 2002). As a dynamic construct, the presence of relational health has been studied mostly among women, men, female youth of color, college students, and adolescent girls and boys (e.g., Frey et al., 2005; Haskins & Appling, 2017; Lenz, 2014; Liang et al., 2002; Liang et al., 2007; Liang et al., 2010; Liang & West, 2011, Storlie et al., 2017; Vandermause et al., 2018). Researchers have also studied the construct of relational health with international samples, yet it has been limited to Hispanic/Latino populations (Lenz et al., 2015) and Asians (Liang et al., 2006). Absent in the literature are studies in which the RHI is adapted for use in Turkish samples. Thus, we suggest that the translation and adaptation of the RHI to the Turkish language may help counselors working with individuals from Turkey to better understand their clients' relational health status and its impacts on other aspects of their life. Also, scholars can utilize this instrument in mental health related research.

2. METHOD

We conducted this study with ethical approval of the Institutional Review Board of Texas A&M University-Corpus Christi and then translated the RHI into Turkish. After administering the scales to undergraduate students at Turkish universities, we analyzed the data to assess the psychometric properties and factorial structure of the instrument.

2.1. Participants

After obtaining the IRB approval, we created an online survey link to recruit participants in Turkey. We contacted three faculty members from two different universities and requested to distribute the survey link with undergraduate students. A total of 350 Turkish-speaking undergraduate students enrolled in either a northern or a northwestern university in Turkey participated in the study. Before data analysis, the data-set was inspected for possible entry errors and missing data. After the inspection, we excluded a total of 137 (39%) participants. Overall, of the 213 remaining participants, the sample consisted of 138 female (65%), 72 male (35%) participants - three participants did not answer the demographic query. The mean age of the participants was 22.29 years (SD = 3.41).

2.2. Data Collection Instrument

2.2.1. Relational health indices

The Relational Health Indices (RHI; Liang et al., 2002) was designed to assess the degree to which individuals are engaging in healthy relationships supporting growth with peers, mentors, and their community (Liang et al., 2002). The 37 self-report items in the RHI are presented in regard to a 5-point Likert-type scale with responses ranging through never, seldom, sometimes, often, and always. Higher scores represent a more exceptional relation quality. Additionally, the RHI has cross-scale outputs for authenticity, empowerment/zest, and engagement those measures are drawn from items across the peer, mentor, and community subscales- in a way they are sub-sub-scales. However, they were not a target within our analyses because (a) they are rarely used in the literature based on scores from the RHI (Frey et al., 2005; Liang et al., 2007; Liang et al., 2010) and (b) as indicated in the Liang et al. (2002), their initial factor analyses were not able to represent all the components of RCT theory; thus, it is an incomplete representation. Therefore, in this study, we also decided not to include those sub-sub-scales within our analyses.

The relational quality with peers subscale includes 12-items, and individuals respond to statements such as “I feel understood by my friends” and “My friendship inspires me to seek other friendships like this one.” The 11-item mentor relationship subscale includes statements such as “I can be genuinely myself with my mentor” and “I feel comfortable expressing my deepest concerns to my mentor.” Lastly, the 14-item community relationships subscale includes statements such as “This community has shaped my identity in many ways” and “It seems as if people in this community like me as a person.” The Cronbach’s alpha values for the peer, mentor, and community subscales were .85, .86, and .90, respectively (Liang et al., 2002).

2.3. Translation of the RHI

Considering the guidelines recommended in the instrument translation literature (e.g., Borsa et al., 2012; van Widenfelt et al., 2005; Wild et al., 2005), we utilized a seven-step process for translation of the RHI from English to Turkish. These steps included (a) instrument selection, (b) forward translation of the RHI from English to Turkish, (c) cross-check for the conceptual meaning of translations, (d) backward translation of the RHI from Turkish to English, (e) examining and revising items, and (f) expert review on instrument’s Turkish version, and (g) final review leading to finalizing the Turkish language version of the instrument.

In the first step, once we selected the instrument, the second author, whose native language is Turkish, completed forward translation from English to Turkish. After a cross-check for the conceptual meaning of the translation, the first author and a doctoral student in a counselor education program, both of whose native language is Turkish, received the items for the back translation. The third and fourth authors, both of whose native language is English, compared the back-translation into English with the original English version. Then as a team, we discussed and revised any problematic items and sent the final Turkish version to two Turkish Literature professionals in Turkey. After their review, we made the last changes and finalized the Turkish version of the instrument for use.

2.4. Data Collection Procedure

The first and second authors contacted three faculties from two higher education institutions in Turkey to request help with the dissemination of the study’s survey. After receiving consent from each faculty with an agreement to collaborate letter indicating their willingness to assist with disseminating the survey, we received the Institutional Review Board approval. Using the Qualtrics research software, we created and shared an online survey link with the faculty members who agreed to distribute the online link to their students. The online survey package comprised of an information sheet, a brief demographic questionnaire, and the RHI-T

(Relational Health Indices-Turkish) scale, as well as additional instruments as a part of a broader study, yet irrelevant to this instrument evaluation. Data were collected over five months and then downloaded from Qualtrics and aggregated into an SPSS file, Version 22 (SPSS; IBM Corporation, 2013) for data analysis.

2.5. Data Analysis

2.5.1. Preliminary analysis

Before data analysis, we cleaned the data set by removing participants who completed less than 75% of the questionnaire. Additionally, cases with less than 25% missing values were replaced using the series mean function in SPSS. Conventional person-series mean function is appropriate when data is normally distributed (Lee et al., 2014). Given that we detected no violation of normality in the data, we deemed the series mean function to impute missing values would be feasible.

2.5.2. Primary analyses

We analyzed the RHI-T scale using the original factor structure and also assessed model fit using the AMOS, Version 22. Following the standards developed by Dimitrov (2012), we examined the values of the CMIN/DF, p , root mean residual (RMR), goodness of fit index (GFI), comparative fit index (CFI), Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA) to determine the degree of model fit. Based on these standards, a strong model fit was found in values for the CMIN/DF < 2 , $p > .05$, RMR $< .08$, GFI $> .90$, CFI $> .90$, TLI $> .90$, and RMSEA $< .10$.

In case the model fit was not following the indicated standards, we inspected modification indices to identify items that could have a covaried error term. When potential items were identified, error terms were covaried, and the analysis was conducted again. Model fit indices were also inspected again. If the model still presented an inadequate fit, we examined correlation loadings of individual items and determined if deletion was necessary. We removed the items with less than .70 correlation coefficients. After identifying the final model, Cronbach alpha coefficients for the RHI-T were computed to estimate the internal consistency of the scores.

3. RESULT / FINDINGS

3.1. Peer Subscale

3.1.1. Primary analysis

Although the hypothesized model revealed a significant chi-square value, $X^2(54) = 158.04$, $p < .001$, it was an unacceptable fit for the data, which was also verified by the fit indices, CMIN/DF = 2.92, RMR = .05, GFI = .89, CFI = .82, TLI = .79, RMSEA = .09.

3.1.2. Final model

After pairing error terms for items 10 and 12 (“Arkadaşımın beni olumlu yönde değiştirdiğini hissediyorum [I feel positively changed by my friend]” and “Arkadaşlığım beni olumlu yönde geliştiriyor [My friendship causes me to grow in important]” respectively); 1 and 8 (“Arkadaşımın zor bir durumu paylaşmam gerekirse ona karşı dürüst olabilirim [Even when I have difficult things to share, I can be honest and real with my friend.]” and “Arkadaşım ile en derin duygu ve düşüncelerimi paylaşmaktan rahatsız olurum [I am uncomfortable sharing my deepest feelings and thoughts with my friend.]” respectively); 2 and 3 (“Arkadaşımın sohbet ettikten sonra, moralimin yükseldiğini fark ediyorum [After a conversation with my friend, I feel uplifted]” and “Arkadaşımın zaman geçirdikçe ona daha yakın hissedirim [The more time I spend with my friend, the closer I feel to him/her]” respectively); and removing item 6 (“Arkadaşımın anlaşamadığımız noktaları yargılanıyor hissetmeden konuşabilirim [I can talk

to my friend about our disagreements without feeling judged]”) an acceptable model fit emerged for scores on the Peer Subscale, $X^2(41) = 73.79$, $p < .01$ which was supported by the fit indices, CMIN/DF = 1.80, RMR = .04, GFI = .94, CFI = .94, TLI = .92, RMSEA = .06. Cronbach’s alpha value for the subscale was within the acceptable range of internal consistency ($\alpha = .78$).

3.2. Mentor Subscale

3.2.1. Primary analysis

Though the hypothesized model demonstrated a significant chi-square value, $X^2(44) = 138.29$, $p < .001$, it was a poor fit for the data, which was also confirmed by the fit indices, CMIN/DF = 3.14, RMR = .04, GFI = .88, CFI = .91, TLI = .89, RMSEA = .10.

3.2.1. Final model

After deleting items 5, 7, 8 and 10 (“Akıl hocam sayesinde kendimi daha iyi tanıdığımı düşünüyorum [I feel as though I know myself better because of my mentor]” “Akıl hocamın değerlerini örnek alıp hayatımda uygulamaya çalışırım [örneğin, sosyal, akademik, dini, fiziksel] [I try to emulate the values of my mentor (such as social, academic, religious, physical/athletic)],” “Akıl hocam ile olan ilişkimin enerjimi arttırdığını ve moralimi yükselttiğini hissediyorum [I feel uplifted and energized by interactions with my mentor],” and “Akıl hocam ile olan ilişkim, beni buna benzer ilişkiler bakmaya/aramaya teşvik eder [My relationship with my mentor inspires me to seek other relationships like this one]”, respectively) an acceptable model fit emerged for scores on the mentor subscale, $X^2(14) = 38.39$, $p < .001$ which was confirmed by the fit indices, CMIN/DF = 2.74, RMR < .03, GFI = .95, CFI = .96, TLI = .99, RMSEA = .09. Cronbach’s alpha coefficient for the resulting scores on the subscale was within the good range of internal consistency ($\alpha = .89$).

3.3. Community Subscale

3.3.1. Primary analysis

Even though the hypothesized model showed a chi-square score, $X^2(77) = 292.66$, $p < .001$, it was a poor fit for the data, which was also supported by the fit indices, CMIN/DF = 3.80, RMR = .14, GFI = .81, CFI = .80, TLI = .76, RMSEA = .11.

3.3.2. Final model

After deleting items 3, 6, 7, 8, 12 and 13 (“Eğer içinde bulunduğum topluluktakiler beni rahatsız eden birşeyi biliyorlarsa, benimle konuşurlar [If members of this community know something is bothering me, they ask me about it]” “Bu topluluktakilerle biraraya geldikten sonra, kişisel ilişkiler için harekete geçmem gerektiğini düşünüyorum [I feel mobilized to personal action after meetings within this community]” “Bu topluluktakilerden saklamam gereken yönlerim olduğunu düşünüyorum [There are parts of myself I feel I must hide from this community]” “Bu topluluktakiler beni seviyormuş gibi görünüyor [It seems as if people in this community really like me as a person]” “Bu toplulukla olan bağım beni başka insanlar ile ilişki kurmaya teşvik ediyor [My connections with this community are so inspiring that they motivate me to pursue relationships with other people outside this community]” and “Bu topluluk kişiliğimi birçok açıdan şekillendirdi [This community has shaped my identity in many ways]” from the model due to distinctly low regression coefficients, a poor model fit emerged for scores on the community subscale, $X^2(20) = 167.13$, $p < .001$ as confirmed by inspection of the fit indices, CMIN/DF = 8.35, RMR = .15, GFI = .83, CFI = .81, TLI = .74, RMSEA = .18. Cronbach’s alpha coefficient for the resulting scores on the subscale was within the good range of internal consistency ($\alpha = .84$).

4. DISCUSSION and CONCLUSION

The purpose of this study was to translate the English version of the RHI (Liang et al., 2002) into Turkish and evaluate all three subscales (peer, mentor, and community) in the new version. After analyzing the data gathered from the Turkish college population, our findings indicated that all three subscales had reliable ($\alpha > .70$) scores. In the item analysis process, we paired and removed items to yield acceptable results, as suggested by Dimitrov (2012). However, these modifications only confirmed the peer and mentor subscales, yet the community subscale was unfit. It is important to remember that the RHI was developed with three relationship scales which could be used independently for studying each type of relationship (Liang et al., 2002, p. 27). Therefore, the peer and mentor subscales of the RHI-T can be utilized independently.

This study confirms earlier endeavors to adapt the RHI-T with other diverse groups. For example, in their work of adapting the RHI-T to the Spanish language, Lenz et al. (2015) found the community subscale to be unfit, even after removing four items. We inspected the removed items in other studies to ascertain whether we removed the same items; however, among the deleted items on the community subscale, only one removed item was common between this study and other studies (Lenz et al., 2015; Liang et al., 2007). Interestingly, though, Lenz et al. (2015) and Liang et al. (2007) found the same four items in the community domain to be problematic and deleted them. Additionally, other researchers suggested the community subscale be problematic when it is used as a unitary construct (Frey et al., 2005), thus may be more useful if this subscale was assessed in two domains: alienation from community and connection with community. However, we did not find similar factor loadings as those reported in Frey et al.'s study.

Based on our findings and the extant literature, it is possible that the way some items and the community subscale were constructed is more applicable to Caucasian and/or American groups. Another possible explanation for these results is that the RHI-T was initially developed for marginalized groups (e.g., women). Though our sample consists of predominantly women college students (65%), we also included men college students for which some of the items may not be applicable. Another reason as to why the community subscale came out as unfit may be that in the original items, the word “community” is used. It is worth noting that the semantic usage of the word community is different in Turkish culture. In other words, people do not use the word “topluluk” to refer to their immediate surroundings, as this is implied by “community” in the American culture.

Additionally, “community” is a frequently used word in the American culture; whereas in the Turkish culture, people rarely use the word “topluluk” as part of their daily language. Correspondingly, the word “mentor” translates to Turkish as “akıl hocası.” However, the concept of mentor is not utilized much in the Turkish culture. Therefore, some items related to mentor and community subscales might not have understood by our participants, or they might have had difficulty relating to the items in these subscales.

4.1. Implications, Limitations, and Suggestions for Future Researchers

Counselors and researchers may utilize RHI-T in their work in various ways. College students from Turkey have been moving to different countries to obtain an international college degree. For instance, there are over ten thousand international students from Turkey in the U.S.A (The Institute of International Education, 2018). One should consider that individuals may feel more comfortable being assessed in their native language and that some of the items or subscales (as community subscale in this study) may not apply to these individuals if the original instrument is validated with a different population. Therefore, counselors can use the RHI-T with this population to better understand their relationship with peers and mentors in treatment settings. The results may help clients to acknowledge and be aware of their strengths in relationships

and social support they receive as well as the issues that need to be addressed in counseling. Counselors in Turkey can also utilize the RHI-T for treatment planning and counseling outcome evaluation. This will empower counselors through expanding their toolbox and enhance their effectiveness by using empirically validated theory-based instruments.

The present study is not free of its limitations. Steven (2009) recommends that in confirmatory factor analysis, ten participants per item should be recruited. However, the sample size of this current study was below the recommended threshold. We recommend that future researchers replicate this study with larger groups. Additionally, future researchers may utilize more heterogeneous samples representing various age groups and individuals from different educational backgrounds. For instance, including adults both with and without a higher education degree may provide more inclusive results as this would be a better representation of Turkey population.

Scholars develop and study theories, and with the help of theory-based assessment tools, researchers have a practical tool to assess the construct of relational health with Turkish speaking individuals. The RHI-T may help researchers and scholars to understand the implication of RCT to a collectivistic culture better. Considering that we had to remove several items and that the community subscale was not a fit even after removing the items, we suggest that future researchers may develop an instrument based on RCT that may be more relevant to the Turkish culture.

4.1. Conclusion

Despite its limitations, this study presents valuable information of the RHI-T for counselors striving to utilize this instrument to understand the relational health of Turkish college students. Our results proved that two sub-dimensions of the RHI-T can be used to measure growth fostering relationships with peers and mentors. The constructs mentorship and community evoke different concepts in Turkish and American cultures. Finally, counselors can utilize the RHI-T with Turkish speaking college students for understanding the degree of the client's growth-fostering relationships, while researchers use it in relational health related studies.

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

Nesime Can: Collected data, drafted and revised the manuscript. **Abdulkadir Haktanir:** Collected data and drafted and revised the manuscript. **A. Stephen Lenz:** Provided supervision and performed the analysis. **Joshua C. Watson:** Provided supervision and consultation. All authors contributed to the process of translation of the instrument.

ORCID

Nesime Can  <https://orcid.org/0000-0002-6448-1275>

Abdulkadir Haktanir  <https://orcid.org/0000-0003-4066-0276>

A. Stephen Lenz  <https://orcid.org/0000-0002-3360-835X>

Joshua C. Watson  <https://orcid.org/0000-0002-3011-3941>

5. REFERENCES

- Borsa, J. C., Domásia, B. F., & Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psychological instruments: Some considerations. *Paidéia*, 22, 423-432.
- Brene, B. (2012). *Daring greatly: How the courage to be vulnerable transforms the way we live, love, parent, and lead*. Penguin Group.

- Comstock, D. L., Hammer, T. R., Strentzsch, J., Cannon, K., Parsons, J., & Salazar, G. (2008). Relational-cultural theory: A framework for bridging relational, multicultural, and social justice competencies. *Journal of Counseling & Development, 86*, 279–287.
- Crumb, L., & Haskins, N. (2017). An integrative approach: Relational cultural theory and cognitive behavior therapy in college counseling. *Journal of College Counseling, 20*(3), 263–277. <https://doi.org/10.1002/jocc.12074>
- Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: American Counseling Association.
- Frey, L. L. (2013). Relational-cultural therapy: Theory, research, and application to counseling competencies. *Professional Psychology: Research and Practice, 44*, 177-185. <https://doi.org/10.1037/a0033121>
- Frey, L., Beesley, D., & Newman, J. (2005). The Relational Health Indices: Reanalysis of a measure of relational quality. *Measurement and Evaluation in Counseling and Development, 38*, 153–163.
- Haskins, N. H., & Appling, B. (2017). Relational-cultural theory and reality therapy: A Culturally responsive integrative framework. *Journal of Counseling & Development, 95*(1), 87–99. <https://doi.org/10.1002/jcad.12120>
- Hofstede, G. (2001). *Culture's consequences? Comparing values, behaviors, institutions, and organizations across nations*, 2nd Ed. Sage Publications.
- Institute of International Education. (2018). *Open doors 2018 fast facts*. Retrieved from <https://theoxfordconclave.org/wp-content/uploads/2019/09/Fast-Facts-2018.pdf>
- Jordan, J. (2008). Recent developments in relational-cultural theory. *Women & Therapy: A Feminist Quarterly, 31*, 1–4. <https://doi.org/10.1080/02703140802145540>
- Joe, J. R., Norman, A. R., Brown, S., & Diaz, J. (2020). The intersection of HIV and intimate partner violence: An application of relational-cultural theory with black and Latina women. *Journal of Mental Health Counseling, 42*(1), 32. <https://doi.org/10.17744/mehc.42.1.03>
- Jordan, J. (2010). *Relational-cultural therapy*. American Psychological Association.
- Jordan, V. J. (2017). Relational-cultural theory: The power of connection to transform our lives. *Journal of Humanistic Counseling, 56*, 228-243.
- Karairmak, Ö. (2008). Multiculturalism, cultural sensitivity and counseling. *Türk Psikolojik Danışma ve Rehberlik Derneği, 3*(29), 115-129.
- Kress, V. E., Haiyasoso, M., Zoldan, C. A., Headley, J. A., & Trepal, H. (2018). The use of relational-cultural theory in counseling clients who have traumatic stress disorders. *Journal of Counseling & Development, 96*(1), 106-114. <https://doi.org/10.1002/jcad.12182>
- Kagitcibasi, C. (2017). Doing psychology with a cultural lens: A half-century journey. *Perspectives on Psychological Science, 12*, 824-832. <https://doi.org/10.1177/1745691617700932>
- Lam, G., & Yeung, M. (2017). The Cultural Obstacles of Counseling Licensure in Hong Kong. *College Student Journal, 51*(2), 193–201.
- L'Abate, L., Cusinato, M., Maino, E., Colesso, W., Scilletta C. (2010). *Relational competence theory: Research and mental health applications*. Springer.
- Lee, M. R., Bartholow, B. D., McCarthy, D. M., Pedersen, S. L., & Sher, K. J. (2014). Two alternative approaches to conventional person-mean imputation scoring of the self-rating of the effects of alcohol scale (SRE). *Psychology of Addictive Behaviors, 29*(1), 231-236. <https://doi.org/10.1037/adb0000015>
- Lenz, A. S. (2014). Mediating effects of relationships with mentors on college adjustment. *Journal of College Counseling, 17*, 195-207. <https://doi.org/10.1002/j.2161-1882.2014.00057.x>

- Lenz, A. S. (2016). Relational-cultural theory: Fostering the growth of a paradigm through empirical research. *Journal of Counseling and Development, 94*, 415-428.
- Lenz, A. S., Holman, R. L., Lancaster, C., & Gotay, S. G. (2016). Effects of relational authenticity on adjustment to college. *Journal of College Counseling, 19*(1), 2-16. <https://doi.org/10.1002/jocc.12027>
- Liang, B., Tracy, A., Glenn, C., Burns, S. M., & Ting, D. (2007). The Relational Health Indices: Confirming factor structure for use with men. *The Australian Community Psychologist, 19*, 35-52.
- Liang, B., Tracy, A., Kenny, M. E., Brogan, D., & Gatha, R. (2010). The Relational Health Indices for youth: An examination of reliability and validity aspects. *Measurement and Evaluation in Counseling and Development, 42*, 255-274. <https://doi.org/10.1177/0748175609354596>
- Liang, B., Tracy, A., Kauh, T., Taylor, C., & Williams, L. M. (2006). Mentoring Asian and Euro-American College Women. *Journal of Multicultural Counseling & Development, 34*(3), 143-154. <https://doi.org/10.1002/j.2161-1912.2006.tb00034.x>
- Liang, B., Tracy, A., Taylor, C. A., Williams, L. M., Jordan, J. V., & Miller, J. B. (2002). The Relational Health Indices: A study of women's relationships. *Psychology of Women Quarterly, 26*, 25-35. <https://doi.org/10.1111/1471-6402.00040>
- Liang, B., & West, J. (2011). Relational health, alexithymia, and psychological distress in college women: Testing a mediator model. *Journal of Orthopsychiatry, 81*, 246-254. <https://doi.org/10.1111/j.1939-0025.2011.01093.x>
- Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychology Bulletin, 128*, 3-72. <https://doi.org/10.1037//0033-2909.128.1.3>
- Sargut, A. S. (2001). *Kültürler arası farklılaşma ve yönetim [Cross-cultural differentiation and management]*. İmge Kitabevi.
- Singh, A. A., & Moss, L. (2016). Using Relational-Cultural Theory in LGBTQQ counseling: Addressing heterosexism and enhancing relational competencies. *Journal of Counseling & Development, 94*(4), 398-404. <https://doi.org/10.1002/jcad.12098>
- Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Routledge.
- Storlie, A. C., Albritton, K., & Cureton, J. L. (2017). Familial and social influences in career exploration for female youth of color: A study of relational cultural theory. *The Family Journal: Counseling and Therapy for Couples and Families, 25*, 351-358. <https://doi.org/10.1177/1066480717732142>
- Sue, D. W., & Sue, D. (2013). *Counseling the culturally diverse: Theory and practice*. Hoboken, John Wiley & Sons, Inc.
- van Widenfelt, B. M., Treffers, P. D. A., de Beurs, E., Siebelink, B. M., & Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clinical Child and Family Psychology Review, 8*, 135-145.
- Vandermause, R., Roberts, M. & Odom-Maryon, T. (2018). Relational health in transitions: Female adolescents in chemical dependency treatment. *Substance Use & Misuse, 53*, 1353-1360. <https://doi.org/10.1080/10826084.2017.1408655>
- Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patients-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health, 8*, 94-104.
- Zaker, B. S. & Boostanipoor, A. (2016). Multiculturalism in counseling and therapy: Marriage and family issues. *International Journal of Psychology and Counselling, 8*(5), 53-57. <https://doi.org/10.5897/IJPC2016.0388>

Performance and Differences in Grading Practices Among Undergraduates at Business Schools

Leiv Opstad ^{1,*}

¹Norwegian University of Science and Technology, NTNU Business School, NO-7491 Trondheim, Norway

ARTICLE HISTORY

Received: Mar. 24, 2021

Revised: Sep. 15, 2021

Accepted: Sep. 22, 2021

Keywords:

Grading standard,
Business School,
Performance,
Mathematical skills,
Gender.

Abstract: If the ranking of students is based on grade scores independent of the selected college or university, it is critical to have an equal national measurement standard. It is a challenge to ensure this if there is a substantial difference in the composition of the students and enrolment requirements among colleges. Based on three different types of colleges in Norway merged into one unit in 2019, this paper examines the grading practices before and after the fusion. By using a regression model to predict the grade depending upon students' academic skills, one can identify different grading practices for the three independent schools and compare the results after they become one unit with identical exams and a common evaluation. The results show significantly more lenient grading practices at small colleges with low entry criteria and that the evaluation is more random, depending upon the instructor. Furthermore, this paper confirms that the grade point average (GPA) from upper secondary school, mathematical abilities and gender are strongly correlated with success in business studies.

1. INTRODUCTION

The purpose of this paper is to analyse whether there are different grading practices among schools within business education in Norway and to examine which factors can explain students' performance in business courses. Some countries rank students depending on which institution they have attended. Norway have chosen a different approach. For the same course, there should be identical evaluation regardless of which university or college the student is graduated from. It is challenging to secure such a scheme. We'll take a closer look at this issue in this article. The research suggests that there are significant differences internationally in grading practice (Broockhart et al., 2016).

The grading system among undergraduates is critical since the ranking of students depends on the grade scores. A grading design that measures students' knowledge and skills would be a good tool for ranking in future studies and would provide the desired information to future employers. Independently of the institution, it is assumed that two students who achieve the same level of result should receive identical grades. The students are ranked according to their grade scores, independently of which college or university they have attended. To ensure this,

*CONTACT: Leiv OPSTAD ✉ leiv.opstad@ntnu.no 📍 Norwegian University of Science and Technology, NTNU Business School, NO-7491 Trondheim, Norway

the Norwegian Ministry of Education has used substantial resources to develop identical grading habits all over the country. All colleges and universities should have the same grading evaluation by following the ECTS (European Credit Transfer and Accumulation System) grading scale system (see Table 1). Two students having the same grade shall perform equally, independently of the chosen college. Grade C shall have the same meaning for the same business course regardless of which school the student has attended and regardless of the student's abilities and academic skills.

It is difficult to justify a system in which students are treated differently depending on background and school and get different grades even there is an equal performance, as long as there is no ranking of the schools. Hence, the grading standard system should compensate for differences across institutions provided that there is no ranking based upon which school the student has attended. The distribution of grades is at the national level (Table 1). The grade for the mean student will vary depending on the performance of the students at the different schools. Hence, the grading system does not give the correct information about the candidates' qualifications. Without further knowledge of the different education institutions the employers do not have the correct qualifications. This can lead to a principal agent problem with asymmetric information.

Table 1. *The Grading System (on National level).*

Grade	Per-cent	Description	General, qualitative description of the evaluation criteria (see: https://www.ntnu.edu/studies/grading)
F		Fail	A performance that does not meet the minimum academic criteria. The candidate demonstrates an absence of both judgement and independent thinking.
E	10	Sufficient	A performance that meets the minimum criteria but no more. The candidate demonstrates a very limited degree of judgement and independent thinking.
D	25	Satisfactory	A satisfactory performance but with significant shortcomings. The candidate demonstrates a limited degree of judgement and independent thinking.
C	30	Good	A good performance in most areas. The candidate demonstrates a reasonable degree of judgement and independent thinking in the most important areas.
B	25	Very good	A very good performance. The candidate demonstrates sound judgement and a very good degree of independent thinking
A	10	Excellent	An excellent performance, clearly outstanding. The candidate demonstrates excellent judgement and a high degree of independent thinking.

1.1. Factors Behind Students' Success in Business Courses

Academic ability is a key factor for success in higher education. A variable that can encapsulate this dimension is grade point average (GPA) scores from upper secondary schools. According to Grove et al. (2006), GPA scores are a proxy estimate of students' academic aptitude in economic education. Many studies have found a positive correlation between GPA and achievement in business studies (Jones et al., 2013; Opstad & Årethun, 2020a). Uyar and Güngörmüş (2011) reported that GPA from upper secondary school was the strongest predictor of success in finance and accounting courses. In comparing GPA with attendance in the courses, the authors observed only a weak positive link between attendance and performance. The association between GPA and success is also substantial and positive for marketing courses

(Marcal et al., 2005). Students with high GPAs also have success in introductory management courses (Brookshire & Palocsay, 2004).

Another key determinant for success in business courses is mathematical skills (Blaylock & Lacewell, 2008; Opstad, 2018). Mathematics is a valuable factor for doing analyses in quantitative courses. Hence, there is a strong significant connection between mathematical abilities and performance in quantitative business courses (Ballard & Johnson, 2004; Mallik & Lodewijks, 2010; Opstad & Årethun, 2019; Uyar & Güngörmüş, 2011). Students' mathematical background seems to be a crucial instrument for managing these subjects. Alcock et al. (2008) and Opstad (2018) also found a positive correlation between mathematical knowledge and performance in non-quantitative business courses like management and marketing even though there are no mathematical tools in presentations in these fields. The reason might be that mathematical strength improves the students' ability to analyse and develop a good structure in their performance in non-analytical courses; however, the mathematical background is not as essential as in quantitative courses (Alcock et al., 2008).

Gender also matters in business courses. Krishna and Orhun (2020) found that females have less success in the quantitative courses, even the female students who have improved their performance over the past years. On the other hand, females seem to perform better in non-quantitative courses (Volchok, 2019). Opstad and Årethun (2020b) observed that women got significant higher scores in marketing course than men.

Other factors linked to performance in business courses are personal characteristics and students' effort. Numerous articles have studied the connection between personal traits and academic success (Trapman et al., 2007). Study effort correlates positively to achievements in business courses (Bonesrønning & Opstad, 2012; Stinebrickner & Stinebrickner, 2008).

Teacher quality seems to be strongly associated with student success. Darling-Hammond (2000) reported a strong positive connection between instructors' skills and student achievement. Odden et al. (2004) found that teachers' qualifications have an impact on students' learning and performance. Other studies have also confirmed this result. The instructors matter, but the influence on student success varies (Bardach & Klassen, 2020).

1.2. What Can Explain Different Grading Practices

In Norway, essays and constructed response questions comprise a considerable part of the final exams. This is in contrast to American colleges where a multiple-choice question format is the main style of examination (Walstad & Miller, 2016). Essay questions make it difficult to ensure equal grading practices across institutions, and many countries experience these same challenges. According to Beenstock and Feldman (2018), differential grading across colleges seems to be the norm rather than the exception.

Admission standards play a critical role to ensure high standards in study programmes (Lawrence & Pharr, 2003). Even if there are similar school enrolment criteria, there might be substantial differences in the qualifications of the students. Some colleges welcome all qualified applications, while others only accept students with high GPAs since there are considerably more applicants than places available. Hence, it is a competition among the students to get an offer of acceptance. Furthermore, an academic school with a good reputation attracts qualified students (Mayer-Foulkes, 2002). Therefore, the differences in enrolment qualifications might remain or even be stronger over time. Well qualified students tend to apply to colleges which accept only applications with good academic skills and high entrance scores. This might cause a bias against programmes which are not attractive and include academically weaker students (Godor, 2017). One might be tempted to give higher grades compared to schools with more popular programmes. Marini et al. (2018) found a substantial difference in grading practices in certain disciplines, depending upon the qualification of the students. According to Godor (2017)

and Opstad (2020), the result can be different grading considerations across the institutes and programmes at the same faculty, depending upon the composition of the students. Academically stronger students tend to get a more stringent grading evaluation.

The role of teachers is crucial for the grading standard (Cheng et al., 2020; Godor, 2017). The instructors' considerations of students' contributions and performance are essential for the students' grades. Bonesrønning (1999) reports that the grading practice of the lecturers are directly linked to their characteristics, such as preferences, attitudes, gender, age, skills and education. Even if the goal is to have an equal grading system, different kinds of teachers can result in differing grading standards.

The effect of a national norm of grading practices implies that courses with students with poor academic skills will result in rather low average grades. Having such students, both the instructor and the school can feel uncomfortable with such an outcome. One can therefore be tempted to follow more or less the composition of the ECTS grading system locally and disregard or not focus on the differences in students' entrance qualifications, and thus one gets less strict in the grading practices. More students achieve the letter grade A or B than they deserve according to the national schedule (see Table 1), which makes more students happy. Those schools can attract students who struggle to achieve enough points required for further studies. Lenient grading practices can also improve instructor ratings and student evaluations of their teacher (Hoefer et al., 2012). Faculty deans can use the grades as a proxy for the instructors' teaching abilities. Also, due to the financial system, colleges involving academically weaker students can be interested in rewarding students with better grades than the national norm. The colleges receive revenue from the Ministry of Education depending on how many students pass their courses. Therefore, the instructor, faculty dean and the administration can benefit from having a less stringent grading standard if the average student has rather low enrollment scores. This can lead to a misallocation of public funds (Bagues et al., 2008).

1.3. Hypotheses

Based on the previous research, we postulated the following hypotheses:

H1: There is a link between performance in business and the students' mathematical background, academic skills and gender.

H2: There are grading differences for equally qualified students among business schools in Norway.

Even though the goal is to have a national standard, local variations can cause diversity in the grading practices

2. METHOD

The data consisted of information about individual background variables and performance in four business courses at three different schools for 3 years: 2016, 2017 and 2019. Earlier on, there were three independent business institutions, but in 2018, they united into one into join school. We will focus on students' outcome before and after the merger was completed. Furthermore, in 2019, an identical design was introduced for all courses with a common exam with same examiners. This makes it possible to compare the results before and after the fusion of the schools. Table 2 presents the data for 2016 and 2017. Since there are clear national standards on the content of the various courses, then there is little variation among the different schools. The pedagogical arrangement is quite identical across the institutions. The exam form is a 4 hour written exam based on response questions. On the other hand, it is the local lecturer who design the exam thesis and conducts the examination. But the rules are that a student should receive an identical grade with the same contribution regardless of institution. Data in this

analysis make it possible to consider it one has managed to have such a grading system. This is administrative data and includes all students who took the exam in the actual courses.

Table 2. Descriptive Statistics.

	All		School A		School B		School C	
	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
Percent S-maths	27.35%		34.73%		21.33%		9.6%	
Percent N-maths	16.70%		22.12%		9.33%		6.21%	
Percent males	46.80%		47.53%		40.00%		50.28%	
GPA Upper Secondary School ¹	4.46	0.52	4.71	0.31	4.06	0.45	3.95	0.60
Microeconomics ²	2.58	1.53	2.97	1.41	1.26	1.51	2.50	1.55
Macroeconomics ²	2.75	1.48	2.87	1.37	2.67	1.72	2.39	1.50
Management ²	3.08	1.17	2.95	1.24	3.03	1.17	3.62	0.69
Marketing ²	2.83	1.27	2.77	1.23	2.73	1.26	3.52	1.29
N ³	860		547		136		177	

Notes. ¹⁾ The grades are from 1 to 6.
²⁾ Mean letter grade (0: F, 1: E, 2: D, 3: C, 4: B, 5: A)
³⁾ The numbers vary depending on the subject.

There are substantial differences between the three schools. School A is located in a rather big city where there is competition among students to gain access to the courses. Therefore, institution A has higher entrance requirement. The GPA from upper secondary school is therefore higher for business school A than for B and C (about three quarters higher) and the variations are also smaller (standard deviation for A is only 0.31 compared with 0.45 and 0.60 for the two others).

Students at upper secondary school can choose between three pathways in mathematics: practical mathematics (P-maths), mathematics for business and social science (S-maths) or mathematics for natural science (N-maths). P-mathematics is practical. The subjects in S-mathematics contain functions, algebra and regression models. N-mathematics is most advanced and theoretical and includes issues like geometry and vectors. The students attending school A have considerably stronger skills in mathematics than those attending B or C. Notice also the variations in scores depending on the institution. Students from A outperform the other institutions in the quantitative courses (macroeconomics and microeconomics), while they tend to underperform in the non-quantitative courses. The mean student from A has the lowest score in management and about the same level as school B in marketing, while the average score for school C is almost one letter grade higher.

2.1. The Model

By using linear regression, we will analyse how the performances at each school are associated with explained variables of GPA, gender and mathematical skills (Model 1):

$$Y_{ij} = a_0 + a_1X_1 + a_2 X_2 + a_3X_3 + a_4X_4 + \varepsilon$$

where,

Y_{ij} : grades in business course i , institution j ;

a_0 : constant; X_1 : gender (0: F, 1: M); X_2 : GPA from secondary upper school;

X_3 : dummy variable for having chosen N-mathematics (0: not chosen, 1: chosen);

X_4 : dummy variable for having chosen S- mathematics (0: not chosen, 1: chosen)

The literature indicates that GPA from secondary upper school, gender and mathematical skills affect students' performance in business courses. Hence, they are chosen as independent variables.

In Model 2 we will use the result to predict performances in the different schools by using dummy variables. The applied Model 2 is:

$$Y_i = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + a_6X_6 + \varepsilon$$

where,

Y_i : grades in business course i . (0: F, 1: E, 2: D, 3: C, 4: B, 5: A);

a_0 : constant; X_1 : gender (0: F, 1: M); X_2 : GPA from secondary upper school;

X_3 : dummy variable for having chosen N-mathematics (0: not chosen, 1: chosen);

X_4 : dummy variable for having chosen S- mathematics (0: not chosen, 1: chosen);

X_5 : dummy variable for belonging to school B (0: not institution B, 1: school B);

X_6 : dummy variable for belonging to school C (0: not school C, 1: school C);

ε : stochastic error

To avoid multicollinearity, no dummy variables for school A were included in the regression model, and this group will belong to the reference category.

3. RESULT / FINDINGS

3.1. Result from the Regression Model (Tables 3-6)

A comparison of the three institutions shows many similarities, such as there is a considerable correlation between GPA score and success regardless of the school. The findings confirm hypothesis 1 (H1). There is a strong correlation between GPA from upper secondary school and outcomes for all courses, but the impact is strongest for the quantitative courses. Comparing microeconomics and marketing, the influence is about twice as big for microeconomics (Model 2). The values depend on the schools. For microeconomics, the unstandardised beta score is 1.37 for school A and around 1.0 for the other institutions.

Therefore, mathematical background from secondary upper school is a good predictor of performance in business courses. Mathematical skills are related to performance and especially in microeconomics ($\beta = 0.54$ for S-maths and 0.44 for N-maths, Model 2). There is also a significant link between mathematical skills and success in marketing and management but a lower impact (β values are between 0.16 and 0.35, and they are higher for N-mathematics than for S-mathematics).

The gender influence is in favour of males in macroeconomics and microeconomics, but the result is opposite for marketing and management (Model 2). Male students underperform compared with female peers.

There are also considerable variations in significance levels and impacts among the schools. For instance, in macroeconomics, N-mathematics is strong related to the success of institution C ($\beta = 1.98$) but negative and with no significant effect for institution A. There is no significant gender effect for B and C but a strong and significant one for A. Looking at marketing, there is no association between mathematical background and scores for institution B. N-mathematics has a significant impact for institution A but not for institution C. For S-mathematics, the opposite is the case. There is a high beta score ($\beta = 0.71$) and significance at the 10 percent level for C but a weak beta score and no significant impact for A. Some of the same differences occur for microeconomics and management.

Table 3. Performance in Macroeconomics (unstandardised β values, standard deviation in parenthesis).

	School A		School B		School C		All (Model 2)	
	β	Sig.	β	Sig.	β	Sig.	β	Sig.
Constant	-22.25 (0.97)	.021 .	.489 (1.208)	.687	-.198 (1.15)	.864	-1.41 (0.68)	.037
Gender	.78 (0.13)	.000	.436 (0.30)	.142	.077 (0.31)	.802	.64 (0.11)	.000
S-maths	.22 (0.14)	.127	.577 (0.34)	.091	1.185 (0.44)	.009	.37 (0.13)	.005
N-maths	-.001 (0.16)	.993	.133 (0.13)	.809	1.981 (0.64)	.003	.16 (0.16)	.316
GPA	1.00 (0.20)	.000	.478 (0.29)	.101	.633 (0.29)	.031	.81 (0.14)	.000
Dummy B							.50 (0.17)	.003
Dummy C							.45 (0.21)	.037
	N = 437, Adj. R ² = 0.06		N = 133, Adj. R ² = 0.037		N = 94 Adj. R ² = 0.176		N = 666, Adj. R ² = .0094	

Table 4. Performance in Microeconomics (unstandardised β values, standard deviation in parentheses).

	School A		School B		School C		All (Model 2)	
	β	Sig.	β	Sig.	β	Sig.	β	Sig.
Constant	-3.85 (0.94)	.000	-1.13 (1.19)	.343	-1.19 (0.90)	.186	-2.25 (0.61)	.000
Gender	.140 (0.12)	.255	.592 (0.26)	.026	.00 (0.25)	.999	.18 (0.10)	.067
S-maths	.58 (0.14)	.000	.194 (0.31)	.528	.96 (0.38)	.012	.54 (0.19)	.000
N-maths	.42	.008	.570 (0.43)	.189	1.12 (0.48)	.020	.44 (0.14)	.002
GPA	1.37 (0.20)	.000	.495 (0.29)	.088	.89	.000	1.03 (0.13)	.000
Dummy B							-.95 (0.16)	.000
Dummy C							.51 (0.17)	.002
	N = 511, Adj. R ² = 0.103		N = 135, Adj. R ² = 0.048		N = 94, Adj. R ² = 0.195		N = 666, Adj. R ² = 0.251	

Table 5. Performance in Marketing (unstandardised β values, standard deviation in parentheses).

	School A		School B		School C		All (Model 2)	
	β	Sig.	β	Sig.	β	Sig.	β	Sig.
Constant	.4574 (0.74)	.542	.26 (0.66)	.690	1.12 (0.79)	.158	.17 (0.46)	.712
Gender	-.24 (0.10)	.015	-.41 (0.15)	.005	-.070 (0.24)	.769	-.27 (0.08)	.001
S-maths	.15 (0.12)	.225	-.02 (0.25)	.950	.71 (0.41)	.086	.13 (0.11)	.203
N-maths	.33 (0.11)	.004	.01 (0.19)	.968	.25 (0.39)	.527	.26 (0.09)	.006
GPA	.50 (0.15)	.001	.63 (0.15)	.000	.59 (0.19)	.002	.56 (0.09)	.000
Dummy B							.26 (0.11)	.015
Dummy C							1.13 (0.15)	.000
	N = 615, Adj. R ² = 0.039		N = 284, Adj. R ² = 0.97		N = 129 Adj. R ² = 0.106		N = 666, Adj. R ² = 0.078	

Table 6. Performance in Management (unstandardised β values, standard deviation in parentheses).

	School A		School B		School C		All (Model 2)	
	β	Sig.	β	Sig.	β	Sig.	β	Sig.
Constant	-.31	-.31	.80 (0.59)	.179	2.57 (0.40)	.000	.59 (0.41)	.147
Gender	-2.28	-2.28	-.48 (0.12)	.000	.14 (0.12)	.245	-.27 (0.70)	.000
S-maths	1.10	1.10	.30 (0.21)	.152	.22 (0.17)	.176	.16 (0.09)	.091
N-maths	3.19	3.19	.415 (0.16)	.010	-.21 (0.18)	.251	.35 (0.09)	.000
GPA	4.46	4.46	.56 (0.14)	.000	.26 (0.10)	.008	.51 (0.08)	.000
Dummy B							.360 (0.09)	.000
Dummy C							1.14 (0.13)	.000
	N = 592, Adj. R ² = 0.057		N = 329, Adj. R ² = 0.115		N = 136, Adj. R ² = 0.06		N = 1059, Adj. R ² = 0.102	

The dummy variables are an indicator of grade standard differences in college B and C compared to A, adjusted for gender and enrolment qualifications (GPA and mathematical backgrounds). Using this method, there seems to be substantial differences in grading practices. Students of the same gender and with the same entrance qualifications receive at institution C at least one letter grade better in in management ($\beta = 1.14$) and marketing ($\beta = 1.13$) than at institution A. For institution B, the difference is much lower for these two courses (β around 0.3). For macroeconomics, this gap is around 0.5, the same difference that exists for microeconomics for college C. For college B, the grading practice has been very strict in microeconomics.

3.2. Results from Common Exam After 2018 (Tables 7-9)

After the fusion in 2018, there was common design for courses with identical exams and grading standard across the campuses after 2019. Tables 7 to 9 present the results for three of the subjects (not available for macroeconomics due to corona and no written exam with grades).

Table 7. *Microeconomics Performance Before (2016–2017) and After Fusion (2019).*

Letter Grade	School A					School B					School C				
	Before fusion		After fusion		Diff.	Before fusion		After fusion		Diff.	Before fusion		After fusion		Diff.
	N	Percent	N	Percent		N	Percent	N	Percent		N	Percent	N	Percent	
F	43	7.9	16	4.3	-3.6	68	45.3	10	8.3	-37	31	17.5	20	17.5	0
E	45	8.2	24	6.4	-1.8	29	19.3	30	24.8	5.5	16	9	25	21.9	12.9
D	81	14.8	80	21.3	6.5	25	16.7	32	26.4	9.7	30	16.9	24	21.1	4.2
C	172	31.4	118	31.5	0.1	10	6.7	36	29.8	23	49	27.7	29	25.4	-2.3
B	131	23.9	120	32	8.1	10	6.7	6	5	-2	36	20.3	13	11.4	-8.9
A	75	13.7	17	4.5	-9.2	8	5.3	7	5.8	0.5	15	8.5	3	2.6	-5.9
Sum	547	100	375	100		150	100	121	100		177	100	114	100	0
Mean ¹	2.97		2.94		0.03	1.26		2.15		-0.99	2.5		1.99		0.51

¹ F: 0, E: 1, D: 2, C: 3, B: 4, A: 5

Table 8. *Management, Performance Before (2016–2017) and After Fusion (2019).*

Letter Grade	School A					School B					School C				
	Before fusion		After fusion		Diff.	Before fusion		After fusion		Diff.	Before fusion		After fusion		Diff.
	N	Percent	N	Percent		N	Percent	N	Percent		N	Percent	N	Percent	
F	29	4.6	14	3.7	-0.9	14	4	62	29.1	25	0	0	16	11.3	11.3
E	49	7.8	54	14.4	6.6	22	6.3	33	15.5	9.2	0	0	23	16.2	16.2
D	122	19.4	55	14.6	-4.8	64	18.2	40	18.8	0.6	9	4.9	44	31	26.1
C	205	32.6	109	29	-3.6	113	32.1	53	24.9	-7	63	34.6	29	20.4	-14.2
B	171	27.2	94	25	-2.2	119	33.8	23	10.8	-23	98	53.8	22	15.5	-38.3
A	62	8.3	50	13.3	5	20	5.7	2	0.9	-5	12	6.6	8	5.6	-1
Sum	628		376	100		352	100	213	100		182	100	142	100	
Mean ¹	2.95		2.97		0.02	3.03		1.76		1.17	3.62		2.61		0.99

¹ F: 0, E: 1, D: 2, C: 3, B: 4, A: 5

Assuming that the composition of students remains the same between the institutions, it gives a picture of how different enrolment qualifications and other differences influence the grade levels with identical exams with the same grade standards. School A has quite stable distributions and grade means before and after the fusion for all three subjects. For school B and C, there are considerable changes in mean grades and the spread of the grades, with the exception of marketing for school B. With the exemption of microeconomics from institution B, the mean grades from before and after the fusion decreased by a half to more than one letter grade. The effect is opposite for microeconomics from school B; the mean grade went up from 1.26 to 2.15

Table 9. *Marketing Performance Before (2016–2017) and After Fusion (2019).*

	School A					School B					School C				
	Before fusion		After fusion		Diff.	Before fusion		After fusion		Diff.	Before fusion		After fusion		Diff.
Letter Grade	N	Percent	N	Percent		N	Percent	N	Percent		N	Percent	N	Percent	
F	49	7.5	11	3.2	-4.3	22	7	16	9.2	2.2	4	2.5	4	5.1	2.6
E	46	7	16	4.6	-2.4	31	9.8	17	9.8	0	13	8.3	7	9	0.7
D	132	20.1	92	26.4	6.3	60	19	31	17.9	-1	13	8.3	27	34.6	26.3
C	240	36.5	143	41.1	4.6	120	38.1	67	38.7	0.6	30	19.1	25	32.1	13
B	161	24.5	71	20.4	-4.1	62	19.7	39	22.5	2.8	62	39.5	14	17.9	-21.6
A	29	4.4	15	4.3	-0.1	20	6.3	3	1.7	-5	35	22.5	1	1.3	-21.2
Sum	657	100	348	100		315	100	173	100		157	100	78	100	
Mean ¹⁾	2.77		2.83		-0.06	2.73		2.60		0.13	3.52		2.52		1.00

¹ F: 0, E: 1, D: 2, C: 3, B: 4, A: 5

Table 10. Comparing Predicted Grade Before the Fusion (2016–2017) With Actual Grade After the Fusion (2019) for the Three Schools A, B and C.

Course	Data from 2016–2017 and applying result from Model 2					Data from common exam 2019/2020		
	Letter Grade ¹			Predicted Grade ²		Actual Grade		
	A	B	C	B	C	A	B	C
Macroeconomics	2.87	2.67	2.39	2.17	1.94	Missing data		
Microeconomics	2.97	1.26	2.50	2.21	1.99	2.94	2.15	1.99
Marketing	2.77	2.73	3.52	2.47	2.39	2.83	2.60	2.52
Management	2.95	3.03	3.62	2.67	2.48	2.97	1.76	2.61

¹ F: 0, E: 1, D: 2, C: 3, B: 4, A: 5

²Actual grade – β (dummy variable)

3.3. Comparing the Predictors From Model 2 With tge Actual Performance After Fusion

Table 10 presents the predicted grade from schools depending on entry qualifications and gender (Model 2) and comparing this with the actual differences after the fusion. If we disregard management from school B, the calculated differences from Model 2 give a good predictor of the students’ actual mean differences from the three campuses depending upon variations in academic skills. Hypotheses 2 (H2) is confirmed.

4. DISCUSSION and CONCLUSION

The results in this study are mainly in line with previous research. GPA is a proxy of academic skills. The GPA scores from school A are between a half and one grade higher than at the two other schools. The variations are also lower at A than at B and C. From Model 2, we notice there is a strong positive correlation between GPA and performance in business courses. The associations are stronger for the quantitative courses (β around 1.0) than for the non-quantitative courses (β around 0.5). GPA is a good predictor for success in business courses (Brookshire & Palocsay, 2005). With similar grade standards across the colleges, this will influence the grading levels. This is the main reason why students from school A deserve higher mean grades than for the two other schools. Another key factor is mathematical background. Table 2 shows that a higher percent of students from A have more theoretical mathematics compared to the two other schools. Especially at school C, few students have a background in advanced mathematics. Around 85 percent have only practical mathematics (P-maths, the alternative to N- and S-maths). Mathematical skills are linked to success in business studies and especially in quantitative courses (Mallik & Lodewijks, 2010). This study confirms this connection with significant positive β values for all courses (Model 2). The impact varies depending upon S- or N-maths and quantitative and non-quantitative courses. For micro- and macroeconomics, the β value is strongest for S-maths. The explanation might be that S-mathematics are adapted and intended for business students. For marketing and management, however, the impact is strongest for N-mathematics. In those courses, one does not use mathematical formulas in the presentation of the subjects. Therefore, one does not need the mathematical tools learned by studying theoretical mathematics. However, studying N-mathematics helps students improve the design and structure of written essays in marketing and management. Hence, the reward in non-quantitative courses is higher grade scores. This result is similar to the finding of Brookshire and Palocsay (2005).

This study shows that gender still matters. There is a plenty of literature on the topic of gender and success in business and economics courses and with some mixed results (Johnson et al., 2014). Even though the gender gap seems to be lessening, there is still a tendency for males to perform better in quantitative courses (Borde, 2017; Mavruk, 2019). On the other hand, many studies show that females outperform males in non-quantitative business courses (Friday et al., 2006; Volchok, 2019). Gender differences in preferences and personal characteristics can probably explain some of the gender gap (Chevalier, 2002).

It looks like there is a different practice among the quantitative and non-quantitative courses, especially for campus C. In the non-quantitative courses, the students get higher scores at institution C than at institution A despite the lower entrance qualification. Our model suggested the grade should be about one letter grade lower if one used the same evaluation and standard as at school A (with a dummy value of $\beta = 1.13$ for marketing and 1.14 for management). This may indicate that one instructor at institution C was not aware of having less qualified students in marketing and management or that one just decided to use ECTS locally. This implies that a more lenient grading practice was used than that in accordance with the national advice. For campus B, the mean grades in non-quantitative courses were lower than for school A before the fusion, but it was not enough to catch up the differences in academic skills (β is around 0.3).

The mean grades for quantitative courses were substantially lower at schools B and C than at school A. The divergence, however, was not sufficient to explain the qualification differences. The grading habit shows a difference of a half letter grade, but Model 2 and the results of the identical exam after the fusion suggest that the difference should be around one letter grade due to the different level of academic abilities. The exception is microeconomics at school B where this study suggests the instructor had been too strict. Findings from model 2 indicate that the students on average deserve almost a higher score by one letter grade ($\beta = -0.95$). With the identical exam, the mean score for this subject increased almost the same (from 1.26 to 2.15, Table 6).

A reason for different grading practices for quantitative and non-quantitative courses could be due to the characteristics of the subjects (Beenstock & Feldman, 2018). Quantitative courses are easier to grade since there normally is only one correct answer. Therefore, it is easy to judge the qualifications. By contrast, in non-quantitative subjects, one can present the essay differently and various presentations and solutions can achieve high scores. There is often more than one way to provide excellent answers. Hence, the instructor can be more likely to give students the benefit of the doubt and reward them if in doubt. Notice also that the results from this study indicated different grading practice between school B and C in management and marketing in 2016–2017.

The results from the regression model (Model 2) seem to be a good predictor of the variation in grading practices among the schools. The actual grading differences after the fusion for the three subject are consistent with the calculated gap between the schools before the fusion but with the exception of management for school B. One reason might be that after the fusion the design of the common exam changed. The instructor did not adjust the course programme to align with this. Therefore, the students at school B did not prepare for the modified exam devised after the fusion.

Although the goal is to apply similar grading practices regardless of the admission criteria and colleges, this study reports that this is not the case. This supports the conclusion of Møen and Tjelta (2010). The composition of the student population influences grading practices, as it is easier to achieve good grades with undergraduates with weaker qualifications. The instructors are less strict in grading if the students are academically weaker. Two students with equal qualities can expect to get different grades depending upon their peers. At school B and C, it was possible for a student to achieve the same result as from school A with less effort. An

average student could expect to improve their scores by choosing a college with low admission standards, and the divergences are huge. A possible explanation for the divergence from the national standard is the self-interest of students, instructors and college administration. Students achieve better grades, and this provides more opportunities for further studies and careers. The instructor can verify good grades, which are indicators of good teaching performance. The dean and the college administration can report good results, which generates more funds from the Ministry of Education and can attract more students.

The policy in Norway is very clear: the directive states that similar students shall receive the same scores regardless of the selected campus, composition of student population and enrolment criteria. The effect of different grading standards is that it gives a wrong signal when applying for work or further studies because the ranking will be incorrect. Therefore, some students are offered entry to programmes at the expense of better-qualified applicants. This is especially true if there are different practices among undergraduate programmes depending on the campus, as the composition of students in master's programmes will be wrong. Skilled students can be rejected by campuses with strict grading habits, resulting in an ineffective use of resources. This study shows that there are good reasons for applying the same exam regardless of the campus, and it probable that the university management was aware of this. Therefore, it was necessary for exams with same instructors to be evaluated by them in order to ensure there would be no differences in grading practice. From the current data, it appears that there were substantial differences in instructors' grading practices and evaluation. Therefore, it is no coincidence that our research suggests substantial differences in practices at small colleges. Both the academical environment and colleagues to notice and adjust for obviously poor judgement in grading are lacking. As examples, the instructor in microeconomics at college B has apparently been too strict, and the instructors at college C have been too lenient in assessing the grades in management and marketing.

Another issue is how a common exam for the three campuses with such big differences in enrolment qualifications and composition of student population will influence the academic level of the courses, the difficulty of the examination, the grading standards, recruitment of students and the academic milieu. Will the result be lower standards and grade inflation at campus A while the instructors will lose some of their motivation at campus B and C? This can be explored in future studies.

This study is based on access to administrative data, and it is likely that many unobserved factors have impacted the results. This can explain why the adjusted R^2 is rather low and the different values of the independent variables associated with success at the different schools. There can be differences in the quality of instructors and their judgement, students' level of effort and personal characteristics and the design of both the course and of the exam format. No data are available to check whether the composition of students from different schools/campuses were the same in 2019 as in 2016–2017, but it is a plausible assumption since the composition had been quite stable over a long period of time.

This analysis shows that GPA, mathematical background and gender are good predictors for performance in business courses. There is a substantial gender distinction between quantitative and non-quantitative courses. Male students have more success than female peers in quantitative course, while the situation is the opposite for non-quantitative courses where the women get higher scores than the men.

The main contribution of this paper was to investigate grade practices between different schools offering the same subject. Despite the national intention to have an equal award system that is independent of the composition of the student body and the colleges, this study reveals a substantial variation in grade standards. Small colleges with academically weaker students tend

to have softer grading practices and considerable variety in grade evaluation, depending upon the instructor.

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

Leiv Opstad  <https://orcid.org/0000-0003-2400-6581>

5. REFERENCES

- Alcock, J., Cockcroft, S., & Frank, F. (2008). Quantifying the advantage of secondary mathematics study for accounting and finance undergraduates. *Accounting & Finance*, 48(5), 697-718.
- Bagues, M., Labini, M. S., & Zinovyeva, N. (2008). Differential grading standards and university funding: Evidence from Italy. *CESifo Economic Studies*, 54(2), 149-176. <https://doi.org/10.1093/cesifo/ifn011>
- Ballard, C. L., & Johnson, M. F. (2004). Basic math skills and performance in an introductory economics class. *The Journal of Economic Education*, 35(1), 3-23. <https://doi.org/10.3200/JECE.35.1.3-23>
- Bardach, L., & Klassen, R. M. (2020). Smart teachers, successful students? A systematic review of the literature on teachers' cognitive abilities and teacher effectiveness. *Educational Research Review*, 30, 100312. <https://doi.org/10.1016/j.edurev.2020.100312>
- Beenstock, M., & Feldman, D. (2018). Decomposing university grades: A longitudinal study of students and their instructors. *Studies in Higher Education*, 43(1), 114-133. <https://doi.org/10.1080/03075079.2016.1157858>
- Blaylock, A., & Lacewell, S. K. (2008). Assessing prerequisites as a measure of success in a principles of finance course. *Academy of Educational Leadership Journal*, 12(1), 51.
- Bonesrønning, H. (1999). The variation in teachers' grading practices: Causes and consequences. *Economics of Education Review*, 18(1), 89-106. [https://doi.org/10.1016/s0272-7757\(98\)00012-0](https://doi.org/10.1016/s0272-7757(98)00012-0)
- Bonesrønning, H. and Opstad, L. (2012). How much is students' college performance affected by quantity of study? *International Review of Economics Education*, 11(2), 46-63. [https://doi.org/10.1016/s1477-3880\(15\)30012-8](https://doi.org/10.1016/s1477-3880(15)30012-8)
- Borde, S. F. (2017). Student characteristics and performance in intermediate corporate finance. *Journal of Financial Education*, 43(1), 1-13.
- Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, 86(4), 803-848. <https://doi.org/10.3102/0034654316672069>
- Brookshire, R. G., & Palocsay, S. W. (2005). Factors contributing to the success of undergraduate business students in management science courses. *Decision Sciences Journal of Innovative Education*, 3(1), 99-108. <https://doi.org/10.1111/j.1540-4609.2005.00054.x>
- Cheng, L., DeLuca, C., Braund, H., Yan, W., & Rasooli, A. (2020). Teachers' grading decisions and practices across cultures: Exploring the value, consistency, and construction of grades across Canadian and Chinese secondary schools. *Studies in Educational Evaluation*, 67, 100928. <https://doi.org/10.1016/j.stueduc.2020.100928>
- Chevalier, A. (2002). Education, motivation and pay of UK graduates: Are they different for women? *European Journal of Education*, 37(4), 347-369.

- Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education Policy Analysis Archives*, 8(1), 1-44. <https://doi.org/10.14507/epaa.v8n1.2000>
- Friday, E., Friday-Stroud, S. S., Green, A. L., & Hill, A. Y. (2006). A multi-semester comparison of student performance between multiple traditional and online sections of two management courses. *Journal of Behavioral & Applied Management*, 8(1), 66-81. <https://doi.org/10.1108/00251740510589742>
- Grove, W. A., Wasserman, T., & Grodner, A. (2006). Choosing a proxy for academic aptitude. *The Journal of Economic Education*, 37(2), 131-147. <https://doi.org/10.3200/JECE.37.2.131-147>
- Godor, B. P. (2017). Revisiting differential grading standards anno 2014: An exploration in Dutch higher education. *Assessment & Evaluation in Higher Education*, 42(4), 596-606. <https://doi.org/10.1080/02602938.2016.1173186>
- Hoefler, P., Yurkiewicz, J., & Byrne, J. C. (2012). The association between students' evaluation of teaching and grades. *Decision Sciences Journal of Innovative Education*, 10(3), 447-459. <https://doi.org/10.1111/j.1540-4609.2012.00345>
- Johnson, M., Robson, D., & Taengnoi, S. (2014). A meta-analysis of the gender gap in performance in collegiate economics courses. *Review of Social Economy*, 72(4), 436-459.
- Jones, C. T., Kouliavtsev, M. S., & Ethridge Jr, J. R. (2013). Lower level prerequisites and student performance in intermediate business courses: Does it matter where students take their principles courses? *Journal of Education for Business*, 88(4), 238-245. <https://doi.org/10.1080/08832323.2012.688777>
- Krishna, A., & Orhun, A. Y. (2020). EXPRESS: Gender (Still) matters in business school. *Journal of Marketing Research*. <https://doi.org/10.0022243720972368>
- Lawrence, J. J., & Pharr, S. (2003). Evolution of admission standards in response to curriculum integration. *Quality Assurance in Education*, 11(4), 222-233. <https://doi.org/10.1108/09684880310501403>
- Mallik, G., & Lodewijks, J. (2010). Student performance in a large first year economics subject: Which variables are significant? *Economic Papers: A Journal of Applied Economics and Policy*, 29(1), 80-86. <https://doi.org/10.1111/j.1759-3441.2010.00051.x>
- Marcial, L. E., Hennessey, J. E., Curren, M. T., & Roberts, W. W. (2005). Do business communication courses improve student performance in introductory marketing?, *Journal of Education for Business*, 80(5), 289-294. <https://doi.org/10.3200/JOEB.80.5.289-294>
- Marini, J., Shaw, E., Young, L., & Ewing, M. (2018). Getting to know your criterion: Examining college course grades and GPAs over time. *The College Board*. Retrieved from <https://files.eric.ed.gov/fulltext/ED582569.pdf>
- Mavruk, T. (2019). Do men outperform women in finance classes? *Journal of International Business Education*, 14, 75-98.
- Mayer-Foulkes, D. (2002). On the dynamics of quality student enrollment at institutions of higher education. *Economics of Education Review*, 21(5), 481-489. [https://doi.org/10.1016/S0272-7757\(01\)00036-X](https://doi.org/10.1016/S0272-7757(01)00036-X)
- Møen, J., & Tjelta, M. (2010). Grading standards, student ability and errors in college admission. *Scandinavian Journal of Educational Research*, 54(3), 221-237. <https://doi.org/10.1080/00313831003764503>
- Odden, A., Borman, G., & Fermanich, M. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education*, 79(4), 4-32. https://doi.org/10.1207/s15327930pje7904_7
- Opstad, L. (2018). Success in business studies and mathematical background: The case of Norway. *Journal of Applied Research in Higher Education*, 10(3), 399-408. <https://doi.org/10.1108/JARHE-11-2017-0136>

-
- Opstad, L. (2020). Why are there different grading practices based on students' choice of business major? *Educational Process: International Journal*, 9(1), 43-57. <https://doi.org/10.22521/edupij.2020.91.3>
- Opstad, L., & Årethun, T. (2019). Factors influencing students' choice of mathematical level at high school and the impact this has on performance on business courses in Norway. *WEI International Academic Conference Proceedings 2019*, WestEastInstitute, 28-40.
- Opstad, L., & Årethun, T. (2020a). Skills, gender, and performance matter when undergraduate business students choose specialisation within business courses. *International Journal of Management, Knowledge and Learning*, 9(1), 95-107.
- Opstad, L., & Årethun, T. (2020b). Factors that explain undergraduate business students' performance in their chosen field. Does gender matter? *Global Conference on Business and Finance Proceedings*, 15(2), 2-21.
- Stinebrickner, R., & Stinebrickner, T.R. (2008). The causal effect of studying on academic performance. *The BE Journal of Economic Analysis & Policy*, 8(1), 1-53. <https://doi.org/10.2202/1935-1682.1868>
- Trapman, S., Hell, B., Hirn, J. W., & Schuler, H. (2007). Meta-analysis of the relationship between the big five and academic success at university. *Journal of Psychology*, 215, 132-151. <https://doi.org/10.1037/e518532013-271>
- Uyar, A., & Güngörmüş, A. H. (2011). Factors associated with student performance in financial accounting course. *European Journal of Economic & Political Studies*, 4(2), 139-154.
- Volchok, E. (2019). Differences in the performance of male and female students in partially online courses at a community college. *Community College Journal of Research and Practice*, 43(12), 904-920. <https://doi.org/10.1080/10668926.2018.1556134>
- Walstad, W. B., & Miller, L. A. (2016). What's in a grade? Grading policies and practices in principles of economics. *The Journal of Economic Education*, 47(4), 338-350. <https://doi.org/10.1080/00220485.2016.1213683>

The Effect of Formative Assessment Practices on Student Learning: A Meta-Analysis Study

Pinar Karaman ^{1,*}

¹Sinop University, Faculty of Education, Department of Educational Sciences, Sinop, Turkey

ARTICLE HISTORY

Received: Jan. 28, 2021

Revised: Aug. 22, 2021

Accepted: Sep. 22, 2021

Keywords:

Formative assessment,
Student learning,
Meta-analysis.

Abstract: The main purpose of this meta-analysis study is to investigate how formative assessment practices promote student learning in Turkey. 32 studies with 47 effect sizes that met the specified criteria such as using true experimental or quasi-experimental design and measuring learning outcomes were included as the final analysis in the meta-analytical review method. The overall mean effect size of the study was obtained as .72 ($SE = .07, p < .05$). Further investigation through subgroup analysis showed that the effect sizes made a significant difference on different types of formative feedback. The effect of features of formative assessment interventions on student learning indicated that student initiated formative feedback ($d = 1.16$) and mixed feedback ($d = .83$) had a large effect, which was followed by a medium effect of adult initiated formative feedback ($d = .69$) and a small effect of computer initiated formative feedback ($d = .42$). On the other hand, education level and publication type had no effect on student academic performance in the study. These findings support the positive effect of formative assessment practices on student learning. Such a result suggests that increasing the number of different types of formative assessment practices in the classrooms would promise a considerable contribution to student learning.

1. INTRODUCTION

Assessment is an important component of effective teaching and learning (Bransford et al., 2000; Hargreaves, 2008). Formative assessment strategy plays a crucial role in supporting the student learning. This assessment strategy provides effective feedback and instructional correctives in the teaching-learning process to improve students' learning, motivation, and self-regulation skills (Black & William, 2009; Cauley & McMillan, 2010; McManus, 2008; Popham, 2008). Formative assessment also known as assessment for learning, diagnostic testing, and feedback is an ongoing process used by teachers, students, and students' peers (Andersson & Palm, 2017a, 2017b; Bennett, 2011). Teachers can adjust their teaching practices to increase student learning through formative assessment (Black & William, 2009; Brookhart, 2009).

Formative assessment has the succeeding three main stages; namely, (1) determining goals, (2) providing feedback to enhance student performance with these goals, and (3) using feedback to

*CONTACT: Pinar KARAMAN ✉ pkaraman@sinop.edu.tr 📍 Sinop University, Faculty of Education, Department of Educational Sciences, Sinop, Turkey

improve further learning of students (Brookhart, 2010). One of the most important components of formative assessment is feedback that helps to provide evidence on student learning (Andersson & Palm, 2017a). Feedback helps students to understand current status of their learning to make further progress (Sadler, 1989). This feedback to advance student learning could come from different agents such as teachers, self-assessment, peer assessment, group assessment, and even computers (Sadler, 1989; Black & William, 1998; Graham et al., 2015; Wiliam, 2018). Feedback may be given to students in different time periods (instantly or delayed) (Andersson & Palm, 2017a). Thus, different types of feedback provide different formative assessment interventions (Hattie & Timperley, 2007). Feedback from teachers and students has an important role in formative assessment practices due to their significant support for student learning (Black & William, 2009), self-regulated learning (Andrade & Brookhart, 2016; Butler & Winne, 1995; Zimmerman & Bandura, 1994), and peer-assisted learning (Gielen et al., 2010). Students' engagement in self-assessment and peer-assessment for effective formative assessment strategies promotes their self-regulated learning skills (Zimmerman, 2002; Weldmeskel & Michael, 2016). In addition to teacher and student initiated formative assessment, computer initiated formative assessment also provides immediate feedback to students (Maier et al., 2016; Van der Kleij et al., 2015). These studies showed that computer-based feedback has an important effect on student learning (Kluger & DeNisi, 1996; Miller, 2009). However, in comparison to formative feedback from teachers and students, computer-based formative assessment is more difficult to apply (Maier et al., 2016).

Several meta-analysis research studies have been conducted to investigate the efficiency of formative assessment strategies. The results of these studies indicate that effect sizes vary with a considerable range (Black & William, 1998; Fuchs & Fuchs, 1986; Graham et al., 2015; King & Nash, 2011; Lee et al., 2020). The magnitude of the effect sizes of differences could come from a variety of the meta-analysis studies that focused on formative assessment types, feedback procedures, and learning subjects. (Maier et al., 2016). Black and William (1998) provided meta-analysis of 250 studies on the effect of formative assessment practices and found positive influence of formative assessment on student achievement with effect sizes ranging from .40 to .70. They argued that formative assessment intervention is more important than other educational interventions to improve student learning. Hattie (2009) examined the factors that were significantly related to student achievement through multiple meta-analysis and found that one of the most important factors is teachers' use of formative assessment strategies. Kingston and Nash (2011), in their meta-analysis research, examined a limited number of studies (a total of 13 studies) to uncover the effect of formative assessment on K-12 student achievement and reported the mean effect size as .20. They suggested that more studies are needed to investigate the relationship between formative assessment and academic achievement. On the other hand, Graham et al. (2015) investigated the effect of formative assessment on students' writing performance and reported a weighted mean effect size of .61. They also reported the impact of feedback from adults ($d=.87$), feedback from students (peer assessment, $d=.58$ and self-assessment, $d=.62$), and feedback from computers ($d=.38$) to student writing performance. Lee et al. (2020) analyzed 33 studies about K-12 education in the USA and reported an overall mean effect size of .29. They found the effectiveness of formative assessment on different subject areas. Moreover, meta-regression analysis denoted that student-initiated self-assessment was the most effective one ($d=.61$) among other interventions. In comparison to informal feedback ($d=.52$), formal formative assessment feedback was more effective on student learning. Briefly, although several meta-analysis studies in the literature concluded that formative assessment has a positive effect on student learning, the effectiveness of different types of formative assessment interventions was not examined adequately in previous meta-analysis studies.

For more than a decade, Turkey has given priority to the improvement of assessment for learning in education programs and offered more support to teachers to encourage them to use this assessment strategy more frequently in their classrooms (Kitchen et al., 2019; MoNE, 2017, 2020). With the growing importance of using formative assessment strategies in classrooms, the number of research studies conducted about the effectiveness of the formative assessment has increased considerably in recent years (Delen & Bellibaş, 2015; Double et al., 2020; Lee et al., 2020; Ozan & Kınca, 2018). In parallel to the publication of more research studies, a meta-analysis research study was developed in the present study. In this regard, the purpose of the study was to provide a synthesis of the experimental and quasi-experimental studies about the effectiveness of formative assessment practices on student learning in Turkey. In addition to the effectiveness of formative assessment in each education level from primary to tertiary, features of formative assessment interventions and publication types were also examined as moderator variables in the study.

In this study, how formative assessment practices in Turkey's education system promote student learning was investigated through meta-analysis. Therefore, the present study is of high importance to gain a better understanding of the effect of formative assessment practices on student learning. Examining the effectiveness of formative assessment and its moderators (i.e. types of formative assessment interventions, education level) would contribute to the literature. In this sense, the following research questions were asked in this study:

- 1) What effect do formative assessment interventions have on student learning according to the findings of the experimental studies applied in Turkey?
- 2) Do the findings of the experimental studies applied in Turkey about the effect of formative assessment interventions on student learning differ significantly according to moderating variables (features of formative assessment interventions, education level, and publication type)?

2. METHOD

Meta-analysis method was conducted in the present study. Meta-analysis is more than a statistical technique that synthesizes a series of research studies answering the same research question in a systematic way (Borenstein et al., 2009; Glass et al., 1981). This statistical method called as quantitative research synthesis helps to summarize and compare the results of the studies. When compared with other research synthesis, meta-analysis focuses on research outcomes to draw conclusions with effect sizes (Card, 2012). ProMeta3 (professional statistical software) was used for data analysis in the present study.

Several steps were carried out to perform meta-analysis (Field & Gillett, 2010); namely, (1) doing a literature review to formulate a problem; (2) specifying inclusion/exclusion criteria; (3) calculation of effect size for each study; (4) doing meta-analysis; (5) assessing moderator variables with advanced analysis; (6) doing publication bias analysis; and (7) writing the results.

2.1. Literature Review

First, research studies that investigated the relationship between formative assessment practices and student learning were collected through a search of databases. Key words were specified in English and Turkish as “formative assessment” and “biçimlendirici değerlendirme”, and “experimental” and “deneysel”, respectively. These databases are Google Scholar, PsycINFO, Turkish Council of Higher Education (YÖK) National Thesis Center, Education Research Complete, ERIC (2020), Web of Science, ULAKBİM (2020), and EBSCO (2020). Peer-reviewed journals, master's theses, and doctoral dissertations were included in the meta-analysis.

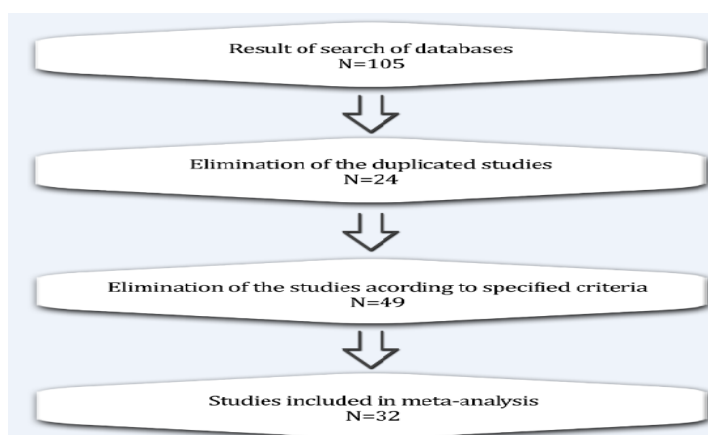
2.2. Selection Criteria

If studies had to meet the following criteria, they were included in the meta-analysis. These criteria were as follows: (1) studies that had true experimental or quasi experimental design with control group and treatment group with formative assessment interventions; (2) studies that measured learning outcomes; (3) studies with enough information to calculate effect sizes; (4) students at different education levels (i.e., primary, secondary, and tertiary); and (5) studies written in English or Turkish language.

A number of 105 records were identified through the search of databases. The number of studies dropped to 81 after removing duplicates and eliminating studies according to the specified criteria (i.e. studies that do not have formative assessment intervention, studies that do not have student learning or academic achievement, and studies that do not have enough statistics). Ultimately, 32 studies with 47 effect sizes that were unpublished theses and peer-reviewed articles that had experimental studies about the effectiveness of formative assessment on student learning were included. A flow chart that summarizes the inclusion of studies through search in the meta-analysis is given in [Figure 1](#). Therefore, the data included 32 studies as shown in [Table 1](#).

Table 1. *The studies included in the meta-analysis.*

Included Study	Number of Effect Sizes	Course	Education Level
Arici and Kaldırım (2015)	1	Language	Tertiary
Atik and Erkoç (2017)	2	Science	Secondary
Aydın et al.(2016)	1	Science	Secondary
Batıbay (2019)	1	Literacy	Secondary
Bayat (2014)	1	Literacy	Tertiary
Bayrak et al. (2019)	2	Science	Secondary
Baysal (2020)	1	Foreign Language	Secondary
Bolat et al. (2017)	1	Computer Science	Tertiary
Demirkesen (2019)	1	Foreign Language	Tertiary
Elvan (2012)	1	Social Sciences	Secondary
Eraz and Öksüz (2015)	1	Mathematics	Primary
Güzel (2018)	1	Science	Secondary
Hotaman (2020)	1	Teacher Training	Tertiary
Kaya and Ateş (2016)	1	Language	Primary
Kıncal and Ozan (2018)	1	Measurement and Evaluation	Tertiary
Korkmaz et al. (2019)	1	Foreign Language	Secondary
Köksalan (2019)	1	Physics	Secondary
Kuzudişli (2019)	2	Science	Secondary
Müldür and Yalçın (2019)	1	Language	Secondary
Ozan and Kıncal (2018)	1	Social Sciences	Secondary
Özgür (2016)	1	Computer Education	Tertiary
Sever and Memiş (2013)	4	Language	Primary
Tavşanlı (2019)	1	Language	Primary
Topal (2020)	1	Educational Sciences	Tertiary
Turan and Sakız (2014)	2	Science	Secondary
Yalaki and Bayram (2015)	1	Chemistry	Tertiary
Yaşar (2018)	4	Mathematics	Secondary
Yıldız and Kılıç Çakmak (2019)	1	Project Management and Application	Tertiary
Yılmaz (2015)	1	Mathematics	Secondary
Yorgancı (2015)	1	Mathematics	Tertiary
Yurdabakan and Cihanoğlu (2009)	6	Foreign Language	Secondary
Yurdabakan and Olgun (2011)	1	Science	Primary

Figure 1. Flow chart of inclusion of the studies.

2.3. Formative Assessment Interventions

Formative assessment interventions have several types of feedback. The sources of formative feedback could come from teachers, self, peers, computers, or mixed (Andrade, 2010; Graham et al., 2015). In the present meta-analysis study, the studies that have various formative feedback from adults (teachers), computers, students, and mixed are coded. The studies that have multiple interventions such as self-assessment, peer assessment, group assessment, adult feedback, and/or computer feedback were coded as mixed.

2.4. Statistical Analysis of Effect Sizes

In meta-analysis, several standardized effect sizes are used to summarize direction and magnitude of effects in research studies such as Cohen's d , Hedge's g or Glass's g (Başol-Göçmen, 2004; Lipsey & Wilson, 2001). Hedge's g also called unbiased d was used to calculate the standardized mean differences between treated groups that have formative assessment interventions and control groups. When comparing Hedge's g statistic to Cohen's d and Glass's g statistic, Hedge's g uses the pooled standard deviation (Hedge, 1981). Hedge's g is preferred since it is better for small samples (<20) and significant for different sample sizes. Hedge's g , Cohen's d , and Glass's g statistic results are interpreted in the same way. Therefore, Cohen's proposal to classify the magnitude of effects was adopted in the study (Cohen, 1987). Magnitude of effects is described as small effect (.18), medium effect (.48), and large effect (.83) in social sciences (Cohen, 1962, 1987).

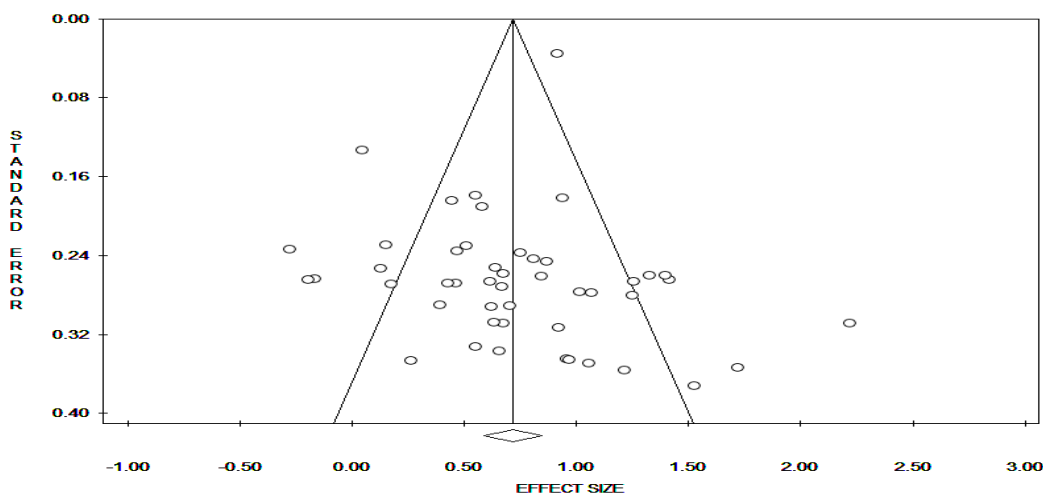
Two statistical models are used in meta-analysis. These models are fixed effect model and random effects model (Borenstein et al., 2009; Hedges & Vevea, 1998). It is assumed that only one true effect size exists for all studies including the meta-analysis with fixed effect model. On the other hand, true effect shows differences from one study to another study for the random effects model. Effect size might change due to the differences of studies such as studies that have different ages, education levels, income levels of participants, or differences of interventions. (Borenstein et al., 2009; Üstün & Eryılmaz, 2014). Due to the differences of studies, different effect sizes may occur in these studies. Therefore, estimating the mean distribution of effects is important for random effects model. Since research studies have different designs of formative assessment interventions and education levels, differences may occur from one study to another study in the present meta-analysis. For that reason, random effects model was employed in the present study. An average weighted effect size was calculated for the efficacy of formative assessment treatment. To test the heterogeneity in effect sizes, Q and I^2 statistic were used. A statistically significant p value for Q statistic means that the true effects vary (Borenstein et al., 2009). In other words, significant p value means that there is a significant variability among the effect sizes. I^2 statistic which gives the amount of

variance across studies due to heterogeneity was also computed (Higgins et al., 2003; Schwarzer et al., 2015).

2.5. Publication Bias

To have an accurate synthesis of studies in meta-analysis, assessing publication bias risk in the studies is important (Borenstein et al., 2009). There are several methods to assess the potential bias for a meta-analysis study. One of the methods is the funnel plot that gives the relationship between the observed effect size of each study and its standard error (Schwarzer et al., 2015). If studies were distributed symmetrically around the mean effect size in the plot, this would be the evidence of absence of publication bias. Funnel plot was used in the present study to inspect whether publication bias exists or not (Figure 2).

Figure 2. A funnel plot indicating standard error and observed effect size.



The funnel plot shows that studies were approximately scattered around the mean effect size. Since the interpretation of funnel plot could be subjective, some of the tests were also used to assess exactly any risks of bias such as Duval and Tweedie's Trim and Fill, Rosenthal's Fail-safe N test, Begg and Mazumdar Rank Correlation Test, and Egger's linear regression test (Begg, & Mazumdar, 1994; Duval, & Tweedie, 2000; Egger et al., 1997; Rosenthal, 1979). Trim and Fill method was used to remove extreme studies and estimate the effect sizes again in order to solve the asymmetry in funnel plot. The results of this method showed that trimming was not performed. Rosenthal's Fail-safe N test estimates how many missing studies would be needed to add to nullify the effect (Rosenthal, 1979). Rosenthal (1979) suggested that if the Fail-safe N test shows that large numbers of studies are needed to nullify the common effect rather than a few studies (i.e. five or ten), it can be concluded that true effect may not be zero in the study (Borenstein et al., 2009). In the present meta-analysis, the number of studies was 5681 according to Rosenthal's method. Therefore, it can be said that the results of the meta-analysis with 47 effect sizes would not be robust if 5681 studies were added. Besides, Egger's linear regression test was not statistically significant ($b = -0.66$, $p = 0.155$). As a result, Funnel Plot, Trim and Fill Method, Fail-safe N Test, and Egger's Linear Regression Test generally showed a low risk of publication bias that could be negligible.

3. FINDINGS

The number of studies included in the meta-analysis and the characteristics of these studies are summarized in Table 2. Most of the studies included in the meta-analysis were journal articles (68.75 %) and master's and doctoral theses (31.25 %) as publication type. 50 % of these studies were conducted at secondary school level, 34.37 % at tertiary level, and 15.62 % at primary

school level. The studies with the treatment groups having various formative assessment interventions were also described in the meta-analysis. The features of formative assessment interventions showed that 37.5 % of these studies have adult (teacher) initiated feedback, 31.25 % have computer initiated feedback, 15.62 % have student initiated feedback, and 15.62 % have mixed feedback (including peer assessment, group assessment, teacher’s feedback, and/or computer feedback).

Table 2. Characteristics of the studies included in the meta-analysis.

		Frequency (f)	Percent (%)
Study type	Thesis (master’s and doctoral)	10	31.25
	Article	22	68.75
Education level	Primary	5	15.62
	Secondary	16	50
	Tertiary	11	34.37
Features of formative assessment interventions	Adult initiated feedback	12	37.5
	Student initiated feedback	5	15.62
	Computer initiated feedback	10	31.25
	Mixed feedback	5	15.62

As summarized in Table 3, the meta-analysis shows the overall effect size as .79 with standard error of .03 in the fixed model. Heterogeneity test was used to investigate the heterogeneity in effect size. The Q value was 188.91 with 46 degrees of freedom, and p value under .05 showed heterogeneity among the studies. In other words, true effect size may have varied across studies. Besides, I^2 statistic was estimated as 75.65% indicating that the percent of variance due to between-subject factors was large. The results revealed that the impact of formative assessment on student learning varied from one study to another. By using random effects model, overall meta-analysis showed that there was a significant effect of formative assessment on student learning ($g = .72$, $SE = .07$, 95% CI = .59; 85, $p < .05$).

Table 3. Overall effect sizes and heterogeneity results related to the effectiveness of formative assessment practices.

Model	k	Mean ES	SE	Lower Limit	Upper Limit	p	Heterogeneity			
							Q value	df	p	I^2
Fixed	47	.79	.03	.74	.84	.00	188.91	46	.00	75.65
Random	47	.72	.07	.59	.85	.00				

* $p < .05$; k= number of effects; ES= effect size

In Figure 3, the forest graph demonstrating the effect size of each study based on the random effects model is presented. 3253 participants were involved in the analysis ($N_1 =$ Experimental group and $N_2 =$ Control group). It can be seen that overall effect size in random effects model across studies has a moderate level in favor of the experimental group ($g = .72$, $p < .05$).

As mentioned in Figure 3, heterogeneity test showed that the effectiveness of formative assessment practices varied from one study to the other. To investigate this variation, subgroup analysis was conducted in the present study. It is hypothesized that this variation may be explained with the studies that used various formative assessment interventions applied to different education levels and publication types. Mixed effect analysis based on random effects weights within subgroups was used to test the model. The results are presented for the features of formative assessment interventions in Table 4.

Figure 3. The forest graph showing the effect size of each study in meta-analysis.

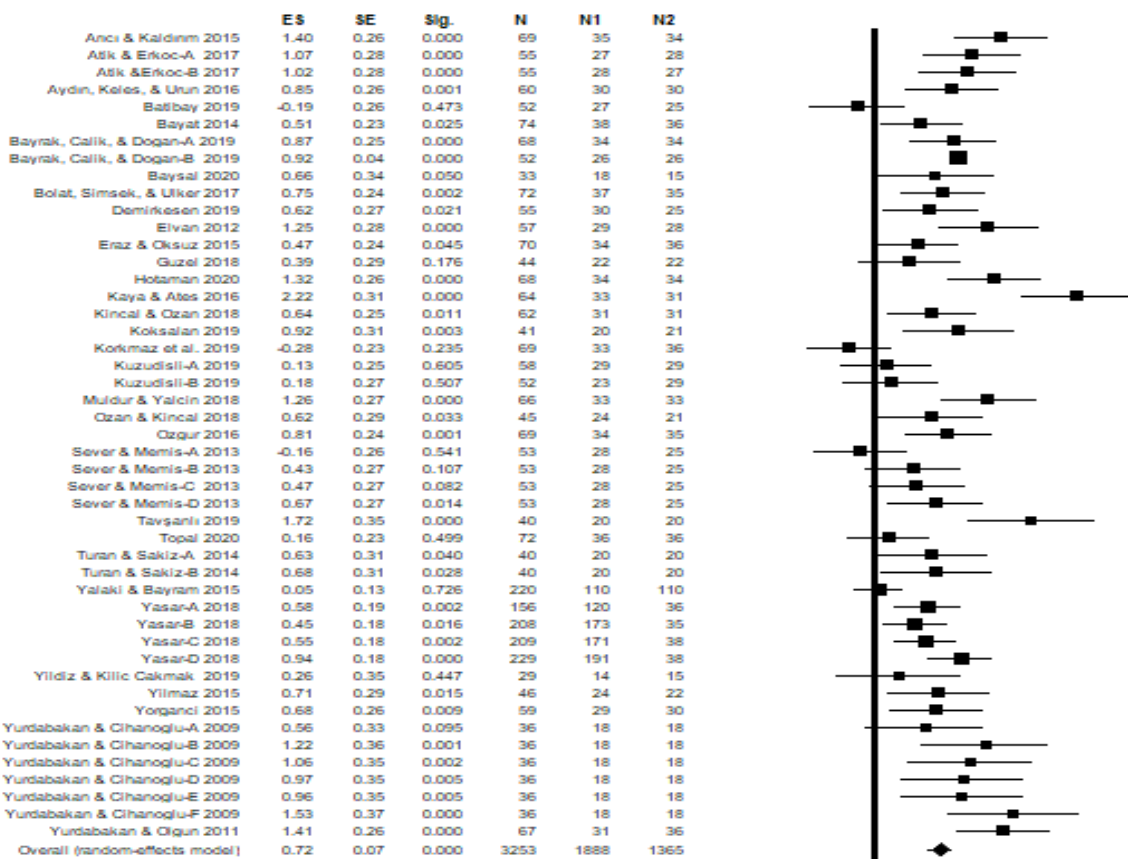


Table 4. Results of the subgroup analysis for features of formative assessment interventions.

	k	Mean ES	SE	Lower Limit	Upper Limit	p	Q value	df	p
Adult initiated feedback	20	.69	.09	.50	.87	.000			
Student initiated feedback	10	1.16	.17	.83	1.49	.000			
Computer initiated feedback	12	.42	.14	.14	.71	.003			
Mixed feedback	5	.83	.19	.46	1.21	.000			
Heterogeneity test							11.54	3	.009

* $p < .05$; k= number of effects; ES= effect size; SE= standard error

Mean effect size for each group was estimated by the mixed effects model. The effect sizes for each formative assessment interventions that varied between 1.16 and .42 were statistically significant. The results of the subgroup analysis might suggest that student initiated formative feedback ($d=1.16, p<.05$) and mixed feedback ($d=.83, p<.05$) had a large effect followed by a medium effect of adult feedback ($d=.69, p<.05$), and a small effect of computer feedback ($d=.42, p<.05$) on student academic performance. To compare the effect size for the subgroups, heterogeneity test was used. Total between tests ($Q=11.54, df=3, p<.05$) showed that the effect size might have varied by formative assessment intervention subgroups. In other words, features of formative assessment interventions such as adult initiated, student initiated, computer initiated, and mixed formative assessment differed significantly in the magnitude of effects.

In Table 5, mixed effects analysis was also used to estimate the effect size of groups in terms of their education levels (primary, secondary, and tertiary). The mean effect sizes that were estimated at different education levels ranged between .89 and .64 and were statistically significant. The results showed that the effect size at primary level had a large effect ($d=.89, p<.05$), while it had a medium effect on student academic performance at secondary level ($d=.71, p<.05$) and tertiary level ($d=.64, p<.05$). The results of the heterogeneity test yielded that comparison of subgroups at different education levels did not make a significant contribution to the variance ($Q=.66, df=2, p=.71$).

Table 5. Results of the subgroup analysis for education level.

Education Level	k	Mean ES	SE	%95 Confidence Interval		p	Heterogeneity		
				Lover Limit	Upper Limit		Q value	df	p
Primary	8	.89	.27	.37	1.41	.00			
Secondary	28	.71	.07	.56	.86	.00			
Tertiary	11	.64	.14	.36	.92	.00			
							.66	2	.71

* $p < .05$

The studies that included the meta-analyses were grouped in terms of publication type: articles, and theses (master’s and doctoral) (see Table 6). Mixed effect analysis showed that effect sizes according to these groups ranged between .78 and .57 and were statistically significant. The magnitude of effect size showed that articles have higher effect ($d=.78, p<.05$) than that of the theses ($d=.57, p<.05$). Heterogeneity test also showed that effect sizes of subgroups according to their publication type did not make a significant contribution to the variance ($Q=2.27, df=1, p=.13$). In other words, the distribution of effect sizes of studies according to publication type was homogeneous.

Table 6. Results of the subgroup analysis for publication type.

	k	Mean ES	SE	%95 Confidence Interval		p	Heterogeneity		
				Lover Limit	Upper Limit		Q value	df	p
Article	33	.78	.08	.63	.94	.00			
Thesis	14	.57	.11	.35	.80	.00			
							2.27	1	.13

* $p < .05$

4. DISCUSSION and CONCLUSION

In the meta-analysis study, 32 studies with a total of 47 effect sizes that met the inclusion criteria were estimated. The results of the study showed the overall mean effect size of .72. The overall mean effect size was consistent with previous meta-analysis results that effect sizes of the effectiveness of formative assessment ranged between .40 and .70 (Black & William, 1998; Graham et al., 2015). Besides, subgroup analysis was used to estimate whether the mean effect size was influenced by the features of formative assessment interventions, education level, and publication type.

The meta-analysis results showed how effective the features of formative assessment interventions were on student learning. The impact of different types of formative assessment interventions on student learning varied. The feedback from the students had the largest effect

but the feedback from the computers had the smallest effect on student learning. Moderator analysis showed that the effect sizes made a significant difference as to different types of formative feedback. Variation in effect sizes may be related to the features of formative assessment interventions in the present study. The impact of features of formative assessment interventions on student learning was examined in a few meta-analysis studies (Graham et al., 2015; Klute et al., 2017; Lee et al., 2020). Lee et al. (2020) examined various formative assessment feedback by using meta-regression. They found a similar result that the effect of student-initiated formative assessment feedback was significantly higher than teacher's formative assessment feedback and mixed feedback from both students and teachers. They implied that learner's active role is important for successful formative assessment based on their findings. Graham et al. (2015) examined four types of formative assessment feedback and found that feedback that came from teachers had the largest impact, but the feedback that came from computers had the smallest impact on student learning. In addition, by using meta-regression they also found that the effect size was not statistically related to grade level, types of formative feedback, or study quality. The present study generally shows similar results with the previous meta-analysis studies. It suggests that various formative assessment interventions in classrooms were effective. When comparing the formative assessment practices, the effect of student initiated formative interventions such as self-assessment, peer assessment, and group assessment was significantly higher than the other formative assessment interventions. Teachers, learners, and peers all have a crucial role for effective formative assessment (Black & William, 2009). The findings specifically indicated that learners' active role is very important for successful formative assessment (Clark, 2012; Lee et al., 2020).

The present study investigated that how mean effect size was in different education levels. While the highest mean effect size was found at primary school level, the lowest mean effect size was found at tertiary level. Mixed effects analysis was used to examine whether group differences were significant or not. The results showed that education level did not make a significant contribution to the variance. Likewise, King and Nash (2011) found that grade level did not make a difference on the effect of formative assessment on student learning. It can be interpreted that formative assessment is effective for all levels of education (Black & William, 1998; King & Nash, 2011). Therefore, the number of using formative assessment in classrooms should be increased in all levels of education.

Lastly, the studies included in the meta-analysis were categorized into two groups (published articles versus theses). Heterogeneity test showed that effect sizes of studies according to their publication type were homogeneous. The magnitude of the effect size did not significantly differ between the published articles and unpublished theses. It can be concluded that this finding was resistant to file drawer treat (publication bias) (Rosenthal, 1979).

Briefly, the present meta-analysis synthesized research studies conducted in Turkey showed that formative assessment interventions have a positive impact on student learning for all education levels. Assessment for learning rather than assessment of learning has been more emphasized in Turkey's curriculum since the 2005 educational reform. Assessment for learning strategies that curriculum requires has become widespread from primary schools to universities (MoNE 2013, 2017, 2018, 2020, YÖKAK 2018, 2019). The present study could give evidence regarding the effectiveness of formative assessment on student learning in Turkey's education system. Besides, there are only a few studies that examined the effectiveness of formative assessment interventions types. Therefore, it suggests that more meta-analysis studies should be conducted on this area (Lee et al., 2020). The present meta-analysis study is promising to provide a significant contribution to literature regarding the effectiveness of formative assessment interventions types. That is why more empirical studies are needed to have evidence regarding the effectiveness of formative assessment practices. Increasing different types of

formative assessment practices especially encouraging learners to have an active role in this process (i.e. self-assessment, peer-assessment, and group-assessment) is crucial. Since the results suggest that use of formative assessment strategies is effective for all education levels, implementation of formative assessment activities efficiently in classrooms is also important. Thus, providing all teachers and college scholars with professional development as to how to use formative assessment tools is highly needed.

The findings of this meta-analysis study were limited by the number of studies on formative assessment conducted in Turkey. Another limitation in the study was examining a few moderator variables such as education level, types of formative assessment interventions, and publication types. In further meta-analysis research, investigation and comparison of more variables such as subject areas, formality of formative assessment, feedback procedures, and feedback time are needed.

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

Pinar Karaman  <https://orcid.org/0000-0002-2218-2701>

5. REFERENCES

- Andersson, C., & Palm, T. (2017a). The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learning and Instruction, 49*, 92-102. <https://doi.org/10.1016/j.learninstruc.2016.12.006>
- Andersson, C. & Palm, T. (2017b). Characteristics of improved formative assessment practice. *Education Inquiry, 8* (2), 104-122. <https://doi.org/10.1080/20004508.2016.1275185>
- Andrade, H. (2010). Students as the definite source of formative assessment: Academic self-assessment and the self-regulation of learning. In G. J. Cizek & H. L. Andrade (Eds.), *Handbook of formative assessment* (pp. 90–105). Routledge Publishing.
- Andrade, H., & Brookhart, S. M. (2016). The role of classroom assessment in supporting self-regulated learning. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation* (pp. 293–309). Springer Publishing.
- Başol-Göçmen, G. (2004). Meta-analiz geneli bir deęerlendirmesi [A general revision of meta-analysis]. *Sakarya Üniversitesi Eđitim Fakóltesi Dergisi, 7*, 186–192.
- Begg, C.B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088–1101. <https://doi.org/10.2307/2533446>
- Bennett, R. E. (2011). *Formative assessment: a critical review. Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. <https://doi.org/10.1080/0969594X.2010.513678>
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. <https://doi.org/10.1080/0969595980050102>
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31. <https://doi.org/10.1007/s11092-008-9068-5>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley and Sons Publishing.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: brain, mind, experience, and school (expanded ed.)*. National Academy Press.
- Brookhart, S. M. (2009). *Exploring formative assessment*. ASCD Publishers.

- Brookhart, S.M. (2010). *Formative assessment strategies for every classroom (2nd ed.)*. An ASCD Action Tool Publishers.
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65, 245-281. <https://doi.org/10.3102/00346543065003245>
- Card, N. A. (2012). *Applied meta-analysis for social science research*. The Guilford Press.
- Cauley, K.M. & McMillan, J.H. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 83 (1), 1-6. <https://doi.org/10.1080/00098650903267784>
- Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, 24(2), 205–249.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cohen, J. (1987). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.
- Delen, I., & Bellibas, M. S. (2015). Formative assessment, teacher-directed instruction and teacher support in Turkey: Evidence from PISA 2012. *Mevlana International Journal of Education*, 5(1), 88-102. <http://dx.doi.org/10.13054/mije.15.01.5.1>
- Double, K., McGrane, J. & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review*, 32 (2), 481–509. <https://doi.org/10.1007/s10648-019-09510-3>
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463. <https://doi.org/10.1111/j.0006-341X.2000.00455.x>
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109), 629-634.
- Field, A.P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63, 665–694. <https://doi.org/10.1348/000711010X502733>
- Fuchs, L., & Fuchs, D. (1986). Effects of systematic formative evaluation on student achievement: A meta-analysis. *Exceptional Children*, 53, 199-208. <https://doi.org/10.1177/001440298605300301>
- Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, 20, 304-315. <https://doi.org/10.1016/j.learninstruc.2009.08.007>
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Sage Publications.
- Graham, S., Hebert, M., & Harris, K.R. (2015). Formative assessment and writing. *The elementary school journal*, 115 (4), 523-547.
- Hargreaves, E. (2008). Assessment. In G. McCulloch, & D. Crook (Eds.), *The routledge international encyclopedia of education (pp. 37–38)*. Routledge Publishing.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 metaanalyses relating to achievement*. Routledge Publishing.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81-112. <https://doi.org/10.3102/003465430298487>
- Hedges, L.V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486-504. <https://doi.org/10.1037/1082-989X.3.4.486>
- Higgins, J., Thompson, S.G., Deeks, J.J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327, 557–560. <https://doi.org/10.1136/bmj.327.7414.557>

- Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28-37. <https://doi.org/10.1111/j.1745-3992.2011.00220.x>
- Kitchen, H., Bethell, G., Fordham, E., Henderson, K., & Li, R.R. (2019). *OECD reviews of evaluation and assessment in education: student assessment in Turkey*, OECD reviews of evaluation and assessment in Education, OECD Publishing. Retrieved August 17, 2021 from <https://doi.org/10.1787/5edc0abe-en>
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). Formative assessment and elementary school student academic achievement: A review of the evidence (REL 2017–259). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. Retrieved January 15, 2021 from <https://files.eric.ed.gov/fulltext/ED572929.pdf>
- Lee, H., Chung, H.Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K-12 education: systematic review. *Applied Measurement in Education*, 33(2), 124-140. <https://doi.org/10.1080/08957347.2020.1732383>
- Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Sage Publications.
- Maier, U., Wolf, N., & Randler, C. (2016). Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education*, 95, 85–98. <https://doi.org/10.1016/j.compedu.2015.12.002>
- McManus, S., Ed. (2008). *Attributes of effective formative assessment*. Council of Chief State School Officers.
- Miller, T. (2009). Formative computer-based assessment in higher education: the effectiveness of feedback in supporting student learning. *Assessment & Evaluation in Higher Education*, 34 (2), 181-192. <https://doi.org/10.1080/02602930801956075>
- MoNE (2013). Early childhood education program [Okul öncesi eğitim programı], Ministry of National Education, Ankara. <https://tegm.meb.gov.tr/dosya/okuloncesi/ooproram.pdf>
- MoNE (2017). The topics of in-service training activities in the last five years, 2012-2016 [Son 5 yılda düzenlenen (2012-2016) hizmetiçi eğitim faaliyetleri konuları], Ministry of National Education, Ankara.
- MoNE (2018). Geography curriculum [Coğrafya dersi öğretim programı], Ministry of National Education, Ankara. <https://mufredat.meb.gov.tr/Dosyalar/2018120203724482-Cografya%20dop%20pdf.pdf>
- MoNE (2020). Strengthening teacher capacity based on school and classroom-based assessment. Social Studies Lesson Teacher's Guide Booklet [Okul ve sınıf tabanlı değerlendirmeye dayalı öğretmen kapasitesinin güçlendirilmesi. Sosyal Bilgiler Dersi Öğretmen Rehber Kitapçığı], Ministry of National Education, Ankara
- Ozan, C., & Kıncal, R. Y. (2018). The effects of formative assessment on academic achievement, attitudes toward the lesson, and self-regulation skills. *Educational Sciences: Theory & Practice*, 18, 85–118. <http://dx.doi.org/10.12738/estp.2018.1.0216>
- Popham, W. J. (2008). *Transformative assessment*. Association of Supervision and Curriculum Development.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results, *Psychological Bulletin*, 86(3), 638-641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18 (2), 119–144. <http://dx.doi.org/10.1007/BF00117714>

- Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R*. Cham: Springer.
- Üstün, U., & Eryılmaz, A. (2014). A research methodology to conduct effective research syntheses. *Education and Science*, 39(174), 1-32.
- Weldmeskel, F.M., & Michael, D.J. (2016). The impact of formative assessment on self-regulating learning in university classrooms. *Tuning Journal for Higher Education*, 4 (1), 99-118. [https://doi.org/10.18543/tjhe-4\(1\)-2016pp99-118](https://doi.org/10.18543/tjhe-4(1)-2016pp99-118)
- Wiliam, D. (2018). Feedback: at the heart of –but definitely not all of–formative assessment. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 3–28). Cambridge University Press.
- Van der Kleij, F., Feskens, R., & Eggen, T.J.H.M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research* 85(4), 1-37. <https://doi.org/10.3102/0034654314564881>
- YÖKAK (2018). Yükseköğretim değerlendirme ve kalite güvencesi 2017 yılı durum raporu [2017 higher education evaluation and quality Assurance Status Report], Ankara: YÖK. https://yokak.gov.tr/Common/Docs/Site_Activity_Reports/2018DurumRaporuv2.pdf
- YÖKAK (2019). Yükseköğretim değerlendirme ve kalite güvencesi 2018 yılı durum raporu [2018 Higher Education Evaluation and Quality Assurance Status Report], Ankara: YÖK.
- Zimmerman, B. (2002). Becoming a self-regulated student: An overview. *Theory into Practice* 41(2), 64-70. https://doi.org/10.1207/s15430421tip4102_2
- Zimmerman, B., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal*, 31(4), 845-862. <https://doi.org/10.3102/00028312031004845>

6. APPENDIX

Studies included in the meta-analysis marked with an *

- *Arici, A.F., & Kaldirim, A. (2015). The effect of the process-based writing approach on writing success and anxiety of pre-service teachers. *Anthropologist*, 22(2), 318-327. <https://doi.org/10.1080/09720073.2015.11891883>
- *Atik, A.D., & Erkoç, F. (2017). The impact of formative tests on student achievement. *Journal of Theory and Practice in Education*, 13(4), 670-692.
- *Ayдын, S., Ural Keleş, P., & Ürün, N. (2016). Süreç değerlendirme yönteminin 7. Sınıf öğrencilerin güneş sistemi ve ötesi: uzay bilmecesi ünitesinde akademik başarıları ve kalıcılık düzeylerine etkisi [The effect of formative assessment on the achievement and retention levels of 7th grade students at the unit of solar system and beyond: mystery in space]. *Türk Eğitim Araştırmaları Dergisi (TURKEAD)*, 1(1), 11-17.
- *Batıbay, E.F. (2019). *Web 2.0 Uygulamalarının Türkçe dersinde motivasyona ve başarıya etkisi: kahoot örneği [The impact of Web 2.0 applications on motivation and success in Turkish course: the example of kahoot]* [Master's thesis]. Hacettepe University.
- *Bayat, N. (2014). The effect of the process writing approach on writing success and anxiety. *Educational Sciences: Theory & Practice*, 14(3), 1123-1141.
- *Bayrak, N., Çalık, M., & Doğan, S. (2019). The effects of smart formative assessment system on academic achievement and course process. *Hacettepe University Journal of Education*. Advance online publication. <https://doi.org/10.16986/HUJE.2019056742>.
- *Baysal, H. (2020). *Altıncı sınıf İngilizce dersinde kavram karikatürleri kullanımının öğrenci başarısına, konuşma becerisine ve motivasyonuna etkisi [The effect of using concept cartoons on students' achievement, speaking skill, and motivation in the sixth grade English]* [Master's thesis]. Balıkesir University.
- *Bolat, Y.İ., Şimşek, Ö., Ülker, Ü. (2017). Oyunlaştırılmış çevrimiçi sınıf yanıtlama sisteminin akademik başarıya etkisi ve sisteme yönelik görüşler [The impact of gamified online classroom response system on academic achievement and views about this system]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 17(4), 1741-1761.
- *Demirkese, B. (2019). *The effects of a mobile phone application on Turkish EFL students' grammar learning* [Unpublished master's thesis]. Necmettin Erbakan University.
- *Elvan, Ö. (2012). *Sosyal Bilgiler öğretiminde çalışma yaprakları kullanılmasının kavram yanlışlarını gidermeye etkisi [The effect of the usage of worksheets for resolving misconceptions in teaching social studies]* [Master's thesis]. Ahi Evran University.
- *Eraz, G., & Öksüz, C. (2015). Effect of primary school teachers' feedback on students' extracurricular mathematics activities. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, 36, 105-119.
- *Güzel, Z. (2018). *Fen bilimleri öğretiminde öz ve akran değerlendirme uygulamalarının yer aldığı probleme dayalı öğrenme yaklaşımının öğrencilerin başarı ve tutumlarına etkisi [The effects of problem based approach practiced through self and peer assessment on student achievement and attitudes in science teaching]* [Master's thesis]. Necmettin Erbakan University.
- *Hotaman, D. (2020). The effect of formative assessment on the academic achievement levels of prospective teachers. *Journal of Curriculum and Teaching*, 9(3), 33-44.
- *Kaya, B., & Ateş, S. (2016). The effect of process-based writing focused on metacognitive skills oriented to fourth grade students' narrative writing skill. *Education and Science*, 41(187), 137-164. <https://doi.org/10.15390/EB.2016.6752>

- *Kıncal, R.Y. & Ozan, C. (2018). Effects of formative assessment on prospective teachers' achievement, attitude and self-regulation skills. *International Journal of Progressive Education*, 14(2), 77-92. <https://doi.org/10.29329/ijpe.2018.139.6>
- *Korkmaz, Ö., Vergili, M., Çakır, R., & Uğur Erdoğan, F. (2019). Plickers Web 2.0 ölçme ve değerlendirme uygulamasının öğrencilerin sınav kaygıları ve başarıları üzerine etkisi [The impact of plickers Web 2.0 assessment and evaluation tool on exam anxiety and academic success of students]. *Gazi Eğitim Bilimleri Dergisi*, 5(2), 15-37. <https://dx.doi.org/10.30855/gjes.2019.05.02.002>.
- *Köksalan, S. (2019). *Sorgulamaya dayalı öğretimde kullanılan biçimlendirici değerlendirmenin öğrencilerin Fizik dersine yönelik tutumlarına ve kavramsal öğrenmelerine etkisinin incelenmesi* [Investigation of the effect of formative assessment used in inquiry-based instruction on students' attitudes towards physics lesson and conceptual learning] [Master's thesis]. Marmara University.
- *Kuzudişli, H. (2019). *Video-içi biçimlendirici değerlendirme ortamında öğrenen değerlendirme etkileşimlerinin incelenmesi* [Investigating of interaction between learner-assessment in the video formative assessment environment] [Master's thesis]. Hacettepe University.
- *Müldür, M., & Yalçın, A. (2019). Öz düzenlemeye dayalı yazma eğitiminin ortaokul öğrencilerinin bilgilendirici metin yazma becerisine, yazmaya yönelik öz düzenleme becerisine ve yazma öz yeterlik algısına etkisi [The effect of self-regulated writing instruction on middle school students' informative writing skills, self-regulated writing skills, and self-efficacy perception]. *Ilkogretim Online*, 18(4), 1779-1804. <https://dx.doi.org/10.17051/ilkonline.2019.639323>
- *Ozan, C., & Kıncal, R. Y. (2018). The effects of formative assessment on academic achievement, attitudes toward the lesson, and self-regulation skills. *Educational Sciences: Theory & Practice*, 18, 85–118. <http://dx.doi.org/10.12738/estp.2018.1.0216>
- *Özgür, P. (2016). Facebook sosyal ağına entegre e-portfolyo yazılımının akademik başarı ve öğretim sürecinde kullanımına yönelik tutuma etkisi [The effect of e-portfolyo software integrated to facebook social network on academic success and attitudes towards its use in teaching process]. *Sakarya University Journal of Education*, 6(1), 38-56.
- *Sever, E., & Memiş, A. (2013). Süreç temelli yazma modellerinin ilkökul dördüncü sınıf öğrencilerinin yazım–noktalama becerisine ve yazma eğilimine etkisi [The Effects Of Process-Based Writing Models On Primary School 4th Grade Students' Spelling-Punctuation Skills And Writing Dispositions]. *Karadeniz Sosyal Bilimler Dergisi*, 5(9), 243- 259.
- *Tavşanlı, F. (2019). *Süreç temelli yazma modüler programının ilkökul 2. sınıf öğrencilerinin yazmaya ilişkin tutum, yazılı anlatım becerisi ve yazar kimliği üzerine etkisi* [The effect of process writing modular program on 2nd grade elementary school students' towards attitudes, writing skills and their author identity] [Doctoral dissertation]. Uludağ University.
- *Topal, M. (2020). *Oyunlaştırma ile zenginleştirilmiş çevrimiçi öğrenmenin başarı, çevrimiçi bağlılık ve öğrenme motivasyonu üzerinde etkisi* [The effect of online learning enhanced with gamification on student's engagement to online learning environment, academic achievement and learning motivation] [Doctoral dissertation]. Sakarya University.
- *Turan, M.A., & Sakız, G. (2014). Fen ve teknoloji dersinde portfolyo kullanımının öğrenci başarısı ve kalıcılığa etkisi [The influence of portfolios on student success and retention level in science and technology class]. *Mersin University Journal of the Faculty of Education*, 10(3), 48-63.

- *Yalaki, Y. & Bayram, Z. (2015). Effect of formative quizzes on teacher candidates' learning in general chemistry. *International Journal of Research in Education and Science (IJRES)*, 1(2), 151- 156.
- *Yaşar, C. (2018). *Geri bildirim verilme zamanının matematik başarısına etkisi [The effect of feedback time on mathematics achievement]* [Master's thesis]. Hasan Kalyoncu University.
- *Yıldız, G., & Kılıç Çakmak, E. (2019). Zenginleştirilmiş e-değerlendirme sisteminin ders başarısına etkisi ve öğrenci memnuniyetinin incelenmesi [The effect of enriched e-assessment system on course success and review of student satisfaction]. *Gazi Journal of Education Sciences (GJES)*, 5, 106-139.
- *Yılmaz, N. (2015). Cebir öğretiminde yazma etkinliklerini kullanmanın ortaokul 7. sınıf öğrencilerinin başarılarına etkisi [The impact of using writing activities in teaching algebra on seventh grade middle school students' achievement]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 15(1), 356-376.
- *Yorgancı, S. (2015). Web tabanlı uzaktan eğitim yönteminin öğrencilerin matematik başarılarına etkileri [The effects of web based distance education method on students' mathematics achievements]. *Kastamonu Eğitim Dergisi*, 23(3), 1401-1420.
- *Yurdabakan, İ., & Cihanoğlu, M. O. (2009). Öz akran değerlendirmenin uygulandığı işbirlikli okuma ve kompozisyon tekniğinin başarı, tutum ve strateji kullanım düzeylerine etkisi. [The effects of cooperative reading composition technique with the applications of self and peer assessment on the levels of achievement, attitude, strategy use]. *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 11(4), 105-123.
- *Yurdabakan, İ., & Olgun, M. (2011). The influence of peer and self-assessment on learning and metacognitive knowledge: Consequential validity. *International Journal on New Trends in Education and Their Implications*, 2(4), 44-57.

Investigation of a Middle School Preservice Teacher's Knowledge of Content and Students

Ebru Ersari ^{1,*}

¹Balikesir University, Necatibey Faculty of Education, Department of Mathematics Education, Balikesir, Turkey

ARTICLE HISTORY

Received: June 01, 2021

Revised: Aug. 09, 2021

Accepted: Sep. 24, 2021

Keywords:

Mathematical Knowledge for Teaching,
Knowledge of Content and Students,
Preservice teacher education.

Abstract: The purpose of this study was to explicate one preservice middle grades mathematics teacher's Knowledge of Content and Students (KCS) in the context of multiple solution strategies. This study's purpose is to underline the importance of preservice teachers' KCS and provide possible investigative methods for evaluating preservice teachers' KCS. Specifically, the research inquiry guiding this study focused on how a middle school preservice mathematics teacher displays KCS when engaging with tasks about pattern recognition and linear functions in the context of multiple solution strategies. The data consisted of three videotaped semi-structured interviews with the preservice mathematics teacher as well as the written work she produced during the interviews. This study explicated one preservice mathematics teacher's performance regarding two important themes of KCS: generating multiple possible solution strategies of middle school students and explaining multiple student solution strategies. In terms of generating multiple solution strategies of middle school students, the study found that the preservice mathematics teacher provided the same solution strategies that she employed when she solved the problems by herself. Regarding explaining multiple student solution strategies, this study revealed that the preservice teacher did not explicate how typical middle school students reason. The preservice teacher had limitations when explaining the possible procedures that students might have used to solve problems when given the final student solutions. With regard to the teacher's abilities to recognise and understand students' typical understandings and misunderstandings, the study demonstrated that the preservice teacher was capable of explaining some solution strategies but not all of them.

1. INTRODUCTION

In the teacher knowledge literature, Shulman's (1986, 1987) categorization is deemed seminal. Shulman (1986) initially organized teacher knowledge into three categories: subject matter content knowledge, pedagogical content knowledge, and curricular knowledge. Pedagogical content knowledge includes the consideration of teaching the content to students and how the content makes sense from the perspective of the students. Moreover, it consists of the knowledge of how teaching one way might have potential pitfalls or advantages regarding students' perspectives and backgrounds. Shulman proposed that pedagogical content knowledge makes a content specialist different from a pedagogue. Drawing on this categorization in his 1986 essay, Shulman (1987) reframed the categorization of knowledge

*CONTACT: Ebru ERSARI ✉ ebru.ersari@balikesir.edu.tr 📧 Balikesir University, Necatibey Faculty of Education, Department of Mathematics Education, Balikesir, Turkey

and identified the seven categories of teacher knowledge as content knowledge, general pedagogical knowledge, curriculum knowledge, pedagogical content knowledge, knowledge of learners and their characteristics, knowledge of educational contexts, and knowledge of educational philosophies. Content knowledge, pedagogical knowledge, and curriculum knowledge are content-specific dimensions of teacher knowledge, whereas the other remaining categories are general dimensions of teacher knowledge and were not the primary focus of Shulman's work (Ball et al., 2008).

Taking Shulman's categorization of teacher knowledge, Ball et al. (2008) developed a model to explore the domains of Mathematical Knowledge for Teaching (MKT). They defined MKT as the "mathematical knowledge needed to carry out the work of teaching mathematics" (p. 395). According to Ball and her colleagues, MKT consists of two categories: Subject Matter Knowledge and Pedagogical Content Knowledge. Subject Matter Knowledge houses three subdomains: Common Content Knowledge (CCK), Knowledge at the Mathematical Horizon, and Specialized Content Knowledge (SCK). Pedagogical Content Knowledge consists of three subdomains: Knowledge of Content and Students (KCS), Knowledge of Content and Teaching (KCT), and Knowledge of Content and Curriculum.

According to Ball et al. (2008), the term Pedagogical Content Knowledge is used differently by various authors and has not been explored in depth. KCS, one component of Pedagogical Content Knowledge, is the combination of both the individual's knowledge of content and the individual's knowledge of students. Ball et al. (2008) explained KCS with "the example of analyzing a student error...[A] teacher might figure it out because she has seen students do this before with this particular type of problem" (p. 403). Teachers' familiarity with and knowledge of possible ways students think about the content is emphasized. The prior emphasis is on teachers' knowledge of students' thinking rather than their knowledge of content by itself. KCS focuses on teachers' knowledge about students' reasoning, how students understand the content, and what types of misconceptions students may have.

Studying KCS, one vital component of MKT, can reveal more accurate descriptions and measures of teachers' KCS, creating clearer distinction between different domains of MKT. This study aims to explore and underline why it is important to better understand preservice teachers' KCS. Preservice teachers need to be more familiar with possible student thinking before they actually begin teaching. In particular, in this study, I examined one preservice middle grades teacher's KCS in the context of three tasks involving linear functions.

1.1. Literature Review

Teaching mathematics is complicated (Boerst et al., 2011; Diez, 2010; Spalding et al., 2011), and teachers need to be responsive to students' mathematical reasoning when they teach (Dyer & Sherin, 2016; Jacobs & Empson, 2016; Thomas et al., 2017). However, studies pointed out that teachers cannot possibly give adequate explanations for every action each student takes when solving a mathematical problem (Nagle et al., 2017; Shin, 2020; Styers et al., 2020). This inability to predict or explain every action taken by a student exists regardless of whether the teacher is preservice (Nathan & Petrosino, 2002; Van Dooren et al., 2002) or experienced (Asquith et al., 2007; Gvozdic & Sadler, 2018). Nevertheless, research shows that teachers who have been trained to work through students' reasoning will be better prepared to notice trends in students' errors (Lee, 2021; Wuttke & Seifried, 2017). Even though teaching multiple ways of solving problems can be challenging for preservice teachers, using multiple solving strategies can have an impact on both high and low achieving preservice teachers by improving their problem solving skills (Gubermen & Leikin, 2013). Using multiple strategies will in turn reproduce both the standard solution method in the course and new solution methods (Leikin & Levav-Waynberg, 2008). Further, using multiple solutions have the potential to impact teacher knowledge. Mathematical Knowledge for Teaching (Ball et al., 2008), one of the

important categorization of teacher knowledge, is an important component of effective mathematics teaching (Bryan et al., 2007). However, researchers have examined MKT differently. Some scholars focused on teachers' overall MKT (Charalambous, 2010; Jacob et al., 2017; Steele & Rogers, 2012), others focused on a specific domain of MKT (Alqahtani & Powell, 2017; Bansilal et al., 2014; Johnson & Larsen, 2012), and others focused on more than one domain of MKT in their examinations (Hill, 2010; Lee et al., 2018; Ni Shuilleabhain, 2016). Ball et al. (2008) distinguished KCS from the other domains of MKT as follows:

consider what is involved in selecting a numerical example to investigate students' understanding of decimal numbers. The shifts that occur across the four domains, for example, ordering a list of decimals (CCK), generating a list to be ordered that would reveal key mathematical issues (SCK), recognizing which decimals would cause students the most difficulty (KCS), and deciding what to do about their difficulties (KCT), are important yet subtle. (p. 404)

KCS will both inform teachers' lessons and instructional methods to preempt students reasoning errors before they become ingrained patterns (Johnson & Larsen, 2012). The ability to understand the nature of students' reasoning errors should be an aspirational goal and core component of teaching. KCS will increase the effectiveness of teacher instruction by helping students understand the mathematical principles and their errors in applying those principles to mathematical problems (Lannin et al., 2007). Different ways of explaining concepts can help different students conceptualize ideas (An et al., 2004). When students were prompted to use multiple solutions, they became more interested in mathematics. This new interest in turn led to greater student mathematical competencies (Schukajlow & Krug, 2014). Multiple teaching approaches are better than one (Guberman & Leikin, 2013) because diverse lesson delivery methods and explanatory approaches should engage a larger number of students and allow them to process information to become better mathematical thinkers. This focus on developing an awareness of multiple ways of teaching mathematical content will in turn address a common teaching tendency, the teacher's reliance on their own personal reasoning strategies as the basis for their lessons (Peterson & Treagust, 1995). Consequently, this study employs KCS because the processes preservice teachers employ in their own problem-solving emerge in their teaching, whereas established teachers employ problem-solving methods learned from exposure to actual student solutions. Studies focusing on teachers' knowledge on students' reasoning found that teachers with limited knowledge lack the ability to listen actively to their students (Johnson & Larsen, 2012), pose problems (Lee et al., 2018), interpret or answer students' responses and questions (Edelman, 2017), and predict students' reasoning (Asquith et al., 2007; Norton et al., 2011). One reason for their lack of knowledge can be the excessiveness of goals in teacher preparation programs that can exacerbate the tendency to omit multiple solution strategies in teachers' pedagogies (Hiebert & Berk, 2020).

In terms of generating and explicating multiple solution strategies, Silver et al. (2005) found that teachers have cognitive (e.g., insecurity) and pedagogical (e.g., teaching difficulty) concerns regarding using multiple solution strategies. Interactive and reflective solutions can enhance teachers' understanding of students' multiple solution strategies (Leikin & Levav-Waynberg, 2007). Having a deeper understanding of both the content and students reasoning abilities are crucial aspects of those solution strategies (Taşdan & Çelik, 2016). There should also be a clear focus on explaining mathematical concepts both procedurally and conceptually. Hiebert and Lefevre (1986) defined conceptual knowledge with an emphasis on relationships as: "a connected web of knowledge, a network in which the linking relationships are as prominent as the discrete pieces of information. Relationships pervade the individual facts and propositions so that all pieces of information are linked to some network" (pp. 3-4). They described procedural knowledge with an emphasis on its two kinds: "... a familiarity with the individual symbols of the system and with the syntactic conventions for acceptable

configurations of symbols... [and] ... rules or procedures for solving mathematical problems” (p. 7). Conceptual and procedural knowledge constructs are generally referred to by mathematics educators as “qualities of knowledge” and by psychologists as “types of knowledge” (Star, Stylianides, 2013, p. 15). Rittle-Johnson et al. (2015) claimed that there is a bidirectional relationship between conceptual and procedural knowledge that both types of knowledge supports each other.

Even though KCS is a vital component of teacher knowledge, there is scarce research investigating preservice mathematics teachers’ KCS (Sitrava, 2020). One of the main reasons for this lack of critical data can be traced to the difficulty of writing KCS items. This difficulty is likely due to the fact that there is no common conceptual understanding and shared definition among researchers regarding what, exactly, KCS is (Hill et al., 2008). Another reason why there is limited research on this subject is because researchers’ have great difficulty findings sample KCS items. With a greater sample of KCS items, researchers could have a better understanding of how to write new KCS items and utilize those items in their research. These sample items predominantly rely on the multiple-choice format. More items with open-ended questionnaires or video interviews could aid resarchers in developing more diverse KCS items. Also, the difficulty involved in distinguishing between the MKT domains could be a contributing factor in the scarcity of data on the topic (Ball et al., 2008).

1.2. Research Questions

The purpose of this study was to scrutinize one preservice teacher’s Knowledge of Content and Students (KCS) using pattern recognition and linear function tasks. To reach this objective, this study investigated the preservice teacher’s knowledge on generating and explaining possible student thinking. Specifically, the research inquiry guiding this study focused on how a middle school preservice mathematics teacher displays KCS when engaging with tasks about pattern recognition and linear functions in the context of multiple solution strategies.

2. METHODOLOGY

2.1. Participant

The study reported in this article is a part of a broader study. The population of the broader study was middle grades preservice mathematics teachers studying at a southern university in the United States. Eight middle grades preservice mathematics teachers volunteered to participate in the study. Convenience sampling (Patton, 2002) was utilized, and four volunteers agreed to participate in the broader study. Selection was based on participants’ schedule availability for the broader study rather than their levels of subject knowledge. Pseudonyms were used for each participant. In this study, the focus was one of the four volunteer middle school preservice teachers (Megan). Megan was selected because she was talkative, and her interviews provided the richest information in terms of preservice teachers’ potential KCS limitations. Time and participant availability necessitated the study only focus on one preservice teacher. For a comparative study, the inservice and preservice teachers’ KCS can be compared to strengthen the findings.

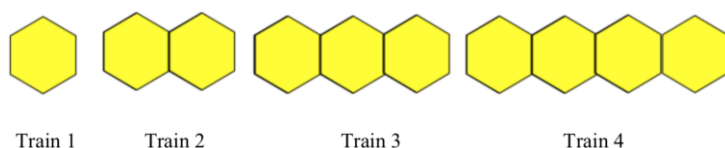
2.2. Data Collection

The data consisted of three videotaped semi-structured interviews (Maxwell, 1996) with Megan as well as the written work she produced during the interviews. Each interview took approximately one and a half hours; therefore, these interview data were supplemented by approximately 5 hours of videotaping, and all of them were transcribed.

Each of the three interviews were conducted around one mathematical task that was selected from the National Council of Teachers of Mathematics (NCTM) teaching and learning modules as part of an online toolkit aligned with the teaching practices contained in *Principles to Actions*

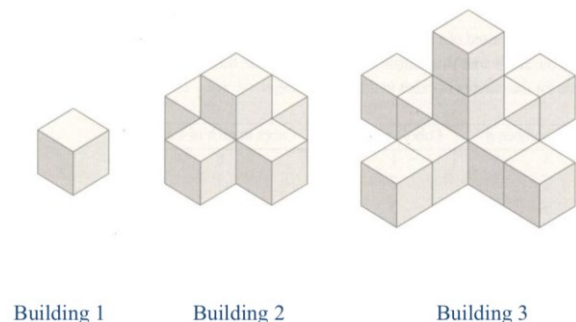
(NCTM, 2014). All the tasks that were selected involved linear relationships. While this presents a limitation in that this study does not have evidence of the participant's KCS in other mathematics topic areas, it has more detailed information about her thinking regarding linear functions. The first interview focused on a Hexagon task. The visual representation of the Hexagon task can be accessed at <https://www.nctm.org/Conferences-and-Professional-Development/Principles-to-Actions-Toolkit/The-Case-of-Patrica-Rossmann-and-the-Hexagon-Task/>. In the Hexagon task, the preservice teacher was asked to find the patterns of the perimeters of trains constructed with regular hexagons. The first four trains, consisting of hexagonal wagons, were visually demonstrated in the problem. The first train consists of one, the second train of two, the third train of three, and the fourth train of four hexagonal wagons. Subsequent hexagons were added linearly to the right edge of the preceding hexagon. The first four trains consisting of hexagonal wagons were visually demonstrated in the problem as follows:

Figure 1. Visual depiction of the the Hexagon task.

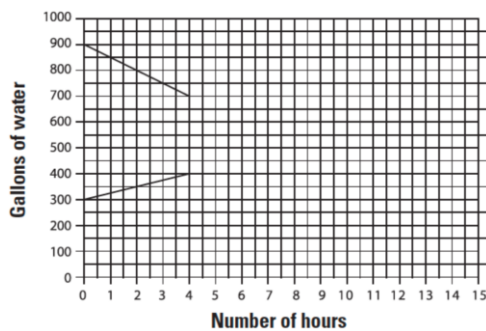


The second interview focused on a Counting Cubes task. The visual representation of the Counting Cubes task can be accessed at <https://www.nctm.org/Conferences-and-Professional-Development/Principles-to-Actions-Toolkit/The-Case-of-Peter-Dubno-and-the-Counting-Cubes-Task/>. In the Counting Cubes task, Megan was asked to elaborate on the patterns of several cubes. The cubes represent buildings consisting of five extensions. Each subsequent building adds an additional cube for each extension of the building. In the task, three buildings were represented visually. Building 1 consists of one cube only; building 2 consists of 5 extensions, an extension on each of the five faces of the cube in building 1; building 3 adds an additional cube in each direction. The first three buildings were shown in the problem as follows:

Figure 2. Visual depiction of the Counting Cubes task.



The third interview focused on a Two Storage Tanks task. The visual representation of the Two Storage Tanks task can be accessed at <https://www.nctm.org/Conferences-and-Professional-Development/Principles-to-Actions-Toolkit/The-Case-of-Elizabeth-Brovey-and-the-Two-Storage-Tanks-Task/>. In the Two Storage Tanks task, Megan was asked to read a graph and find the amount of water in two storage tanks, one losing water and the other gaining water at different rates. The amount of water in both tanks over a period of time was demonstrated with the number of hours presented on the x axis and gallons of water in the tank presented on the y axis. The graph was provided in the problem as follows:

Figure 3. Visual depiction of the Two Storage Tanks task.

Specialized Content Knowledge (SCK) items developed by Hiebert et al.'s (2019) were modified for the interview protocol of this study, and their rubrics were adapted to evaluate the participant's KCS. Hiebert et al. (2019) specifically focused on three topics: multiplying two-digit whole numbers, subtracting fractions, and dividing fractions. This study focused on two topics: linear functions and pattern recognition. These sets of topics have transferrable qualities and shared concepts like integers, functions, and variables. Also, the relationship between quantities is the focus for both studies' topics; therefore, both studies are matching. The only widely known bank of KCS items is a set developed by Ball and colleagues, but those items are not available for public use. Although Hiebert et al.'s (2019) SCK items were developed for use with elementary school teachers (preservice and in-service), the basic structure of the items was transferrable to a middle grade context. Namely, this study drew heavily on Hiebert et al.'s (2019) SCK items and Ball et al.'s (2008) distinction of SCK and KCS for constructing KCS items and rubrics for this study. Hiebert et al.'s (2019) study used three-point rubrics 0 meaning no knowledge to 2 meaning extensive knowledge. This study also used three-point rubrics 0 referring to limited or lack of explanation; 1 referring to having some valid explanation but partial explanation; 2 having an adequate and elaborated explanation. Scoring rubrics are provided in the [Appendix](#).

2.3. Exploring KCS of the Preservice Teacher

To compare and contrast the participant's KCS, her performance on tasks requiring her to 1) generate multiple solution strategies and 2) explain multiple solution strategies was evaluated. To generate multiple solution strategies, Megan was asked to generate three correct solution strategies she thought middle school students would likely use for each of the three tasks. In the generating multiple solution strategies part on the Hexagon task, the preservice teacher was asked to show and explain three different ways that students could correctly solve the perimeter of any train in the pattern; on the Counting Cubes task, three different ways that students could correctly find the number of cubes in the n^{th} building; on the Two Storage Tanks task, three different ways that students could correctly find the time at which the two tanks contain the same amount of water.

The Generating Multiple Solution Strategies part was used for exploring the preservice teacher's KCS and required her to predict and anticipate the typical middle school students' solution strategies. To evaluate how accurately and elaborately the participant generated multiple solution strategies, her responses to the Generating Multiple Solution Strategies theme of all of the three tasks were investigated. To determine how accurately and elaborately the participant explained multiple solution strategies, her responses to the Explaining Multiple Solution Strategies theme of the Hexagon and Counting Cubes tasks were focused. Students' complete work was not given on these two tasks and the preservice teacher was asked to anticipate students' work. The preservice teacher was given three hypothetical students' final answers without providing the answers' solutions. She was asked to explain the possible

procedures that each hypothetical student might have come up with for each of the three solutions. On the two tasks, the focus was on students' reasoning or challenges while examining students' work. If the task did not include the steps of the solutions but required the preservice teacher to predict the steps, the task was regarded as exploring the participant's KCS. In other words, for the Hexagon and Counting Cubes tasks, the hypothetical student work consisted only of hypothetical students' responses and required the participant to anticipate possible solution strategies, which aligned with KCS. Rachel, Sam, and Jason are the hypothetical students whose responses were provided in the Hexagon task. David, Emily, and Mary are the hypothetical students whose responses were provided in the Counting Cubes task. In the Hexagon task, Rachel's response was given as $2n + 2n + 2$ or $2(2n) + 2$; Sam's response was given as $5 + 4(n - 2) + 5$ or $4(n - 2) + 5 + 5$, or $4(n - 2) + 10$; and Jason's response was given as $6n - 2(n - 1)$ as the perimeter of any train in the pattern. In the Counting Cubes task, David's response was given as $n + 4(n - 1)$; Emily's response was given as $1 + 5(n - 1)$, and Mary's response was given as $5n + 1$ as the number of cubes in the n^{th} building. For the Two Storage Tanks task, the hypothetical students' work was already given, the preservice teacher was not required to predict student work, and asked to explain student work mathematically, which aligned with SCK and is not the focus of this study.

3. RESULTS / FINDINGS

In this section, the findings of the participant's KCS on the tasks were presented regarding both the Generating Multiple Solution Strategies and the Explaining Multiple Solution Strategies themes. First, the participant's performance on the Hexagon, Counting Cubes, and Two Storage Tanks tasks were presented in the context of the Generating Multiple Solution Strategies theme. Preservice teacher's performance patterns on the tasks regarding the Generating Multiple Solution Strategies theme were then presented. Next, the participant's performance patterns on the Hexagon and Counting Cubes tasks were presented in the context of the Explaining Multiple Solution Strategies theme. Finally, the preservice teacher's performance patterns on the tasks in the Explaining Multiple Solution Strategies theme were presented.

3.1. Generating Multiple Solution Strategies Theme

3.1.1. Hexagon task

Even though Megan was asked to explain three different ways that students could solve the perimeter of any train in the pattern, she provided only two possible student solution strategies. Megan considered the students' potential solutions to the Hexagon task as follows: 1) finding the perimeter of each hexagon first and then subtracting the shared sides and 2) finding the perimeter of a set of hexagons (i.e., grouping 2 trains as a set or grouping 4 trains as a set) and then excluding the number of shared sides between sets of hexagons. Following is her response regarding students' possible first strategy:

Figure 4. Megan's response regarding students' possible first strategy on the Hexagon task.

Show/explain three different ways that students could **correctly** solve the perimeter of any train in the pattern.

1. One way

- find Peremiter of each train
 - Subtract the # of sides shared w/ another hexagon

In the first strategy, the hexagons were considered separately, whereas in the second strategy, the hexagons were considered as a group. Following is her response regarding students' possible second strategy:

Figure 5. Megan's response regarding students' possible second strategy on the Hexagon task.

2. A second way

- find a set perimeter of trains (whole)
- divide up trains based on that whole
- finding # of sides shared

In her responses regarding students' possible strategies, Megan did not mention whether students could possibly find the relationship between the number of trains and the perimeter of hexagons in the task or not. In her own solution strategies, Megan grouped the hexagons by 2 hexagons in her third strategy and by 4 hexagons in her fourth strategy and then excluded the sides that were shared between the group of hexagons.

4.1.2 Counting cubes task

Even though Megan was asked to explain three different ways that students could solve the perimeter of any train in the pattern, similar to the Hexagon task, she provided only two possible student solution strategies. The following shows what Megan considered as students' possible solutions in the Counting Cubes task: 1) using the expression $5n - 4$ and 2) using the expression $5(n - 1) + 1$. Following is her explanation of students' possible first strategy:

Figure 6. Megan's explanation of students' possible first strategy on the Counting Cubes task.

Show/explain three different ways that students could correctly find the number of cubes in the n^{th} building.

1. One way

- looking at each face of the cube
- finding where the ⁵ new cubes will go in the next building

$5n - 4$

In terms of the first possible student strategy, Megan described the strategy as “looking at each face of the cube and finding where the 5 new cubes will go in the next building.” The first student solution strategy that she came up with was the same as her first strategy. She was not clear what -4 represented when she described using expression $5n - 4$ in both of her own solution strategy and the students' possible solution strategies. She was incorrectly considering that each cube had 4 open faces and -4 represented those missing cubes that could come next to each of the open faces. She did not recognize that the last cubes in the extensions had 5 open faces as well as that -4 did not represent the cubes that could come next to the open faces. Her explanation of the second possible student solution strategy is as follows:

Figure 7. Megan's explanation of students' possible second strategy on the Counting Cubes task.

2. A second way

$$5(n-1) + 1$$

- looking at 5 different extensions and counting the # of cubes in each extension then adding the 1 cube in the middle.

Megan described the second possible student strategy, using the expression $5(n - 1) + 1$, as “looking at 5 different extensions and counting the number of cubes in each extension then adding the 1 cube in the middle.” This second student solution strategy was the same as her own second strategy, and Megan assumed that students would use the same methods she did.

3.1.3. Two storage tanks task

Different from the Hexagon and the Two Storage Tanks tasks, Megan provided three possible students' solutions to the Two Storage Tanks task as follows: 1) using $y = mx + b$ for finding the equations of both lines and finding the y value for the same x value by plugging in different x values until finding the same x and y values and 2) using $y = mx + b$ for finding the equations of both lines, set them equal to each other, and find the intersection point, 3) extending the lines and finding the intersection point.

Her explanation of the first students' possible solution strategy is as follows:

Figure 8. Megan's explanation of students' possible first strategy on the Two Storage Tanks task.

Show/explain three different ways that students could correctly solve the Two Storage Tanks task.

1. One way

- Using " $y = mx + b$ " equation same for both lines & finding the y value ~~for~~ that both equations have for the for the same x value

Megan thought using the equations $y = -50x + 900$ for tank T and $y = 25x + 300$ for tank W and plugging in numbers until getting the same x and y value as the first solution strategy students would likely try. She, however, considered that this strategy was too hard to apply if the x value was a big number.

Her explanation of the second possible student solution strategy is as follows:

Figure 9. Megan's work on students' possible second strategy on the Two Storage Tanks task.

2. A second way

- Using $y = mx + b$ & Systems of equations

As a more systematic way to solve the problem, Megan explained the second solution strategy as finding the equations of each line using $y = mx + b$ and using systems of equations to find the x and y values for the intersection point.

Her explanation of the third possible student solution strategy is as follows:

Figure 10. Megan’s work on students’ possible third strategy on the Two Storage Tanks task.

3. A third way

- continue to graph each line out
 - find intersection point

The third solution strategy Megan thought students might try was extending each graphed line until they meet then finding the intersection point.

3.1.4. Patterns among the tasks regarding Generating Multiple Solution Strategies theme

Megan provided correct expressions as possible student solutions. However, she could not explain what some numbers and variables referred to in the expressions that she provided. She also explained some of the meanings of the variables inaccurately. For instance, she related -4 with the open faces of the cubes even though -4 was not related with the number of open faces. Even though Megan grappled with explaining the meaning of -4 in the expression $5n - 4$, she still reported as one possible solution strategy. Her challenges when solving the tasks herself were similar to her challenges when explaining the possible student solutions. Also, Megan did not offer explanations about which solutions would be easier or more difficult for students.

Megan showed some possible student solution strategies, but her student solutions lacks variety for different levels of students. She provided similar strategies to the ones she came up with when solving the tasks herself. However, she could not relate the student solution strategies that she generated to middle school students’ reasoning. She did not hypothesize which strategy might have been more common and which strategy might have been less common among middle school students and did not provide reasons. She explained procedurally what students might possibly have done; however, she did not explain conceptually what they might have done.

3.2. Explaining Multiple Solution Strategies Theme

3.2.1. Hexagon task

Megan first explained Jason’s response because it was similar to her solution. Following is her explanation of Jason’s response:

Figure 11. Megan’s explanation of Jason’s response.

Jason’s response: $6n - 2(n - 1)$

- find perimeter of 1 hexagon
 Subtract Side being shared

Megan described Jason’s response, $6n - 2(n - 1)$, as finding the perimeter of one hexagon and subtracting $2(n - 1)$, which represented the number of sides being shared. Next, Megan explained Sam’s response. Megan did not write down any explanation; however, she underlined some parts of his work. Following shows Megan’s work on Sam’s response:

Figure 12. Megan’s work on Sam’s response.

Sam’s response: $5 + 4(n - 2)$ + 5 , or $4(n - 2)$ + $5 + 5$, or $4(n - 2)$ + 10

In Sam's response, $5 + 4(n - 2) + 5$, or $4(n - 2) + 5 + 5$, or $4(n - 2) + 10$, she stated that 5 was the number of sides in the first hexagon and the other 5 was the number of sides in the last hexagon; $4(n - 2)$ was number of the sides in the middle hexagon, where 4 was the number of sides that were on the perimeter for each middle hexagon and $n - 2$ was the number of hexagons in the middle.

Even though Rachel's response was written first among other possible solution strategies, Megan preferred to explain Rachel's response last.

Figure 13. Megan's work on Rachel's response.

Following are the answers of Rachel, Sam, and Jason. Explain the possible procedure(s) that each student might come up with these solutions.

Rachel's response: $2n + 2n + 2$ or $2(2n) + 2$

$$\begin{array}{l} 2(4) + 2(4) + 2 \\ 8 + 8 + 2 \\ 16 + 2 = 18 \end{array} \quad \begin{array}{l} 2(3) + 2(3) + 2 \\ 6 + 6 + 2 \\ 12 + 2 \\ 14 \end{array}$$

In Rachel's response, $2n + 2n + 2$ or $2(2n) + 2$, Megan struggled to explain what 2s referred to in the expressions. She checked that Rachel's expression was correct by plugging in 3 and 4 for n in the expression and finding the correct perimeters. She stated in terms of Rachel's response that, "My only thing that I think I can come up for her is that she is taking a hexagon and saying that is $2n$. And so, then she is adding up $2n + 2n$ and then 2 would be the sides in the middle." When she was asked what $2n$ represented, she explained that "I know n definitely represents the number of trains, and then this is just another way of showing this part of the equation of subtracting the sides out [showing $-2(n - 1)$ in Jason's response]." Because Megan said that $-2(n - 1)$ was the number of shared sides when she was explaining Jason's response, she was asked if $2n$ in Rachel's response represented the shared sides. She explained that, "No, I think she is representing that in a different way, but it is very similar to the $2n$ down here, but it is represented in a different way, to me, not as clear as Jason's responses." She asserted that none of the 2s in Rachel's explanation was clear, and Rachel needed more explanation regarding what those numbers referred to. Ultimately, Megan was still struggling to explain what the numbers in Rachel's representation referred to.

3.2.2 Counting cubes task

Because Emily's response $1 + 5(n - 1)$ was similar to her own response, Megan first started explaining Emily's solution by describing 1 as the 1 cube in the middle, 5 as the number of extensions, and $n - 1$ as the number of cubes in each extension. Following is Megan's work on Emily's response:

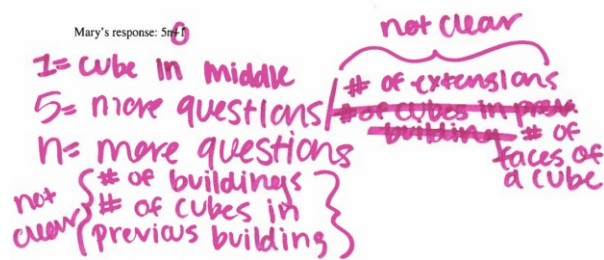
Figure 14. Megan's work on Emily's response.

Emily's response: $1 + 5(n - 1)$

$$\begin{array}{l} 1 = 1 \text{ cube in middle} \\ 5 = \# \text{ of extensions} \\ n - 1 = \# \text{ cubes in each extension} \end{array}$$

Then, Megan described Mary's response, $5n + 1$, describing 1 as the cube in the middle. Following is Megan's comments on Mary's response:

Figure 15. Megan’s work on Mary’s response.



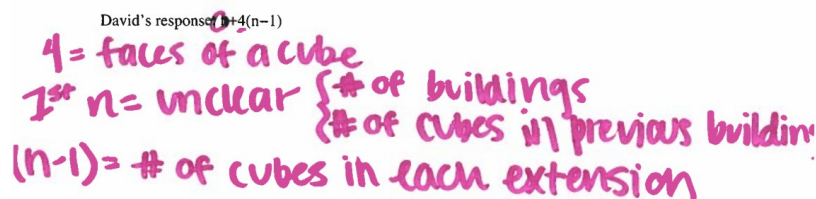
Megan asserted that the meaning of 5 and n was not clear in Mary’s expression because 5 could be either the number of extensions or the number of faces of a cube and n could be either the number of buildings or the number of cubes in the previous building. She did not recognize that n could be the number of cubes in each extension excluding the middle cube and +1 was the cube in the middle. Her description of what 5 and n could mean again shows that she conflated the number of cubes in a building with the number of open faces in a cube. Megan’s lack of knowledge on the task impacted her incorrect assumptions about Mary’s response.

In terms of David’s response $n + 4(n - 1)$, Megan thought that 4 was the number of open faces of a cube; $(n - 1)$ was the number of cubes in each extension. Following is Megan’s comments on David’s response:

Figure 16. Megan’s work on David’s response.

Task 3b: Explain Multiple Solution Strategies

Following are the answers of David, Emily, and Mary. Explain the possible procedures that each student might come up with these solutions.



She said the meaning of the first n was not clear and could refer to either the number of buildings or the number of cubes in the previous building. As Megan counted the number of faces of a cube in her own solution, she could relate David’s response with her own solution. Ultimately, Megan struggled to explain the meaning of the numbers in both David’s and Mary’s responses.

3.2.3. Patterns among the tasks regarding explaining multiple solution strategies theme

Megan preferred to explain the solutions that she was familiar with first and the solutions that she struggled with the most last. Megan struggled to explain some of the students’ solution strategies. For instance, Megan failed to explain what the 2s referred to in the expression $2n + 2n + 1$ or $2(2n) + 2$ in the Hexagon task. Megan sometimes represented the variables in student solutions inaccurately. For instance, she thought that 4 in the expression $n + 4(n - 1)$ in the Counting Cubes task represented the number of faces of a cube. However, 4 in the expression $n + 4(n - 1)$ represented the number of extensions.

Megan provided explanations of some of the student solution strategies. However, she had limited conceptual understanding of the solution strategies. She preferred to start explaining students’ solutions based on the ones that were similar to her own strategy.

4. DISCUSSION and CONCLUSION

In the teacher knowledge literature, Shulman's (1986, 1987) categorization of teacher knowledge is prominent. Shulman's seven categories provide the basis for numerous explanations of teacher knowledge. Developing Shulman's influential teacher knowledge categorization, Ball et al. (2008) constructed MKT. KCS, one of the domains of MKT, plays an important role in shedding light on teachers' knowledge. In spite of the scarcity of research on KCS, studies have shown that teachers lack KCS (Edelman, 2017; Johnson & Larsen, 2012; Lee et al., 2018). The purpose of this study is to underline the importance of the investigation of preservice teachers' KCS. Specifically, this research focused on a middle school preservice teachers' KCS when engaging with tasks about pattern recognition and linear functions in the context of multiple solution strategies. Generating multiple solution strategies and explaining multiple solution strategies are the two themes this study uses to explore the preservice teacher's KCS.

In terms of predicting students' thinking and confusion, the preservice teacher, Megan, provided the same solution strategies that she provided when she solved the problems by herself. This study shows that the preservice teacher's performance on predicting students' reasoning might depend on her own knowledge. In general, Megan's student predictions mirrored her own solutions to the tasks. Similar to this study, Norton et al. (2011) found that there can be a relationship between prospective elementary teachers' prediction of students' work and their own mathematical knowledge. Regarding the acquaintance with students' mathematical reasoning, Megan did not explicate how typical middle school students reason. This finding is consistent with Asquith et al.'s (2007) finding that teachers have difficulties predicting students' understanding and reasoning. In Asquith et al.'s (2007) study, middle school teachers struggled to predict students' understanding of the equal sign and variable, whereas in this study, the preservice teacher had difficulties predicting middle school students' possible reasoning about pattern recognition and linear functions. With regard to students' typical understandings and misunderstandings, Megan was able to explain some solution strategies, but she could not explain others. For instance, Megan struggled to explain how the student produced the expression $2n + 2n + 2$ or $2(2n) + 2$ for finding the perimeter of the hexagons in the Hexagon task. Megan stated that she did not know where any of the 2s come from in the expression $2n + 2n + 2$ or $2(2n) + 2$. She could not explain that in this strategy, the student considers the tops of the hexagons as two times the train number and the bottoms. Also, Megan could not describe that since there are two sides on top of each hexagon, the number of top sides on any train is $n \times 2$ (n hexagons \times 2 top sides per hexagon) or $2n$. Similarly, she did not explain that the number of bottom sides is also $2n$ in the student's solution.

Regarding multiple solution strategies, Silver et al. (2015) showed that veteran middle grades mathematics teachers were also concerned about explaining multiple solution strategies to their students. Those veteran teachers stated that some students might have had difficulties understanding different ways of solving problems. Therefore, multiple solution strategies can be challenging for teachers in terms of their ability to solve problems in multiple ways by themselves, as well as their ability to explain multiple solutions to students with limited understanding. Similar to this study, the teacher in the study of Johnson & Larsen (2012) had constraints on understanding her students' struggles. Different from this study, the participant in Johnson & Larsen's study (2012) had higher content knowledge. In their study, the teacher was a mathematician and got his PhD in mathematics. Therefore, this study and Johnson & Larsen's (2012) study showed that teachers might struggle to understand from students' perspectives no matter how knowledgeable they are in terms of the content. In the Counting Cubes task example, Megan was able to explicate some solution strategies, but she could not explain some other solution strategies and provided inaccurate explanations regarding what the

variables could represent. For instance, when she explained the expression $5n + 1$ in the Counting Cubes task, she said that 5 could be either the number of extensions or the number of faces of a cube. However, 5 did not represent the number of faces of a cube in the expression. When Megan explained the expression $n + 4(n - 1)$, she thought that 4 was the number of faces on a cube and that the first n could be either the number of buildings or the number of cubes in the previous building. However, neither 4 represented the number of faces of a cube, nor did n represent the number of buildings.

In sum, this study explicated one preservice teacher's performance regarding two important themes of KCS: generating multiple solution strategies and explaining multiple solution strategies. More research is needed to construct KCS items to elaborate more on what constitutes teachers' KCS. Also, teachers' performance on the components of KCS requires more investigation. In this study, no professional development or interventions were employed. Thus, the impact of professional development on KCS can be explored. More research is needed to understand how to improve teachers' KCS. Additionally, preservice teachers' KCS can be explored further for different content and different grade levels with more participants by using more tasks. Also, the relationship between other domains of teacher knowledge can be investigated. How KCS impacts preservice and in-service teachers' teaching strategies could be examined. In order to improve the quality of teacher education programs, future studies can explore the relationship between preservice teachers' knowledge and their teaching strategies. Also, researchers need to conduct studies using well-designed KCS tasks. For this purpose, they can collaborate together to develop KCS tasks for their future studies and produce KCS items that are publicly available.

Acknowledgements

This study is based on the dissertation study conducted by the author. I want to thank my advisor Dr. Denise Spangler and my committee members Dr. Anna Conner and Dr. Steve Oliver for their guidance. I also would like to thank Dr. Andrew Izsák for his feedback. I owe special thanks to the participant, Megan, who accepted to be part of this study. Additionally, I want to thank the Ministry of National Education of the Republic of Turkey for supporting me financially during my study in the United States.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. All scientific responsibility of the manuscript belongs to the author. Institutional Review Board reviewed and approved that the study is aligned with research publication ethics (IRB ID: STUDY00006215, Approval Date: 21/06/2018).

ORCID

Ebru ERSARI  <https://orcid.org/0000-0002-0324-3185>

5. REFERENCES

- Alqahtani, M. M., & Powell, A. B. (2017). Mediation activities in a dynamic geometry environment and teachers' specialized content knowledge. *The Journal of Mathematical Behavior*, 48, 77-94.
- An, S., Kulm, G., & Wu, Z. (2004). The pedagogical content knowledge of middle school, mathematics teachers in China and the US. *Journal of Mathematics Teacher Education*, 7(2), 145-172.
- Asquith, P., Stephens, A. C., Knuth, E. J., & Alibali, M. W. (2007). Middle school mathematics teachers' knowledge of students' understanding of core algebraic concepts: Equal sign and variable. *Mathematical Thinking and Learning*, 9(3), 249-272.

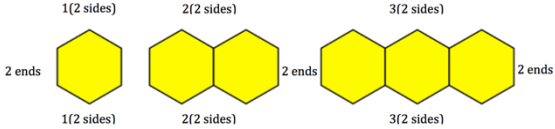
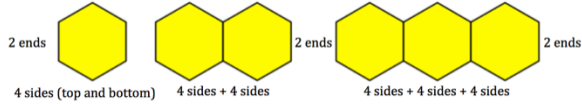
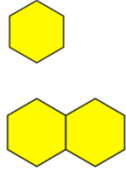
- Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching. *Journal of Teacher Education*, 59(5), 389-407.
- Baki, A. (2020). *Matematiği öğretme bilgisi* (3rd ed.). PegemAkademi
- Bansilal, S., Mkhwanazi, T., & Brijlall, D. (2014). An exploration of the common content knowledge of high school mathematics teachers. *Perspectives in Education*, 32(1), 34-50.
- Boerst, T. A., Sleep, L., Ball, D. L., & Bass, H. (2011). Preparing teachers to lead mathematics discussions. *Teachers College Record*, 113(12), 2844-2877.
- Bryan, C. A., Wang, T., Perry, B., Wong, N., & Cai, J. (2007). Comparison and contrast: Similarities and differences of teachers' views of effective mathematics teaching and learning from four regions. *ZDM*, 39(4), 329-340.
- Charalambous, C. (2010). Mathematical knowledge for teaching and task unfolding: An exploratory study. *The Elementary School Journal*, 110(3), 247-278.
- Diez, M. E. (2010). It is complicated: Unpacking the flow of teacher education's impact on student learning. *Journal of Teacher Education*, 61(5), 441-450.
- Dyer, E. B., & Sherin, M. G. (2016). Instructional reasoning about interpretations of student thinking that supports responsive teaching in secondary mathematics. *ZDM*, 48(1-2), 69-82.
- Edelman, J. (2017). How preservice teachers use children's literature to teach mathematical concepts: focus on mathematical knowledge for teaching. *International Electronic Journal of Elementary Education*, 9(4), 741-752.
- Guberman, R., & Leikin, R. (2013). Interesting and difficult mathematical problems: changing teachers' views by employing multiple-solution tasks. *Journal of Mathematics Teacher Education*, 16(1), 33-56.
- Gvozdic, K., & Sander, E. (2018). When intuitive conceptions overshadow pedagogical content knowledge: Teachers' conceptions of students' arithmetic word problem solving strategies. *Educational Studies in Mathematics*, 98(2), 157-175.
- Hiebert, J., & Berk, D. (2020). Foreword: Building a profession of mathematics teacher education. *The Mathematics Enthusiast*, 17(2), 325-336.
- Hiebert, J., Berk, D., Miller, E., Gallivan, H., & Meikle, E. (2019). Relationships between opportunity to learn mathematics in teacher preparation and graduates' knowledge for teaching mathematics. *Journal for Research in Mathematics Education*, 50(1), 23-50.
- Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1-27). Lawrence Erlbaum Associates.
- Hill, H. C. (2010). The nature and predictors of elementary teachers' mathematical knowledge for teaching. *Journal for Research in Mathematics Education*, 41(5), 513-545.
- Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, 39(4), 372-400.
- Jacob, R., Hill, H., & Corey, D. (2017). The impact of a professional development program on teachers' mathematical knowledge for teaching, instruction, and student achievement. *Journal of Research on Educational Effectiveness*, 10(2), 379-407.
- Jacobs, V. R., & Empson, S. B. (2016). Responding to children's mathematical thinking in the moment: An emerging framework of teaching moves. *ZDM Mathematics Education*, 48(1-2), 185-197.
- Johnson, E., & Larsen, S. P. (2012). Teacher listening: The role of knowledge of content and students. *The Journal of Mathematical Behavior*, 31(1), 117-129.
- Lannin, J. K., Barker, D. D., & Townsend, B. E. (2007). How students view the general nature of their errors. *Educational Studies in Mathematics*, 66(1), 43-59.

- Lee, M. Y. (2021). Using a technology tool to help pre-service teachers notice students' reasoning and errors on a mathematics problem. *ZDM*, 53(1), 135-149.
- Lee, Y., Capraro, R. M., & Capraro, M. M. (2018). Mathematics teachers' subject matter knowledge and pedagogical content knowledge in problem posing. *International Electronic Journal of Mathematics Education*, 13(2), 75-90.
- Leikin, R., & Levav-Waynberg, A. (2007). Exploring mathematics teacher knowledge to explain the gap between theory-based recommendations and school practice in the use of connecting tasks. *Educational Studies in Mathematics*, 66(3), 349-371.
- Leikin, R., & Levav-Waynberg, A. (2008). Solution spaces of multiple-solution connecting tasks as a mirror of the development of mathematics teachers' knowledge. *Canadian Journal of Science, Mathematics and Technology Education*, 8(3), 233-251.
- Maxwell, J. A. (1996) *Qualitative research design: An interactive approach*. Sage Publications.
- Nagle, C., Moore-Russo, D., & Styers, J. (2017) Teachers' interpretations of student statements about slope. In E. Galindo, & J. Newton (Eds.), *Proceedings of the 39th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, (pp. 589-596).
- Nathan, M. J., & Petrosino, A. (2003). Expert blind spot among preservice teachers. *American Educational Research Journal*, 40(4), 905-928.
- National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. NCTM.
- National Council of Teachers of Mathematics *Principles to Actions Professional Learning Toolkit*. <https://www.nctm.org/PtAToolkit/>
- Ni Shuilleabhain, A. (2016). Developing mathematics teachers' pedagogical content knowledge in lesson study. *International Journal for Lesson and Learning Studies*, 5(3), 212-226.
- Norton, A., McCloskey, A., & Hudson, R. A. (2011). Prediction assessments: Using video-based predictions to assess prospective teachers' knowledge of students' mathematical thinking. *Journal of Mathematics Teacher Education*, 14(4), 305-325.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods*. (3rd ed.). SAGE.
- Peterson, R., & Treagust, D. (1995). Developing preservice teachers' pedagogical reasoning ability. *Research in Science Education*, 25(3), 291-305.
- Rittle-Johnson, B., Schneider, M., & Star, J. R. (2015). Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of mathematics. *Educational Psychology Review*, 27(4), 587-597.
- Schukajlow, S., & Krug, A. (2014). Do multiple solutions matter? Prompting multiple solutions, interest, competence, and autonomy. *Journal for Research in Mathematics Education*, 45(4), 497-533.
- Shin, D. (2020). Preservice Mathematics Teachers' Selective Attention and Professional Knowledge-Based Reasoning About Students' Statistical Thinking. *International Journal of Science and Mathematics Education*, <https://doi.org/10.1007/s10763-020-10101-w>
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4-14.
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1-22.
- Silver, E. A., Ghouseini, H., Gosen, D., Charalambous, C., & Strawhun, B. T. (2005). Moving from rhetoric to praxis: Issues faced by teachers in having students consider multiple solutions for problems in the mathematics classroom. *The Journal of Mathematical Behavior*, 24(3-4), 287-301.


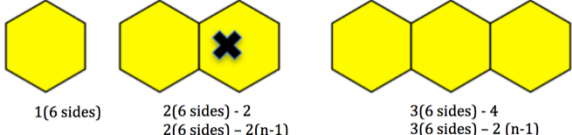


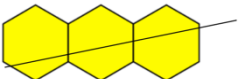
-
- Sitrava, R. T., (2020). Middle School Mathematics Teachers' Reasoning about Students' Nonstandard Strategies: Division of Fractions. *International Journal for Mathematics Teaching and Learning*, 21(1), 77-93.
- Spalding, E., Klecka, C. L., Lin, E., Wang, J., & Odell, S. J. (2011). Learning to teach: It's complicated but it's not magic. *Journal of Teacher Education*, 62(1), 3-7.
- Star, J. R., & Stylianides, G. J. (2013). Procedural and conceptual knowledge: Exploring the gap between knowledge type and knowledge quality. *Canadian Journal of Science, Mathematics and Technology Education*, 13(2), 169-181.
- Steele, M. D., & Rogers, K. C. (2012). Relationships between mathematical knowledge for teaching and teaching practice: The case of proof. *Journal of Mathematics Teacher Education*, 15(2), 159-180.
- Styers, J. L., Nagle, C. R., & Moore-Russo, D. (2020). Teachers' noticing of students' slope statements: attending and interpreting. *Canadian Journal of Science, Mathematics and Technology Education*, 20(3), 504-520.
- Taşdan, B. T., & Çelik, A. (2016). A Conceptual Framework for Examining Mathematics Teachers' Pedagogical Content Knowledge in the Context of Supporting Mathematical Thinking. *European Journal of Education Studies*, 2(5), 90-120.
- Thomas, J., Jong, C., Fisher, M. H., & Schack, E. O. (2017). Noticing and knowledge: Exploring theoretical connections between professional noticing and mathematical knowledge for teaching. *The Mathematics Educator*, 26(2), 3-25.
- Van Dooren, W., Verschaffel, L., & Onghena, P. (2002). The impact of preservice teachers' content knowledge on their evaluation of students' strategies for solving arithmetic and algebra word problems. *Journal for Research in Mathematics Education*, 33(5), 319-351.
- Wuttke, E., & Seifried, J. (Eds.). (2017). *Professional error competence of preservice teachers: Evaluation and support*. Springer.

6. APPENDIX †

Scoring Rubric for the Hexagon task

Task 2: Generate Multiple Solution Strategies		
Topic	Valid Strategies	Description of Coding
Finding the pattern in the Hexagon task	Tops and bottoms plus ends	<p>In this strategy, the student considers the tops of the hexagons as two times the train number and the bottoms. Since there are two sides on top of each hexagon, the number of top sides on any train is $n \times 2$ (n hexagons \times 2 top sides per hexagon) or $2n$. Similarly, the number of bottom sides is also $2n$. Then, the two end sides are considered separately.</p> <p>Possible representations: Verbal description. Equations: $2n + 2n + 2$ or $2(2n) + 2$.</p> 
	Tops and bottoms of each plus ends	<p>In this strategy, the student considers the tops of each hexagon and the bottoms of each hexagon. Then, the two end sides are considered separately.</p> <p>Possible representations: Verbal description. Equations: $4n + 2$, or $(2 + 2)n + 2$</p> 
	Insides and Outsides	<p>In this strategy, the student considers the end hexagons, noticing that each contributes five to the perimeter. Then, they consider that each internal hexagon contributes four.</p> <p>Possible representations: Verbal description. Equations: $5 + 4(n - 2) + 5$, or $4(n - 2) + 5 + 5$, or $4(n - 2) + 10$.</p>  <p>The second train is the two end hexagons that will be separated. They have 5 sides each (not including the one shared in the middle). Thus, the perimeter is $5 + 5$ or 10.</p>

† All the scoring rubrics are made based on the task solution paths on NCTM’s Professional Learning Toolkit

		 <p>The hexagon is the middle being added. This adds two on top and two on the bottom for 4 sides total. For every added hexagon, 4 more sides need to be added to the second train whose perimeter is 10.</p> <p>Note: the formula does work on Train 1, even though one can't really see the 5 and 5 in the train.</p>
	<p>Total minus shared sides</p>	<p>In this strategy, the student considers that each hexagon has six sides and notices that sides between hexagons are no longer on the perimeter.</p> <p>Possible representations: Verbal description. Equations: $6n - 2(n - 1)$</p> 
	<p>Symmetry split</p>	<p>In this strategy, the student considers the top sides and one end side as a unit and the bottom sides with the other end side.</p> <p>Possible representations: Verbal description. Equations: $(2n + 1) + (2n + 1)$ or $2(2n + 1)$</p>  <p>There are 3 sides above (two sides and one end) and it repeats below it.</p>  <p>There is two sides on the top of each, plus an end...and the same on the bottom</p>  <p>This time its 3 plus 2 plus 2 on the top...times two, because it's also on the bottom.</p> <p>$(2n + 1) + (2n + 1)$ if you think top and bottom $2(2n + 1)$ if you think doubling</p>

	<p>Increases by four</p>	<p>In this strategy, the student notices that the perimeter values increase by four with each additional hexagon.</p> <p>Possible representations: Verbal description. Equations: $4n + 2$. Table: list values., notice an increase of 4 each time (may conclude equation is $n + 4$, which is correct if n is the perimeter of the $n - 1^{th}$ train) Graph: plot points.</p> <div data-bbox="702 537 1165 896" data-label="Figure"> <table border="1"> <caption>Data points from the Hexagon Train Perimeters graph</caption> <thead> <tr> <th>Train Number</th> <th>Perimeter</th> </tr> </thead> <tbody> <tr><td>0</td><td>2</td></tr> <tr><td>1</td><td>6</td></tr> <tr><td>2</td><td>10</td></tr> <tr><td>3</td><td>14</td></tr> <tr><td>4</td><td>18</td></tr> <tr><td>5</td><td>22</td></tr> <tr><td>6</td><td>26</td></tr> <tr><td>7</td><td>30</td></tr> <tr><td>8</td><td>34</td></tr> <tr><td>9</td><td>38</td></tr> <tr><td>10</td><td>42</td></tr> </tbody> </table> </div> <p>Plotted the points for each of the trains after counting the perimeter of each, and realized the pattern was linear, increasing by four as the train number increases by one as the slope.</p> <p>$y = mx + b$</p> <p>$y = 4x + b$ (used a point and guess and check to solve for b)</p> <p>$y = 4x + 2$</p> <p>[or could connect all the points on graph with a straight edge and see that when $x = 0, y = 2$]</p>	Train Number	Perimeter	0	2	1	6	2	10	3	14	4	18	5	22	6	26	7	30	8	34	9	38	10	42
Train Number	Perimeter																									
0	2																									
1	6																									
2	10																									
3	14																									
4	18																									
5	22																									
6	26																									
7	30																									
8	34																									
9	38																									
10	42																									

Task 3b: Explain Multiple Solution Strategies		
Topic	Component	Description of Coding
Finding the pattern in the Hexagon task	Explains <i>procedurally</i> what students might have done	<ul style="list-style-type: none"> Explains how students get each solution pattern. <p>For Rachel's response, indicates that $2n + 2n + 2$ or $2(2n) + 2$ can be found by adding tops and bottoms plus ends.</p> <p>For Sam's response, indicates that $5 + 4(n - 2) + 5$, or $4(n - 2) + 5 + 5$, or $4(n - 2) + 10$ can be found by adding the insides and outsides.</p> <p>For Jason's response, indicates that $6n - 2(n - 1)$ can be found by adding all sides minus shared sides.</p>
	Explains <i>conceptually</i> what students might have done	<ul style="list-style-type: none"> Indicates that there is a pattern between the number of train and the perimeter of the hexagons. Indicates that the pattern consists of both multiplying (by 4) and adding (by 2).

Scoring Rubric for the Counting Cubes task

Task 2: Generate Multiple Solution Strategies									
Topic	Valid Strategies	Description of Coding							
Finding the pattern in the Counting Cubes task	Arms plus middle cube	In this strategy, the student considers that the figure has 5 arms and the number of cubes in each arm is 1 less than the number of building number. $1 + 5(n - 1) = 1 + 5n - 5 = 5n - 4$							
	Arms minus 4 cubes	In this strategy, the student considers that the figure has 5 arms and the number of cubes in each arm is the same as the building number. Then, students subtract 4 cubes since in the pattern, there is 1 cube less in each arm. $5n - 4$							
	Arms plus tower	In this strategy, the student considers that there is one tower in the middle and 4 arms in the figure. The tower has n cubes (n : building number) and each arm has $n - 1$ cubes. $n + 4(n - 1) = n + 4n - 4 = 5n - 4$							
	Table method	Table method: <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Building number</th> <th>Number of Cubes</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>1</td> </tr> <tr> <td>2</td> <td>$1 + 5 = 6$</td> </tr> <tr> <td>3</td> <td>$1 + 5 + 5 = 11$</td> </tr> </tbody> </table> <p>By using the table, students find the number of cubes as the following (n is the building number)</p> $1 + 5(n - 1) = 5n - 4$	Building number	Number of Cubes	1	1	2	$1 + 5 = 6$	3
Building number	Number of Cubes								
1	1								
2	$1 + 5 = 6$								
3	$1 + 5 + 5 = 11$								

Task 3b: Explain Multiple Solution Strategies		
Topic	Component	Description of Coding
Finding the pattern in the Hexagon task	Explains <i>procedurally</i> what students might have done	<ul style="list-style-type: none"> Explains how students get each solution pattern. <p>For David's response, indicates that $n + 4(n - 1)$ can be found as one tower in the middle and 4 arms in the figure. The tower has n cubes (n: building number) and each arm has $n - 1$ cubes.</p> <p>For Emily's response, indicates that $1 + 5(n - 1)$ can be found by considering n as the building number and 1 as the middle cube.</p> $1 + 5(n - 1) = 5n - 4$ <p>For Mary's response, indicates that $5n + 1$ can be found by adding all sides (n as the number of arms) plus one cube in the middle.</p>
	Explains <i>conceptually</i> what students might have done	<ul style="list-style-type: none"> Indicates that there is a pattern between the building number or cubes in the tower in the middle or arms and the total number of cubes. Indicates that the pattern consists of both multiplying (by 5) and adding (by -5).

Scoring Rubric for the Two Storage Tanks task

Task 1: Identify Concepts Underlying Procedures			
Topic	Concept	Score	Description of Coding
Two Storage Tanks Task	Linear Equations in two variables	0	<ul style="list-style-type: none"> • Makes no mention of the linear equations.
		1	<ul style="list-style-type: none"> • Makes a general statement that students should consider that the graphs are linear and there are two variables (x: number of hours, y: gallons of water)
		2	<ul style="list-style-type: none"> • Sees that there is a linear relationship between variables x and y variables. Provides further explanations on what x and y refers to and gives specific examples by using the graph provided in the task.
	Initial Value of the function	0	<ul style="list-style-type: none"> • Makes no mention of the initial value of the function.
		1	<ul style="list-style-type: none"> • Makes a general statement that students should consider the initial value of the function for finding the intersection point of two equations.
		2	<ul style="list-style-type: none"> • Sees how the initial value might impact the intersection point and provides further explanations by giving specific numbers from the given graph.
	The rate of change	0	<ul style="list-style-type: none"> • Makes no mention of the rate of change.
		1	<ul style="list-style-type: none"> • Makes a general statement that the rate of change impact the intersection point of two linear equations.
		2	<ul style="list-style-type: none"> • Sees that the rate of change is the slope of the function and how different rate of changes might impact on the intersection point. • Provides examples on how small or big rate of change might impact the steepness of the graph and what it means to be steeper by comparing different rates of change.

Assessing Measurement Invariance of Achievement Emotions Questionnaire for Teachers in Prospective Teacher Sample

Sevilay Kilmen ^{1,*}

¹Bolu Abant İzzet Baysal University, Faculty of Education, Department of Educational Sciences, Bolu, Turkey.

ARTICLE HISTORY

Received: June 02, 2020

Revised: July 28, 2021

Accepted: Sep. 13, 2021

Keywords:

Measurement invariance,
Emotions about teaching,
Convergent validity,
Achievement emotions,
Self-efficacy.

Abstract: The purpose of the current study is to determine whether the Achievement Emotions Questionnaire for Teachers (AEQT) is a psychometrically sound instrument to measure prospective teachers' teaching-related emotions. The three-factor model of the AEQT was confirmed in a prospective teacher sample. Also, reliability results showed that the AEQT is a reliable measurement tool. Measurement invariance results revealed that configural, metric, and scalar invariance were provided across gender. These findings support the use of the AEQT when examining differences based on achievement emotions across gender. For teacher training programs, only configural invariance was provided. Although configural invariance suggests that the three-factor structure of the AEQT is the same across the teacher training programs, the lack of metric invariance indicates that the relationship between the items and the underlying latent variable the AEQT factors is not the same across these groups. The observed variables are not related to the latent variable equivalently across teacher training programs. This result does not allow the comparison of path coefficients and covariances between observed and latent variables across teacher training programs. Also, the lack of scalar invariance indicates that different teacher training programs may interpret some items differently and prevent a comparison of averages between these groups.

1. INTRODUCTION

One of the research topics in educational settings is emotions about teaching. Examining the factors influencing emotions about teaching is of considerable importance, given that the investigation of emotions about teaching enables researchers and teacher trainers to better understand and predict prospective teachers' behavior. Indeed, to date, a large body of studies examined the factors related to emotions about teaching in different populations and contexts. The previous research results emphasized that emotions about teaching are related to many important teaching-related factors in educational settings (Henao-Arias et al., 2017), such as burnout (Frenzel et al., 2016), job satisfaction (Moè et al., 2010), teacher-student relationship, classroom discipline, students' engagement (Hagenauer et al., 2015), and self-efficacy (Eren, 2014).

The Achievement Emotions Questionnaire for Teachers (AEQT, Frenzel et al., 2010) is one of the most commonly used instruments for measuring different facets of teachers' achievement

*CONTACT: Sevilay Kilmen ✉ kaplansevilay@yahoo.com 📍 Bolu Abant İzzet Baysal University, Faculty of Education, Department of Educational Sciences, Bolu, Turkey

emotions. The AEQT was used to measure teachers'/prospective teachers' achievement emotions on many different culture samples (Becker et al., 2015; Frenzel et al., 2009; Hong et al., 2016; Klassen et al., 2012). The research results revealed that the three-factor measurement model of the AEQT was confirmed on different culture samples. However, despite the AEQT's widespread use in various countries, studies on testing the measurement invariance of the AEQT across gender and different teacher training programs are missing.

The first aim of the current research is to examine the factor structure of the AEQT by using both parallel analysis and confirmatory factor analysis on a prospective teacher sample. The second aim of the current research is to provide convergent validity evidence by investigating the relationships of prospective teachers' achievement emotions with their professional self-efficacy beliefs on the Turkish prospective teacher sample. The third aim is to examine the measurement invariance of the AEQT across gender and different teacher training programs.

The current study is crucial for three reasons: First, to the author's knowledge, this is the first study to examine if the measurement invariance is established across gender and teacher training programs in emotions about teaching. In previous studies, although participants were compared according to their demographic features, the measurement invariance of the AEQT was not addressed in these studies. If the measurement invariance was not established, it means that comparison groups do not perceive and interpret items in the same way. Therefore, conducting these comparisons may not be proper to see real differences between groups. Examining the measurement invariance of the AEQT provides an evaluation of whether the AEQT measures the same latent construct(s) in different groups (Raykov et al., 2012). Therefore, the results of the current study may be especially useful for studies which compare teachers'/prospective teachers' teaching emotions according to gender and teacher training programs.

Second is that the previous studies investigating emotions about teaching by using the AEQT were often conducted on in-service teacher samples such as German teachers (Becker et al., 2015, Frenzel et al., 2009), Japanese and Korean teachers (Hong et al., 2016), Greek teachers (Karagianni & Papaefthymiou-Lytraand, 2018), and Canadian teachers (Klassen et al., 2012). Studies examining emotions about teaching by pre-service teacher samples were much rarer (e.g., Eren, 2014). The current study provides concrete contributions to the studies which aim to conduct group comparisons by using the AEQT, on a prospective teacher sample, by focusing on evaluating the psychometric quality of the AEQT on a prospective teacher sample.

Third, as mentioned before, previous studies using the AEQT focused on the teacher samples, not prospective teachers. Therefore, convergent validity pieces of evidence were obtained from the teacher sample. Besides measurement invariance, the current study provides supportive evidence for the convergent validity of the AEQT on prospective teacher samples by examining the relationship between prospective teachers' achievement emotion and self-efficacy beliefs.

1.1. Achievement Emotions About Teaching

Achievement emotions were examined in educational settings by dividing them into categories in terms of their features. According to a number of study results, achievement emotions are divided into two primary dimensions as valence and activation. In terms of valence, emotions are divided into two categories: positive versus negative. On the other hand, achievement emotions are classified as activating versus deactivating in terms of activation. For instance, while enjoyment, hope, and pride are positive activating emotions, relief is a positive deactivating emotion. Anger, anxiety, and shame are negative activating emotions while hopelessness is a negative deactivating emotion (Pekrun et al., 2004). In this study, the three-factor structure of AEQT was examined on a prospective teacher sample. According to mentioned explanations about achievement emotions, the AEQT framework focuses on one

positive and two negative activating emotions (i.e., enjoyment, anger, and anxiety) which are prominent achievement emotions for teachers (Frenzel et al., 2009; Sutton, 2004; Sutton & Wheatley, 2003).

Numerous studies have shown that emotions about teaching are the key concepts to closely relate to classroom climate and teaching quality. For example, relevant literature reveals that negative achievement emotions about teaching (i.e., anger and anxiety) are negatively related to key concepts about teaching such as teachers' self-efficacy and enthusiasm (Frenzel et al., 2016, Frenzel et al., 2009; Kunter et al., 2008). Also, teachers' emotions about teaching are related to their perceptions about student characteristics. For example, whereas teachers' perceptions of students' performance, motivation, and discipline during the lessons were positively related to their positive emotions about teaching (i.e., enjoyment), they were negatively related to their negative emotions about teaching (i.e., anger and anxiety, Frenzel et al., 2009). Moreover, teachers' emotions are also closely related to their students' emotions. Indeed, a recent study's results based on longitudinal data demonstrates evidence of the reciprocal transmission of teacher and student emotions (Frenzel et al., 2018).

1.2. Convergent Validity of the AEQT

In this study, to test the convergent validity of the AEQT in the Turkish prospective teacher sample, the relationships between the prospective teachers' emotions about teaching and their professional self-efficacy beliefs were examined. Teacher efficacy belief refers to the "judgment of teachers' capabilities to bring about desired outcomes of students' engagement and learning, even among those students who may be difficult or unmotivated" (Tschannen-Moran & Woolfolk Hoy, 2001, 783). In a number of studies, the relationships between teachers' self-efficacy beliefs and their emotions about teaching were examined (e.g., Hascher & Hagenauer, 2016). Previous study results showed that emotions and self-efficacy are related variables (e.g., Burić et al., 2020). In a study examining the relationships between self-efficacy and emotions about teaching based on a sample who are in the teaching practicum, it was found that prospective teachers' teaching enjoyment in teaching practicum is positively predicted by their self-efficacy, whereas anxiety is negatively predicted (Hascher & Hagenauer, 2016).

On the other hand, teacher self-efficacy was addressed as one latent variable, in some previous studies (e.g., Hascher & Hagenauer, 2016), while in some, this variable was examined in its dimensions (Hagenauer et al., 2015). In the current study, to assess prospective teachers' self-efficacy beliefs, the three-factor teacher self-efficacy beliefs framework described by Tschannen-Moran and Woolfolk Hoy (2001) was adopted. This framework includes three dimensions: self-efficacy for instructional strategies, self-efficacy for classroom management, and self-efficacy for student engagement. Relevant literature shows that specific types of self-efficacy may affect specific teaching emotions (e.g., Hagenauer et al., 2015). Specifically, study results showed that teachers who held high self-efficacy beliefs had more positive emotions (e.g., enjoyment, pride) and less anger and anxiety, compared to teachers who had low self-efficacy beliefs (e.g., Hong et al., 2016). Therefore, by considering that prospective teachers' different types of self-efficacy beliefs may have a different influence on their emotions about teaching, in the current study, the roles of three types of self-efficacy beliefs on emotions about teaching were addressed separately (i.e., instructional strategies, classroom management, and student engagement).

Based on the previous study findings, it is reasonable to hypothesize that the prospective teachers' positive achievement emotion about teaching would positively associate with their self-efficacy beliefs and negative achievement emotion about teaching would negatively associate with their self-efficacy beliefs. Therefore, while testing convergent validity of the AEQT by using prospective teachers' professional self-efficacy, it was expected that the

enjoyment subscale of the AEQT would be related to prospective teachers' self-efficacy beliefs positively and anger and anxiety would be related negatively.

1.3. Measurement Invariance

Measurement invariance is the level of perception and interpretation of scale in the same way across groups (Byrne & Watkins, 2003). When comparing scores obtained from a scale, ensuring measurement invariance between groups is a prerequisite (Marsh et al., 2014). If the scale items are perceived and interpreted differently by the groups, the scores obtained from the comparison of these groups may be misinterpreted. There are four hierarchical types of measurement invariance levels: configural, metric, scalar, and strict invariance (Vandenberg & Lance, 2000).

- Configural invariance: tests if the factor structure of a scale is the same across comparison groups.
- Metric invariance: examines if factor loadings of a scale besides factor structure are equal across comparison groups.
- Scalar invariance: examines if intercepts of a scale besides factor structure and factor loadings are equal across comparison groups.
- Strict invariance: examines if residual variances of a scale besides factor structure, factor loadings, and intercepts are equal across comparison groups.

In the present study, in the configural invariance stage, the AEQT was tested for whether the same factor structure across gender and the teacher training program groups exist. In the metric invariance model stage, the factor loadings of the AEQT items were constrained to be equal across gender and teacher training program groups. In the scalar invariance stage, the AEQT item intercepts were constrained to be equal across the groups in addition to factor loadings. In the last level of measurement invariance procedure, to test strict invariance, the error variances were constrained across groups in addition to the factor loadings and intercepts.

1.4. The Present Study

The purpose of the current study is threefold. The first purpose is to test the factor structure of the AEQT with the Turkish prospective teacher sample. The second purpose is to examine whether the three-factor measurement model of the AEQT had measurement invariance across gender and teacher training programs in the Turkish prospective teacher sample. The third purpose is to provide evidence of the convergent validity of the AEQT in the Turkish prospective teacher sample. The research questions of this study are:

- 1) Is the factor structure of the AEQT similar to the original scale?
- 2) Are the configural, metric, scalar, and strict measurement invariance of the AEQT provided across gender?
- 3) Are the configural, metric, scalar, and strict measurement invariance of the AEQT provided across teacher training program groups?
- 4) Is there any relationship between prospective teachers' emotions about teaching and their professional self-efficacy, as an indicator of the AEQT?

2. METHOD

2.1. Study Group

To determine the current study's research sample, convenience sampling was used. The general research sample consists of 560 prospective teachers (407 females) majoring in science teaching (n = 107), social sciences teaching (n = 108), English language teaching (n = 138), special education teaching (n = 106), and mathematics teaching (n = 101) in the Faculty of Education of a university located in the north-west of the Black Sea region in Turkey. These were the

participants in the current study. The research sample consists of 133 freshmen, 170 sophomores, 137 juniors, and 120 seniors. Their ages range from 17 to 37 ($M = 20.54$, $S = 2.31$).

The research sample was randomly divided into two samples to conduct exploratory and confirmatory factor analyses and examine the convergent validity of the AEQT in two different samples separately. Sample 1 consists of 271 prospective teachers (194 females) who participated in the present research majoring in science teaching ($n = 77$), social sciences teaching ($n = 25$), English language teaching ($n = 114$), and special education teaching ($n = 55$). There are 73 freshmen, 13 sophomores, 92 juniors, and 93 seniors in Sample 1. Their ages range from 17 to 29 ($M = 20.61$, $S = 1.59$).

Sample 2 consists of 289 prospective teachers (213 females) majoring in science teaching ($n = 30$), social sciences teaching ($n = 83$), English language teaching ($n = 24$), special education teaching ($n = 51$), and mathematics teaching ($n = 101$). The research sample consisted of 73 freshmen, 13 sophomores, 92 juniors, and 93 seniors. Their ages range from 17 to 29 ($M = 20.48$, $S = 2.81$). Measurement invariance analyses were conducted by merging Sample 1 and Sample 2.

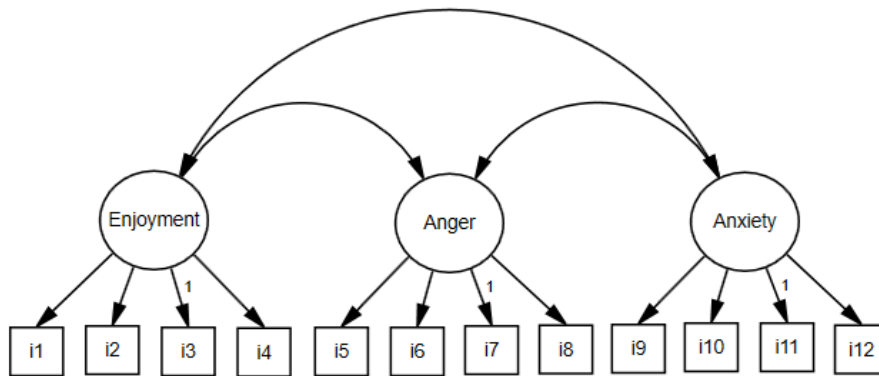
2.2. Research Instruments

The AEQT (Frenzel et al., 2009) and the Ohio State Teacher Efficacy Scale (OSTES, Tschannen-Moran & Woolfolk-Hoy, 2001) were used as measurement tools in the current study. These scales were applied to undergraduate students in the 2018-2019 fall semester. How to answer items in these scales was explained briefly to the participants before the administration, and any questions from the prospective teachers were responded to by the researcher. All participants were informed that their data would not be shared with anyone. All participation was voluntary. The participants completed the scales approximately in 10 min. The general features of the AEQT (Frenzel et al., 2009) and the OSTES (Tschannen-Moran & Woolfolk-Hoy, 2001) were then introduced.

2.2.1. *The achievement emotions questionnaire for teachers (AEQT)*

The AEQT (Frenzel et al., 2010) is a self-report scale with 12 items used to measure teachers' achievement emotions about teaching. The original AEQT was developed to measure in-service teachers' achievement emotions about teaching. The AEQT was adapted to Turkish by Eren (2014). As the AEQT was administered to prospective teachers in Eren's (2014) research, all the AEQT items were converted to the future tense form except for one item (i.e., I feel uneasy when I think about teaching; Eren, 2014). The AEQT consists of three first-order factors (see Figure 1): enjoyment (four items, e.g., I will teach with enthusiasm), anger (four items, e.g., I will get really mad while I teach), and anxiety (four items, e.g., Preparing to teach will cause me to worry). Possible responses range from 1 (strongly disagree) to 5 (strongly agree). The findings showed that all dimensions of the AEQT resulted in satisfactory reliability coefficients (see Table 1).

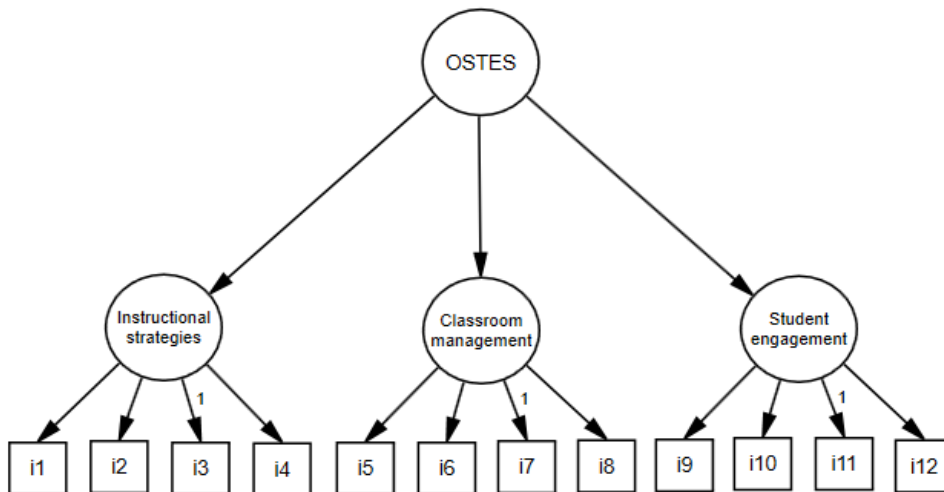
Figure 1. The AEQT measurement model.



2.2.2. The Ohio state teacher efficacy scale (OSTES)

The OSTES (Tschannen-Moran & Woolfolk-Hoy, 2011) is a self-report scale with 12 items test used to measure teacher self-efficacy. The OSTES consists of one second-order factor and three first-order factors (see, Figure 2): self-efficacy for instructional strategies (four items, e.g., To what extent can you craft good questions for your students?), self-efficacy for classroom management (four items, e.g., How well can you establish a classroom management system with each group of students?), and self-efficacy for student engagement (four items, e.g., How much can you do to get students to believe they can do well in schoolwork?). Prospective teachers responded to items using a 9-point Likert-type scale ranging from 1 (nothing) to 9 (a great deal).

Figure 2. The OSTES measurement model.



In the current study, CFA was conducted to see whether the three-factor structure of the OSTES fit the data. CFA results show that the second-order self-efficacy model had a good fit to the current data ($\chi^2_{(51)} = 218.98$; comparative fit index (CFI) = .95; Tucker-Lewis index (TLI) = .94; standardized root-mean-square residual (SRMR) = .06). To evaluate the reliability of the subscales of the OSTES, Cronbach's alpha coefficients were calculated. For all subscales and the whole scale, satisfactory reliability coefficients ranging from .82 to .88 were obtained.

2.3. Data Analysis

Little’s missing completely at random (MCAR) test (Little, 1988) was used to examine if missing values are completely at random or not. After non-significant Little’s MCAR test

results, the Expectation-Maximization algorithm, which is a technique that uses maximum likelihood estimates for incomplete data, was performed. Original AEQT has a three-factor structure (Frenzel et al., 2010). In order to explore the factor structure of the AEQT on the Turkish prospective teachers, the parallel analysis (Horn, 1965), which is commonly used for scale dimensionality (Timmerman & Lorenzo-Seva, 2011), was conducted on Sample 1. Following the parallel analysis, to examine if the three-factor measurement model of the AEQT was confirmed by the research data, the confirmatory factor analysis was carried out using a different sample (Sample 2) (Kline, 2005). Model fit was evaluated using chi-square (χ^2), comparative fit index ($CFI \geq .90$), and Tucker-Lewis index ($TLI \geq .90$), and standardized root-mean-square residual ($SRMR \leq .08$) (Brown & Cudeck, 1993; Kline, 2005; Hu & Bentler, 1999). To evaluate scales' reliability, Cronbach's alpha was computed.

As the achievement emotions model for teachers obtained from the AEQT was considered to be able to be interpreted differently by the sub-groups in the prospective teacher sample, configural, metric, scalar, and strict measurement invariance across gender and teacher training programs were examined in the present study by using multi-group confirmatory factor analysis. Measurement invariance model comparisons were assessed using ΔCFI cutoff criteria ($\Delta CFI \leq .01$; Chen, 2007; Cheung & Rensvold, 2009).

3. RESULTS

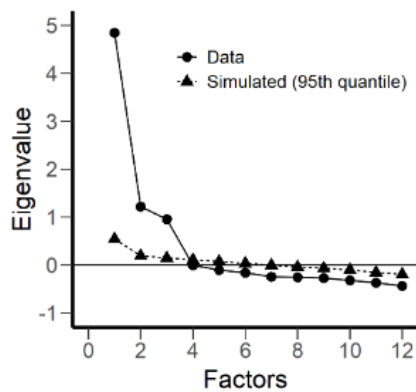
3.1. Factor Structure of the AEQT

3.1.1. Parallel analysis results

The parallel analysis was applied to the AEQT scores obtained by "Sample 1" ($n = 271$) to reveal the factor structure of the AEQT in the research sample. The parallel analysis results suggested the three-factor model proposed by Frenzel, Pekrun, and Goetz (2010) ($\chi^2_{(33)} = 129.123$, $p < .001$, see Figure 3 and Table 1). It was found that item loadings on these three dimensions were above .69.

Table 1. Parallel analysis, confirmatory factor analysis, and reliability results of the AEQT.

	Parallel analysis results (Sample 1)			Confirmatory factor analysis results (Sample 2) Standardized parameter estimations	Reliability results (Cronbach's alpha)		
	Factor loadings				Sample 1	Sample 2	Entire sample
Enjoyment1	.89	-.03	.05	.65			
Enjoyment2	.92	-.06	-.07	.76			
Enjoyment3	.86	.00	.00	.69	.91	.91	.92
Enjoyment 4	.73	-.12	.01	.58			
Anxiety1	-.10	.69	.06	.78			
Anxiety2	.06	.91	-.07	.91	.83	.83	.83
Anxiety3	.04	.84	.08	.95			
Anxiety4	-.22	.69	.04	.88			
Anger1	.04	.11	.70				.44
Anger2	.02	.04	.92				.49
Anger3	.06	-.03	.94		.85	.85	.44
Anger4	-.28	-.03	.71				.54
	$\chi^2_{(33)} = 129.123$, $p < .001$			$\chi^2_{51} = 259.060$, $p < .001$ CFI=.95, TLI=.94, SRMR=.06			

Figure 3. Parallel analysis eigenvalues.

3.1.2. Confirmatory factor analysis results

For the confirmatory factor analysis, the three-factor measurement model identified by Frenzel, Pekrun, and Goetz (2010) was used as the baseline for confirmatory factor analysis which was performed on the “Sample 2” ($n=289$). Each item was specified to reflect the corresponding factor and the three first-order factors were allowed to correlate. When the results were evaluated, it was found that the three-factor measurement model provided a good fit to the data ($\chi^2_{51}=259.060$, CFI = .95, TLI = .94, SRMR = .06, see Table 1), suggesting that the three first-order achievement emotions measurement model offered a reasonably good representation of the data. All standardized factor loadings were above .44. In addition, the findings indicated a moderately latent correlation between anger and anxiety sub-dimensions ($r = .495$, $p < .001$). These findings suggested that two negative factors of the AEQT can be the first-order factors of a second-order factor. Therefore, a higher-order model was formed. Model 2 comprised one second-order latent factor overarching anger and anxiety, and enjoyment as separate first-order latent factors. But the findings showed that the higher-order model did not yield acceptable fit indexes ($\chi^2_{32}=404.324$, TLI = .78, CFI = .83, SRMR = .17). Therefore, the higher-order model was rejected. Consequently, the comparison of Model 1 and Model 2 suggested the adoption of Model 1 with three first-order factors (i.e., enjoyment, anger, and anxiety) measured by 12 indicators.

3.2. Measurement Invariance Results

3.2.1. Measurement invariance across gender

In the current study, multiple-group confirmatory factor analysis was used to assess configural, metric, scalar, and strict invariance across gender and then across teacher training programs for the three-factor achievement emotion measurement model. The results showed that the configural invariance model across gender presented an acceptable fit to the data ($\chi^2_{102}=361.402$, CFI = .934, see Table 2). Following the configural invariance model, to examine metric invariance, factor loadings of the AEQT items were constrained to be equal across gender. As seen in Table 2, compared with the configural invariance model as a baseline model, the metric invariance model did not demonstrate any change in CFI ($\chi^2_{111}=392.236$, CFI = .923, $\Delta\text{CFI} = .000$, Chen, 2007; Cheung & Rensvold, 2009). This finding showed that metric invariance was supported in research data. Following the metric invariance model, the scalar invariance model was tested by constraining both item factor loadings and item intercepts to be equal across gender. The results showed that the χ^2 change between metric and scalar invariance model was not statistically significant, and CFI value did not decrease in the scalar invariance model ($\chi^2_{120}=399.709$, CFI = .928, $\Delta\text{CFI} = .000$). When configural invariance, metric invariance, and scalar invariance models were evaluated together across gender, it was observed that ΔCFI demonstrated no significant reduction in model fit. That is, the AEQT factor

structure, factor loadings, and intercepts did not differ significantly across gender. These findings showed that males and females responded to items of the AEQT in the same way. However, strict measurement invariance implying invariance of residual variances yielded a poor fit to data across gender ($\chi^2_{132} = 577.709$, CFI = .896, $\Delta\text{CFI} = .039$). That is, the error variances are different across gender.

Table 2. *Measurement Invariance Results of the AEQT.*

	χ^2	<i>df</i>	CFI	ΔCFI
Gender				
Configural	361.402	102	.934	-
Metric	392.236	111	.934	-
Scalar	399.709	120	.935	-
Strict	577.709	132	.896	.039
Teacher training programs				
Configural	616.850	255	.920	-
Metric	720.810	291	.904	.036
Scalar	762.790	327	.903	.001
Strict	968.880	375	.856	.047

3.2.2. Measurement invariance across teacher training programs

The same measurement invariance routine was also applied to test measurement invariance for teacher training programs. The AEQT's measurement invariance across science teaching, social sciences teaching, English language teaching, special education teaching, and mathematics teaching programs was investigated. According to the results, the unconstrained configural invariance model data fit was obtained ($\chi^2_{255} = 616.850$, CFI = .920). This finding suggested that the three-factor measurement model is similar across different teacher training programs. Following configural invariance, the metric invariance model was tested by constraining item factor loadings to be equal across teacher training programs. Results showed that the metric invariance model resulted in a significant loss of fit ($\chi^2_{255} = 720.810$, CFI = .904, $\Delta\text{CFI} = .036$). The loss of fit results suggested that item factor loadings are different across the teacher training programs. The scalar invariance model implying invariance of intercepts also yielded a poor fit to data, showing that item intercepts are different across teacher training programs ($\chi^2_{327} = 790.810$, CFI = .903, $\Delta\text{CFI} = .001$). Lastly, it was found that strict invariance is not supported. That is, the error variances are different across the teacher training programs ($\chi^2_{375} = 968.880$, CFI = .856, $\Delta\text{CFI} = .047$).

3.3. Convergent Validity Results

The correlation analysis results showed that the enjoyment component of the AEQT was negatively correlated with anger and anxiety (see Table 3) for both Sample 1 and Sample 2. The convergent validity results showed that, as expected, enjoyment is significantly and positively correlated with instructional strategies, classroom management, and student engagement subscales of the OSTES. Anger and anxiety subscales are significantly and negatively correlated with instructional strategies, classroom management, and student engagement subscales of the OSTES. Of note, the enjoyment subscale of the AEQT, compared with the anger subscale, and anxiety subscales showed a stronger correlation with the OSTES and its subscales.

Table 3. *Convergent validity results of the AEQT.*

	Enjoyment	Anxiety	Anger
Sample 1 (n=271)			
1.OSTES	.40	-.31	-.21
2.Instructional strategies	.33	-.30	-.19
3.Classroom management	.30	-.21	-.07
4.Student engagement	.40	-.29	-.27
Sample 2 (n=289)			
1.OSTES	.49	-.28	-.28
2.Instructional strategies	.42	-.24	-.27
3.Classroom management	.35	-.27	-.11
4.Student engagement	.48	-.22	-.34

4. CONCLUSION

There has been recent interest in examining the impact of teaching-related emotions in teacher training environments. As a result of this interest, researchers need psychometrically-sound items to assess teaching-related achievement emotions. The purpose of the current study was to determine whether the AEQT is a psychometrically sound instrument to measure prospective teachers' teaching-related emotions. From a theoretical point of view, the emerging factor structure implies the existence of highly correlated but also distinct emotions of the AEQT. Indeed, the three-factor model suggested a better fit for the data than did the higher-order model. This finding provides evidence that three teaching-related emotions were distinct. When the reliability results were examined, it was seen that the AEQT is a reliable measurement tool.

The results revealed that configural, metric, and scalar invariance were established across gender. These findings support the use of the AEQT when examining differences based on achievement emotions across gender (Brown, 2006). However, the current analyses suggested that while the AEQT demonstrated configural invariance (equal factor structure) across five teacher training programs, metric invariance (equal factor loadings) was not supported.

Although configural invariance suggests that the three-factor structure of the AEQT is the same across the teacher training programs, the lack of metric invariance indicates that the relationship between the items and the underlying latent variable the AEQT factors is not the same across these groups. That is, the observed variables are not related to the latent variable equivalently across teacher training programs. This result does not allow the comparison of path coefficients and covariances between observed and latent variables across teacher training programs (Chen et al., 2005). Also, the lack of scalar invariance indicates that different teacher training programs may interpret some items differently and prevent a comparison of averages between these groups (Van de Schoot et al, 2012; Vandenberg & Lance, 2000).

Strict invariance was established neither across gender nor across teacher training programs. Establishing configural, metric, and scalar invariance across gender could let the researchers compare the latent variables based on gender. But as strict invariance was not established for gender and the teacher training programs, the latent variables are measured with different amounts of error between groups (Van de Schoot et al., 2012). This result could cause a difference in factor score averages across gender and the teacher training programs even when true values of the underlying construct are the same (Brown, 2006). Therefore, it is important to be careful when the AEQT factor scores are compared across these teacher training programs in future investigations. The convergent validity was supported by results that revealed that

self-efficacy and achievement emotions are significantly and selectively related to each other. Enjoyment as a positive emotion is positively associated with self-efficacy, while anger and anxiety as negative emotions are negatively associated with self-efficacy. These results supported many research results that revealed the relationships between teachers'/prospective teachers' emotions and self-efficacy (Brigidoa et al., 2013; Hascher & Hagenauer, 2016; Moè et al., 2010; Stephanou et al., 2013).

4.1. Limitations

As with most educational studies, the current study has some limitations. First, the prospective teacher sample is unbalanced in terms of gender (most of them are female). It could work with a more balanced sample in terms of gender in future studies. Second, the current study focused only on three main emotions about teaching. In future studies, it could be interesting to examine how teachers'/prospective teachers' other emotions about teaching are influenced by gender and self-efficacy beliefs.

Third, in the present study, all analysis which was conducted was limited to 560 pre-service teachers majoring in science teaching, social sciences teaching, English language teaching, special education teaching, and mathematics teaching. The results of the current study showed that metric, scalar, and strict invariance of the AEQT was not provided across teacher training programs. A possible reason of these results may be sample size. Because, in the current study, to examine measurement invariance across teacher training programs, study sample was divided into five categories. Therefore, the number of participants in each category was highly decreased. Model-data fit measures for metric, scalar, and strict invariance of the AEQT may have been decreased depending on this reason. To increase generalizability of research findings and to reassess the measurement invariance, the AEQT can be used in larger and different samples. That is, future studies are needed to cross-validate the results with other samples.

Finally, given that those prospective teachers in the current study have high enjoyment and low anger and anxiety, the current study did not examine the specific relationships between self-efficacy and emotions about teaching in a sample that has low enjoyment and high anger and anxiety. However, it could be helpful to gain a deeper understanding of the relationships between professional self-efficacy and emotions about teaching in future studies using a homogeneous sample.

Acknowledgments

The author is grateful to Prof. Dr. Altay Eren for his help during the data collection process.

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

Sevilay Kilmen  <https://orcid.org/0000-0002-5432-7338>

5. REFERENCES

- Becker, E. S., Keller, M. M., Goetz, T., Frenzel, A. C., & Taxer, J. (2015). Antecedents of teachers' emotions in the classroom: An intraindividual approach. *Frontiers in Psychology, 6*, 1-12.
- Brígidoa, M., Borrachero, A. B., Bermejo, M. L., & Mellado, V. (2013) Prospective primary teachers' self-efficacy and emotions in science teaching. *European Journal of Teacher Education, 36*(2), 200-217.

- Brown, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In: K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, Sage.
- Brown, T. (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.
- Burić, I., Slišković, A., & Sorić, I. (2020). Teachers' emotions and self-efficacy: A test of reciprocal relations. *Frontiers in Psychology, 11*, 1650.
- Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34*(2), 155-175.
- Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modelling, 14*(3), 464-504.
- Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*(3), 471-492.
- Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.
- Eren, A. (2014). Relational analysis of prospective teachers' emotions about teaching, emotional styles, and professional plans about teaching. *Australian Educational Researcher, 41*(4), 381-409.
- Frenzel, A. C., Goetz, T., Stephens, E. J., & Jacob, B. (2009). Antecedents and effects of teachers' emotional experiences: An integrated perspective and empirical test. In P.A. Schutz & M. Zembylas (Eds.), *Advances in teacher emotion research: The impact on teachers' lives* (pp. 129-152). Springer.
- Frenzel, A. C., Pekrun, R., & Goetz, T. (2010). *Achievement emotions questionnaire for teachers (AEQ-T teacher)-user's manual*. Program of Psychology, University of Munich.
- Frenzel, A.C., Pekrun, R., Goetz, T., Daniels, L.M., Durksen, T.L., Becker-Kurz, B., & Klassen, R.M. (2016). Measuring teachers' enjoyment, anger, and anxiety: the teacher emotions scales (TES). *Contemporary Educational Psychology, 46*, 148-163.
- Frenzel, A. C., Becker-Kurz, B., Pekrun, R., Goetz, T., & Lüdtke, O. (2018). Emotion transmission in the classroom revisited: A reciprocal effects model of teacher and student enjoyment. *Journal of Educational Psychology, 110*(5), 628-639.
- Hagenauer, G., Hascher T., & Volet, E. (2015). Teacher emotions in the classroom: associations with students' engagement, classroom discipline, and the interpersonal teacher-student relationship. *European Journal of Psychology of Education, 30*, 385-403.
- Hascher, T., & Hagenauer, G. (2016). Openness to theory and its importance for preservice teachers' self-efficacy, emotions, and classroom behavior in the teaching practicum. *International Journal of Educational Research, 77*, 15-25.
- Henaó-Arias, J. F., Marin-Rodríguez, A. E., & Vanegas-García, J. H. (2017). Education hinging on emotions: An emotional view of education. *Educación y Educadores [online]*. 20(3), 451-465.
- Hong, J., Nie, Y., Heddy, B., Monobe, G., Ruan, J., You, S., & Kambara, H. (2016). Revising and validating achievement emotions questionnaire-teachers (AEQ-T-T). *International Journal of Educational Psychology, 5*(1), 80-107.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.
- Hu, L. & Bentler, P. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.
- Karagianni E., Papaefthymiou-Lytra S. (2018). EFL Teachers' Emotions: The driving force of sustainable professional development. In A. J. Martínez (Ed.), *Emotions in Second Language Teaching* (pp. 385-401). Cham: Springer.

- Klassen, R. M., Perry, N. E., & Frenzel, A. C. (2012). Teachers' relatedness with students: An underemphasized component of teachers' basic psychological needs. *Journal of Educational Psychology, 104*(1), 150-165.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling*. Guilford Publications, Inc.
- Kunter, M., Tsai, Y. M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction, 18*(5), 468-482.
- Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198-1202.
- Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85-110.
- Moè, A., Pazzaglia, F., & Ronconi, L. (2010). When being able is not enough. The combined value of positive affect and self-efficacy for job satisfaction in teaching. *Teaching and Teacher Education, 26*(5), 1145-1153.
- Pekrun, R., Goetz, T., Perry, R. P., Kramer, K., Hochstadt, M., & Molfenter, S. (2004). Beyond test anxiety: Development and validation of the Test Emotions Questionnaire (TEQ). *Anxiety, Stress & Coping, 17*(3), 287-316.
- Pitkäniemi, H. (2017). A teacher's practical theories, self-efficacy, and emotions - What connections do they have, and how can they be developed? *Nordisk Tidskrift för Allmän Didaktik, 3*(1), 2-23.
- Raykov, T., Marcoulides, G. A., & Li, C-H. (2012) Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement, 72*(6), 954-974.
- Stephanou, G., Gkavras, G., & Doulkeridou, M. (2013). The role of teachers' self- and collective-efficacy beliefs on their job satisfaction and experienced emotions in school. *Psychology, 4*(3A), 268-278.
- Sutton, R. E., & Wheatley, K. F. (2003). Teachers' emotions and teaching: A review of the literature and directions for future research. *Educational Psychology Review, 15*(4), 327-358.
- Sutton, R.E. (2004). Emotional regulation goals and strategies of teachers. *Social Psychology of Education, 7*(4), 379-398.
- Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220.
- Tschannen-Moran, M., & Woolfolk-Hoy, A. (2001). Teacher Efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*(7), 783-805.
- Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492.
- 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.

Factor structure and measurement invariance of the TIMSS 2015 mathematics attitude questionnaire: Exploratory structural equation modelling approach

Seyma Uyar ^{1,*}

¹Burdur Mehmet Akif Ersoy University, Faculty of Education, Department of Educational Sciences, Burdur, Turkey

ARTICLE HISTORY

Received: Sep. 18, 2020

Revised: July 31, 2021

Accepted: Sep. 23, 2021

Keywords:

Exploratory structural equation model,
Measurement invariance,
TIMSS,
Multi-group confirmatory factor analysis,

Abstract: In the current study, the appropriateness of the Mathematics Attitude Questionnaire administered to middle school 8th grade students in the TIMSS 2015 application to the exploratory structural equation and confirmatory factor analysis models was examined. The study was conducted on 6079 students making up the sample of Turkey. In the TIMSS 2015 application, the attitude items are presented under four headings called students' interest in mathematics, students' views on engaging teaching in mathematics lessons, students' self-confidence in mathematics, and students' value mathematics. As a result of the investigation of the factor structure of these items, the attitude questionnaire with its 5 factors and 35 items was accepted to be suitable for the Exploratory Structural Equation Model (ESEM). Moreover, invariance of the TIMSS 8th grade mathematics attitude questionnaire depending on gender was investigated at six stages as configural, weak (metric), strong (scalar), strict, variance-covariance, and latent mean invariance through ESEM. It was concluded that the questionnaire satisfied all the invariance conditions.

1. INTRODUCTION

Comparative studies are thought to have a large share in shaping the education policies of countries. For this reason, it is seen that many countries take part in comparative studies, which include international measurement and evaluation studies such as PISA (Programme for International Student Assessment), TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study), and TALIS (The OECD Teaching and Learning International Survey). For example, TIMSS is a survey research conducted by the International Education Achievement Assessment Organization (IEA) for the comparative evaluation of the knowledge and skills acquired by the 4th and 8th grade students in the fields of mathematics and science at four-year intervals. In the TIMSS application, information about students' performances, education systems, curricula, student characteristics, characteristics of teachers, and schools is collected, and student achievement is compared with

*CONTACT: Şeyma Uyar ✉ syuksel@mehmetakif.edu.tr 📍 Burdur Mehmet Akif Ersoy University, Faculty of Education, Department of Educational Sciences, Burdur, Turkey

e-ISSN: 2148-7456 /© IJATE 2021

other countries or in different subgroups constructed based on gender and socioeconomic level in the same sample (Ministry of National Education [MNE], 2016). However, the biggest problem in such studies is whether the measurement tools applied to the compared groups are really equivalent in terms of the measured property. When a cognitive or behavioural feature is to be measured under different conditions (measurement time, test application methods or group), this construct may mean different for each condition (Bornstein, 1995). In this case, it is not easy to completely distinguish the difference between individuals from measurement time, measurement method or group membership (Horn & Mcardle, 1992). In order to be able to compare a construct correctly and appropriately in the given conditions, it is necessary to examine the invariance of the meaning of the construct under these conditions (Putnick & Bornstein, 2016). For example, when a factor structure shows similar conformity for data obtained at different times, it is necessary to talk about a measurement invariance within time (longitudinal) (Little, 2013) or if the factor structure remains the same as a result of administration of a test on the Internet environment or as a paper and pencil test, it is necessary to talk about the invariance in terms of the measurement method (Whitaker & McKinney, 2007). Similarly, when a factor structure remains the same in subgroups constructed on the basis of gender (male-female), country or socioeconomic level for different groups, it may indicate that measurement invariance between groups is ensured (Kline, 2005). In tests, when measurement invariance is achieved, it is possible to base the difference obtained in terms of the measured property on individual characteristics (Başusta & Gelbal, 2015).

Measurement invariance studies are generally carried out within the scope of SEM with multi-group confirmatory factor analysis (MGCFA) or item response theory (IRT) approaches (Chung et al., 2016). In fact, confirmatory factor analysis (CFA) is recommended instead of IRT in order to examine measurement invariance in measurement tools consisting of ordinal items (Stark et al., 2006). CFA is basically a factor analysis and one of the aims of factor analysis is to reveal the validity of the scores. Validity responds to the question of whether the measuring tool provides a score for the desired dimension. At the same time, it questions whether the items in one dimension really and only measure this dimension (Thompson, 2004). When answers are found to these questions, factor analysis can also be the evidence of content validity, construct validity, and even face validity. For this reason, factor analysis is considered to be the heart of psychological constructs (Nunnally, 1978).

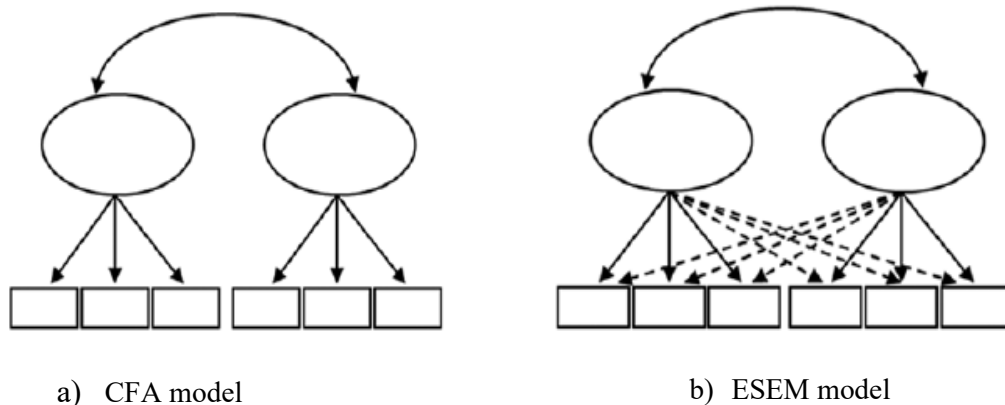
Factor analysis is a method originally developed to explain the characteristics called unobserved (latent) variables or factors underlying a performance related to observed variables. These factors cannot be observed directly, but instead the observed variables are considered to be indicators of latent constructs. Statistically, the purpose of factor analysis is to reveal the maximum variance by identifying new constructs (factors and dimensions) that may occur in fewer numbers through the relationships between observed variables (Brown, 2006; Büyüköztürk, 2002; Özdamar, 2004; Reykov & Marcoulides, 2008). Thus, the fact that factors are interpretable to a large extent on the basis of data and for which reasons observed variables have high levels of correlations with each other can be explained (Reykov & Marcoulides, 2008).

The most used types of factor analysis are Exploratory Factor Analysis (EFA; Jennrich & Sampson, 1966) and Confirmatory Factor Analysis (CFA; Joreskog, 1969) (Özdamar, 2004; Stapleton, 1997). EFA and CFA are basically similar in terms of explaining the observed relationships between indicators with fewer latent variables. However, they are different techniques in terms of the number and nature of the priori features and limitations of the model (Brown, 2006). EFA is thought of as a precursor of CFA used to describe and distinguish basic psychological constructs (Cudeck & MacCallum, 2007). In EFA, the researcher is not expected to determine a construct about the data in advance. Although the researcher has some expectations,

EFA may not provide a suitable model for these expectations, and the analysis process should not be affected by the expectations of the researcher (Brown, 2006; Thompson, 2004). EFA is used as an exploratory and descriptive technique in terms of which observed variables are reasonable indicators of latent dimensions. CFA does not aim to discover or reveal factors, but it is used to verify, test or quantify a hypothetical construct predicted among a set of measurements. In the area of interest of CFA lies the examination of the model for the relationship of factors with each other and with observed variables (Reykov & Marcoulides, 2008).

CFA is a type of structural equation modelling and deals with measurement models. CFA tries to model the relationships between test items, test scores, or the observation levels of behaviour called indicators or observed variables and latent variables or factors (Brown, 2006). In fact, CFA is hypothesis-based in nature. CFA tests a theory, while EFA generates a theory (Brown, 2006; Stapleton, 1997). Cross-loadings allowed between variables by EFA are forced to be zero in CFA. There is a limitation in CFA that argues that the scores are only related to the relevant factor; however, the imposition that a latent variable is only related to the target item and not to other dimensions gives rise to some difficulties for empirical validation. In reality, such a pure relationship is unlikely. Although a factor plays a dominant role in explaining the target observed variable, it is very difficult to say that this item is explained only by the relevant factor. The fact that there is such a restrictive requirement in CFA may cause inflated factor correlations and thus the structural relationships between factors to be damaged (Asparouhov & Muthen, 2009; Marsch et al., 2009). Therefore, it can be seen that a model that is thought to be well defined with EFA is not confirmed by CFA (Guo et al., 2019). In CFA, the suitability of the models is evaluated by looking at the model fit indices. When the indices are not within the acceptable limits, this often causes the absence of confirmation or doubts about the reliability and reproducibility of the models (Asparouhov & Muthen, 2009; Booth & Hughes, 2014). The Explanatory Structural Equation Model (ESEM) is a technique developed to overcome this limitation. Its most important advantage is that it allows different latent variables to cross-load with different items (Asparouhov & Muthen, 2009; Gomes et al., 2017). When the structure of ESEM model is examined in Figure 1 (b), it is seen that ESEM brings together the advantages of EFA and CFA (Marsh et al., 2014).

Figure 1. CFA and ESEM representations for a two-factor model (Booth & Hughes, 2014).



ESEM, like EFA, is flexible when testing measurement models and includes possible rotations for the construct matrix. It tests measurement models of latent variables using EFA instead of CFA. It is a more flexible type of structural equation model that models using an explanatory approach considering which of the factor rotations are appropriate (Schmitt, 2011). In CFA, all parameters are defined a priori by the researcher. It also presents different hypotheses for the relationships between observed and latent constructs. ESEM, on the other hand, requires only the number of factors as a priori knowledge and freely estimates all the other parameters (Booth

& Hughes, 2014). ESEM uses EFA while creating the measurement model that is part of a factor model and calculating the variance of residuals, rotated factor loadings, factor variances, and covariances (for example, regressions of latent factors on independent variables). Unlike CFA, an item does not necessarily load on a single factor. Small but statistically significant cross-loadings are not forced to be zero in the analysis with ESEM.

Existing research shows that ESEM model provides model data fit better than CFA in personality scales (Kristjansson et al., 2011). In addition, some studies show that better model fit coefficients are obtained compared to CFA (Rosellini & Brown, 2011; Mattson, 2012). ESEM models are also expanded to multi-group analysis, allowing factorial structures to be compared in terms of measurement invariance or differential item functioning (Marsch et al., 2009; Tomás, et al., 2014). ESEM is a preferred model to reveal the factor structure when it exhibits better model data fit than CFA (Marsh, et al., 2014). In a study comparing CFA and ESEM in terms of model-data fit, it was stated that the data in the subgroups of culture, socioeconomic level and social capital fit better with ESEM model in the PIRLS 2006 and PISA 2009 applications (Caro et al., 2014). In another study where the factor structure of the Academic Motivation Scale was compared with CFA and ESEM, it was stated that the ESEM approach fit better with the data and the pattern of factor correlations was ranked more appropriately to the theoretical framework (Guay et al., 2015). Joshanloo and Lamers (2016), in their study examining the construct of well-being with CFA and ESEM, revealed that the factor structure in CFA is not clear enough, and that the two dimensions are very well separated with ESEM. Marsh et al. (2011), on the other hand, evaluated the 11-factor construct addressed under two main headings called cognitive and affective related to the academic motivation and responsibility scale. They stated that although the number and pattern of factors obtained with CFA and ESEM were the same, better fit indices were obtained with the ESEM model.

In the literature, it is stated that testing the factor structure of a measurement tool with ESEM is more advantageous than CFA. On the other hand, Marsch et al. (2009) stated that when ESEM and CFA are tested together and the model fit is acceptable for both, it is more appropriate to continue the analysis by using CFA. For this reason, in the current study, it is aimed to determine which of ESEM and CFA models is more suitable for revealing the factor structure and construct validity of TIMSS 8th grade mathematics attitude questionnaire. In previous studies examining the TIMSS questionnaire, it is seen that CFA was used for model testing (Bofah & Hannula, 2015; Ertük & Erdiñç-Akan, 2018; Polat, 2019). In the current study, it is thought that the comparability of the results to be obtained with ESEM with previous studies will contribute to revealing the factor structure of the TIMSS mathematics attitude questionnaire. In the current study, it is also aimed to test the measurement invariance of the factor structure of the TIMSS attitude questionnaire depending on gender with the accepted model. Consistent with the TIMSS 2015, 2011 and 2007 results, it was stated that mathematics achievements of the students who have interest in mathematics, who value mathematics, who have self-confidence in mathematics and who love mathematics in the Turkish sample are high (Raport, 2015). In some studies, it has been seen that the gender factor is important in terms of beliefs about mathematics (Bofah & Hannula, 2015; Simpkins et al., 2005). Watt (2004) stated that girls' interest in mathematics is higher than that of boys, but that there is no difference between their mathematic performances. According to him, gender is more important in affecting students' self. The current study focuses on invariance by gender in order to be a reference for future research and to make more reliable interpretations on the relationship between the mathematics attitude questionnaire and achievement. In this connection, in the current study, answers to the following questions were sought: "Do the Turkish data obtained from the TIMSS 2015 8th grade mathematics attitude questionnaire fit the ESEM and CFA models? Does the questionnaire achieve measurement invariance in female and male student groups?"

2. METHOD

In this study, factor structure of the TIMSS 8th grade mathematics attitude questionnaire and its measurement invariance by gender are examined. The current study is a descriptive research aimed at revealing the existing state.

2.1. Study Group

A total of 238 schools having 8th graders participated in the TIMSS 2015 study from Turkey. In the period when the TIMSS 2015 study was conducted, there was a total of 1,187,893 eighth grade students. The number of students included in the TIMSS sample was 6079. In the Turkish sample, 48.4% (2943) were females and 51.6% (3136) were males. The current study was conducted by using all the data obtained from 6079 students.

2.2. Data Collection Tool

In the TIMSS application, science and mathematics achievement tests and student, teacher, school, and parent questionnaires were used to determine the knowledge and skill levels of the 8th grade students. In the current study, an attitude questionnaire applied to the 8th grade students was used. The questionnaire was administered under four titles “Students’ Interest in Mathematics”, “Students’ Views on Engaging in Mathematics Lessons”, “Students’ Self-Confidence in Mathematics” and “Students’ Value Mathematics”. There is a total of 37 items in the questionnaire including 9 items (2 are negative) to measure students interest in mathematics, 10 items to measure their views on engaging teaching in mathematics lessons, 9 items for their self-confidence in mathematics (5 are negative), and 9 items to measure their value of mathematics. The scale items are in the form of 4-point Likert scale with the following response options: ‘1’ strongly disagree, ‘2’ a little disagree, ‘3’ a little agree, and ‘4’ strongly agree.

2.3. Data Analysis

Before starting the analyses, missing data and assumptions were examined and the data were made ready for analysis. In the TIMSS 8th grade data, there was missing data less than 1% in each variable. In all the data, the rate of missing data was 1.02%. This rate is lower than 5% (Çokluk et al., 2010; Raykov & Marcoulides, 2008). However, it is necessary to investigate whether the missing data are completely random or random and which technique will be applied to the missing data should be determined. In studies, it is stated that such data should be approached with caution regardless of the method (Allison, 2003; Graham, 2012; Little & Robin, 1987). The distribution of the missing data can be determined with the Little Missing Completely at Random (MCAR) test. In the current study, the MCAR test was found to be significant, meaning that the data were not completely random. In this case, assigning data can be seen as a more reliable method than deleting the data. For this reason, missing data were assigned with Expectation Maximization (EM) (Allison, 2003).

Since CFA and ESEM, two types of the structural equation modelling as multivariate statistical methods, are used in the current study, it is necessary to check whether the data can satisfy the multivariate normality assumption or not. Therefore, the data were evaluated in terms of univariate and multivariate outliers, univariate and multivariate normality, linearity, covariance, and multicollinearity. In order to determine the univariate outliers, the z values of the variables were calculated and those outside the ± 3 range were examined. It was seen that there were not any univariate outliers in the data. For multivariate outliers Mahalanobis distances were calculated and it was found that Mahalanobis values did not exceed the critical chi-square value at $p < 0.001$ ($\chi^2_{36,0.001} = 66.62$). For multicollinearity, the following conditions were taken into consideration: a confidence interval-CI value lower than 30, variance inflation factor (VIF) lower than 10, and the tolerance values greater than .20 (Tabachnick & Fidell, 2007). No mul-

ticollinearity was found in the data. However, based on the Levene test results, it was determined that some items did not provide covariance between groups. On the other hand, since most items had skewness and kurtosis values outside the range ± 1 , it was accepted that the normality assumption was not satisfied in the current study. For this reason, Robust Maximum Likelihood (MLR) estimation method was used in the Mplus 7.0 program.

In order to examine the factor structure of the TIMSS mathematics attitude questionnaire, the models analysed by CFA and ESEM were compared. The TIMSS questionnaire items were presented to the students under four headings. Therefore, in the current study, the questionnaire items were tested as 3-factor, 4-factor and 5-factor ESEM and 4-factor and 5-factor CFA and the results were compared. In order to decide the model having the best fit, Bayesian Information Criterion (BIC) values (Kuha, 2004), differences between adjusted chi-square (*adjusted χ^2*) values (Asparauhov & Muthen, 2006), fit indices, and factor loadings were examined. In addition, the level of correlation between the factors was also taken into account. In order to evaluate model fit, the following goodness-of-fit indices can be used: χ^2 , χ^2 /degree of freedom (χ^2/sd), root-mean-square error of approximation (RMSEA), the comparative fit index (CFI), goodness of fit index (GFI), adjusted goodness fit index (AGFI), the Tucker–Lewis index (TLI), and standard root mean square (SRMR). Although the use of such a variety of indices (especially when they take different values) creates a conflict about the fit of the model with the observed data, it can be decided about model fit by considering some suggested value ranges (Schermelleh-Engel et al., 2003). In addition, some studies indicate that CFI, TLI, and RMSEA indices are independent of the sample size (Hu & Bentler, 1995; Marsch et al., 2005). In the current study, it was decided that the model would be acceptable if the value of RMSEA and SRMR was smaller than 0.05 and CFI value was greater than 0.90, and TLI value was greater than 0.90 (Browne & Cudeck, 1993; Schermelleh-Engel & Moosbrugger, 2003). It was also accepted that the one with a lower BIC value of the two compared models would be the one with a better fit (Krueger et al., 2007). Moreover, it was decided that the model in which the correlations between the factors are smaller than 0.70 would be accepted to have a better fit with the data (Marsh et al., 2011; Guay et al., 2015).

After the selection of the model, the invariance of the attitude questionnaire was examined in the male and female groups. The females were taken as the reference group. By imposing restrictions in parameters in the males, the modal invariance was investigated. At this stage, the limited model and the less limited model were compared in terms of fit. The stages of the measurement invariance were hierarchically investigated as configural, weak, strong, strict, variance-covariance, and latent mean invariance (Guay et al., 2015; Marsch et al., 2010; Meradith, 1993; Morin & Maïano, 2011; Steenkamp & Baumgartner, 1998).

In the configural invariance stage, it is examined whether the factor model is equal for groups or not. In other words, factor loadings, factor means (intercept), and error variances are set free in both groups, latent variances are equated to 1, and latent means are equated to 0 in the reference group. In the weak invariance stage, as different from the previous model, restriction of equality between groups is imposed on factor loadings and cross-loadings. In the strong invariance stage, along with factor loadings, factor means are also restricted and forced to be equal between the groups compared. In the strict invariance stage, restriction of equality between groups is imposed on measuring errors at item level. In fact, the invariances ensured up to this stage prove that the properties of a measurement tool are the same between groups. In the current study, the model continued to be restricted and variance/covariance invariance was examined by restricting the variance/covariance matrix to be equal to 1 in all groups. In the last stage, the latent mean invariance was equated to 0 and latent mean invariance test was conducted (Morin & Maïano, 2011).

In order to find evidence for invariance, the differences between χ^2 , CFI, TLI, and RMSEA values obtained from hierarchical models can be used. Since the MLR was used to estimate the parameters in the current study, Satorra-Bentler $\chi^2 (S - B_{\chi^2})$ value was obtained. For this reason, in order to calculate the difference between χ^2 values, it is necessary to calculate TR_d values by making adjustments. When the obtained TR_d value is greater than the critical value at the relevant degree of freedom and 0.05 significance level, the null hypothesis is rejected and it is interpreted that the models are different from each other (Asparouhov & Muthen, 2010; Bryant & Satorra, 2012; Satorra & Bentler, 2010). The formula in Equation 1 is used in calculating the TRd value.

$$c_d = [(d_0 * c_0) - (d_1 * c_1)] / (d_0 - d_1) \tag{1}$$

$$TR_d = [(T_0 * c_0) - (T_1 * c_1)] / c_d$$

d_0 : degree of freedom obtained for the restricted model, c_0 : scaling factor of the restricted model, d_1 : degree of freedom of the compared model, and c_1 : scaling factor of the compared model. T_0 : $S - B_{\chi^2}$ value of the restricted model; T_1 : $S - B_{\chi^2}$ value of the compared model.

It is stated in the literature that CFI and RMSEA values are more reliable than χ^2 because they are not sensitive to the sample size. For this reason, it is appropriate to evaluate other fit indices together with χ^2 in model comparisons. Chen (2007) states that measurement invariance can be achieved when the decrease in the CFI value is .01 or less, or the increase in the RMSEA value is .015 or less. These values are suggested for the Maximum Likelihood (ML) estimation; however, Sas et al. (2014) showed that CFI and RMSEA values gave similar results to ML in MLR method. Therefore, in model comparisons, besides χ^2 test, ΔCFI and $\Delta RMSEA$ values were examined and thus decision was made about the invariance (Guay et al., 2015; Jung, 2019).

3. FINDINGS

3.1. Model Fit of the TIMSS 8th Grade Attitude Questionnaire

The TIMSS 8th grade attitude questionnaire items were presented to the students under four different headings. Therefore, Table 1 presents the goodness-of-fit indices of the models tested as four-factor models and of the models constructed as alternatives.

Table 1. Goodness-of-fit indices obtained for the alternative models and information criteria

	χ^2/sd	CFI	TLI	RMSEA	BIC	SRMR	<i>sf</i>
3-factor ESEM	14119.9/558	0.853	0.825	0.063	515966.04	0.042	1.318
4-factor ESEM	8075.98/524	0.918	0.896	0.049	508445.10	0.027	1.337
5-factor ESEM	5246.81/491	0.948	0.930	0.040	504986.23	0.020	1.343
4-factor CFA	14636.47/623	0.848	0.838	0.061	516245.98	0.072	1.329
5-factor CFA	10046.53/619	0.898	0.890	0.05	510174.40	0.059	1.329

sf: scaling factor

As can be seen in Table 1, the model having the highest goodness-of-fit indices is the 5-factor ESEM model ($\Delta\chi^2 = 4944,33$; $\Delta df = 128, p < .05$; RMSEA = 0.040, CFI = 0.948, TLI = 0.930, and SRMR = 0.020). At the same time, the lowest BIC value (BIC = 504986.23) was obtained in this model. After this model, the model having the best fit is the 5-factor CFA model (RMSEA = .05, CFI = .898, TLI = .890 and SRMR = .059). As the TR_d value indicating the $S - B_{\chi^2}$ difference between two models was calculated to be 4944.33, it can be argued that

there is a significant difference between the models ($\Delta\chi^2(128) = 155.40, p < .05$). In order to decide on the final situation, factor loadings obtained by the 5-factor ESEM and 5-factor CFA and correlations between factors were compared and the analysis results are given in Table 2 and Table 3.

Table 2. Factor loadings and cross loadings of the mathematics attitude questionnaire obtained with the 5-factor ESEM and 5-factor CFA.

Factors	Item	F1	F2	F3	F4	F5	CFA
Students' interest in mathematics	1	0.721	0.036	-0.017	0.066	0.037	0.805
	2	-0.527	0.010	0.361	0.167	-0.042	-0.561
	3	-0.622	-0.017	0.374	0.133	-0.006	-0.679
	4	0.463	0.111	0.086	0.039	0.084	0.548
	5	0.849	-0.020	-0.046	0.037	0.011	0.880
	6	0.674	0.012	0.079	0.089	0.019	0.713
	7	0.662	-0.030	-0.010	0.215	0.004	0.811
	8	0.700	0.061	-0.009	0.130	-0.026	0.810
	9	0.696	0.011	-0.106	0.177	-0.016	0.868
Students' views on engaging in mathematics les- sons	1	-0.041	0.465	-0.013	0.206	0.043	0.549
	2	0.025	0.724	-0.033	0.030	-0.023	0.741
	3	0.179	0.477	-0.002	0.077	0.090	0.655
	4	0.095	0.337	0.155	0.227	-0.011	0.466
	5	0.009	0.757	-0.034	0.026	0.013	0.779
	6	0.051	0.770	-0.033	-0.041	-0.011	0.773
	7	-0.016	0.705	0.039	0.014	0.003	0.693
	8	0.085	0.649	0.050	-0.007	-0.008	0.682
	9	-0.053	0.746	-0.010	-0.001	0.023	0.719
	10	-0.020	0.752	-0.009	-0.036	0.028	0.732
Students' mathe- matics anxiety	1	0.110	0.016	0.651	-0.195	-0.010	0.685
	2	-0.086	0.005	0.609	-0.208	-0.033	0.766
	3	0.059	-0.088	0.616	0.031	0.020	0.543
	4	-0.045	0.026	0.684	-0.153	0.019	0.785
	5	-0.125	-0.032	0.684	-0.100	0.016	0.800
Students' self- confidence in mathematics	1	0.174	0.043	-0.074	0.605	0.036	0.820
	2	0.103	0.075	-0.020	0.652	0.050	0.798
	3	0.064	-0.023	-0.009	0.752	0.020	0.773
	4	-0.002	0.151	-0.037	0.694	0.003	0.761
Students' value mathematics	1	0.296	0.168	0.058	0.000	0.295	0.587
	2	0.223	0.068	0.114	0.030	0.440	0.630
	3	-0.052	-0.035	-0.023	0.081	0.733	0.695
	4	-0.055	-0.035	-0.026	0.089	0.767	0.733
	5	0.214	-0.064	-0.008	0.303	0.354	0.627
	6	0.140	0.002	-0.006	-0.032	0.714	0.786
	7	0.061	0.022	-0.012	-0.039	0.720	0.740
	8	-0.111	0.100	0.038	0.028	0.521	0.493
	9	0.048	0.126	-0.032	-0.071	0.631	0.679
Cross Loadings		$\overline{ X } = .085$			$\overline{SD} = .109$		

F1: Factor 1, F2: Factor2, F3: Factor 3, F4: Factor4, F5: Factor5

When the factor loadings obtained with ESEM in Table 2 are examined, it is seen that the factor loadings of the items in the first factor called “students’ interest in mathematics” vary between .463 and .849. The cross-loadings of these items on the other factors are close to zero. In other words, these items do not exhibit high loading values in the other factors. The factor loadings of the items in the second factor called “students’ views on engaging in mathematics lessons” were found to be ranging from 0.337 to 0.752. The cross-loadings of these items are close to zero. There are five items in the third factor called “mathematics anxiety” and the factor loadings of these items were found to be ranging from 0.609 to 0.684. These items exhibit factor loadings close to zero in the other factors. According to the results of ESEM, in the fourth dimension called “students’ self-confidence in mathematics”, there are four items and the factor loadings of these items were found to be ranging from 0.605 to 0.752. Finally, in the fifth factor called “students’ value mathematics”, there are 9 items and the factor loadings of the items were found to be ranging from 0.295 to 0.767. However, two items in this factor were found to be exhibiting high factor loadings in the first factor called “students’ interest in mathematics” (main factor loading value is 0.295, cross-loading value is 0.296). The fifth item in the factor called “students’ value mathematics” exhibits a similar loading value in the factor called “students self-confidence in mathematics” and shows a high cross-loading value (main factor loading is 0.354, cross loading value is 0.303). This situation casts doubt on the validity of both of the items. According to Marsch et al. (2011) cross-loadings of items should be as close to zero as possible. The discriminant validity is considered to be poor if the cross-loading of the item moves away from zero and gives a high loading on a factor other than its own (Hair et al., 2010). It may be appropriate to remove such items from the measurement tool. At this point, although ESEM follows a strict path in eliminating weak items, it actually wants to increase the validity of the factors. The mean cross-loading of factor loadings obtained by ESEM as absolute value is .085, while the standard deviation is .109. When the factor loadings obtained with CFA are examined, it is seen that the lowest factor loading is .548, while the highest factor loading is .880. The factor loadings obtained with CFA are generally higher than the factor loadings obtained with ESEM. Correlations between the factors in the model are given in Table 3.

Table 3. Correlations between the factors obtained with ESEM and CFA.

	Factor	Factor 2	Factor 3	Factor 4	Factor 5
ESEM	1	0.472	-0.332	0.660	0.518
	2		-0.043	0.328	0.450
	3			-0.410	-0.097
	4				0.430
CFA	1	0.531	0.810	-0.565	0.640
	2		0.482	-0.174	0.552
	3			-0.611	0.597
	4				-0.262

When the correlation values given in Table 3 are examined, it is seen that the correlation values between the first and fourth factors (-.565), between the second and fourth factors (-.174), and between the fourth and fifth factors (-.262) in the CFA model are lower compared to those of ESEM. The correlation found between the first and third factors in the CFA is .810, while the same correlation was calculated to be -.332 in ESEM. The high correlation between the two factors in the CFA model creates doubts about whether they measure similar features and whether the model has five or four factors. For this reason, after examining factor loadings and correlation values, the ESEM model was accepted, but two items, "mathematics will help me"

and "job involving mathematics", were excluded from the analysis due to their high cross-loadings. The fit indices obtained after the items were removed are shown in Table 4.

Table 4. Fit indices obtained for the five-factor and 35-item TIMSS 8th grade attitude questionnaire with ESEM

	CFI	TLI	RMSEA	BIC	SRMR	sf.
ESEM (35 items)	0.953	0.935	0.039	474776.122	0.019	1.350

When the fit indices given in Table 4 are examined, it can be said that ESEM is consistent with the data (RMSEA = 0.039, CFI = 0.953, TLI = 0.935 and SRMR = 0.019). Therefore, it was accepted that the TIMSS 8th grade mathematics attitude questionnaire conforms to ESEM with 35 items and 5 factors and the measurement invariance of the questionnaire between male and female groups was examined with ESEM. The results obtained are given in Table 5.

3.2. Measurement Invariance in terms of Gender

As can be seen in Table 5, when the fitting of the ESEM model in male and female groups is examined, it can be said that the model is acceptable (for males: RMSEA = 0.097, CFI = 0.959, TLI = 0.943, SRMR = 0.018; for females: RMSEA = 0.043, CFI = 0.947, TLI = 0.926, SRMR = 0.02). Since the model was confirmed separately in each group, the invariance phase was initiated.

Table 5. Fit indices obtained for gender groups and invariance stages with ESEM.

	$S - B_{\chi^2}$	df	sf	CFI	TLI	RMSEA	SRMR	ΔCFI	$\Delta RMSEA$
Males	2241.46	430	1.361	0.959	0.943	0.037	0.018	-	-
Females	2750.71	430	1.304	0.947	0.926	0.043	0.020	-	-
M1	4981.13	860	1.332	0.953	0.935	0.040	0.019	-	-
M2	5418.74	895	1.319	0.948	0.931	0.041	0.025	-0.004	0.001
M3	5610.14	1040	1.317	0.948	0.940	0.038	0.025	0	-0.003
M4	5887.86	1075	1.337	0.945	0.939	0.038	0.028	-0.003	0
M5	6100.86	1090	1.337	0.943	0.937	0.039	0.048	-0.002	0.001
M6	6233.90	1095	1.335	0.941	0.936	0.039	0.046	-0.002	0

df: degree of freedom M1: Configural invariance, M2: Weak invariance, M3: Strong invariance, M4: Strict invariance, M5: Variance/covariance invariance, M6: Latent mean invariance

When the first stage (M1); configural invariance is examined, it can be said that the model is confirmed and the configural invariance is accepted (RMSEA = .040, CFI = .953, TLI = .935, SRMR = .019). It can be said that in the weak invariance stage (M2), fit indices are within the acceptable ranges (RMSEA = .041, CFI = .948, TLI = .931, SRMR = .025). The TR_d value for the chi-square difference between the two models was calculated to be 512.67. Accordingly, it can be said that there is a significant difference between the two models ($\Delta\chi^2(35) = 49.80, p < .05$). However, the difference between fit indices obtained in configural and weak invariance stages is acceptable ($\Delta CFI = -.004, \Delta RMSEA = .001$). This shows that weak invariance is achieved. When M3 (strong invariance) is examined for the third stage, it can be said that fit indices of the model are at an acceptable level (RMSEA = .038, CFI = .948, TLI = .940, SRMR = .025). TR_d value between M3 and M2 models was calculated to be 184.90. Accordingly, there is a significant difference between the two models ($\Delta\chi^2(145) = 174.1, p < .05$). When the change of fit indices is examined in addition to χ^2 difference value, it can be

said that the difference between the models is not significant ($\Delta CFI = 0$, $\Delta RMSEA = -.003$). In this case, it can be concluded that the strong invariance of the TIMSS attitude items is tenable across male and female student groups. In the fourth stage, after the restriction of equality of error variances between the groups is imposed, it can be stated that the fit indices of the strict invariance (M4) model remain within acceptable ranges (RMSEA = .038, CFI = .945, TLI = .939, SRMR = .028). The TR_d value between M4 and M3 was calculated to be 250.36 and a significant difference was found between the models as ($\Delta\chi^2(35) = 49.80$). However, the difference of the fit indices between M4 and M3 provides support for strict invariance across gender ($\Delta CFI = -.003$, $\Delta RMSEA = 0$). When the fit indices obtained for M5 are examined at the next stage, it can be said that the model produces acceptable fit (RMSEA = .039, CFI = .943, TLI = .937, SRMR = .048). TR_d value calculated between M5 and M4 was found to be 213.0. Although, there is a significant difference between the two models ($\Delta\chi^2(15) = 24.99$), the change of fit indices are within the acceptable range ($\Delta CFI = -.002$, $\Delta RMSEA = .001$). The fit indices belonging to M6 model constructed to test whether latent means are invariant or not can be said to be in acceptable ranges (RMSEA = .039, CFI = .941, TLI = .936, SRMR = .046). When M6 and M5 are compared, it can be said that the change in M6 is not significant compared to M5. Thus, latent mean invariance is achieved: ($\Delta CFI = -.002$, $\Delta RMSEA = 0$).

4. DISCUSSION and CONCLUSION

In the current study, the fit of TIMSS 8th grade mathematics attitude questionnaire to the ESEM and CFA models was examined in the first stage and it was seen that the data fit the 5-factor ESEM model the best. ESEM is a model that allows cross-loadings. In CFA, an item's factor loading (even very small values) is forced to be zero in factors other than its own factor. This may cause high correlation values between factors and poor model fit (Marsh et al., 2009). Stromeier et al. (2015) stated that cross-loadings allow for better model estimation based on such indicative information rather than causing pollution to the structure. As a matter of fact, in the current study, the 5-factor ESEM model exhibited fit better than the 5-factor CFA and 4-factor CFA models. In addition, the correlation between the 2nd and 3rd factors in the CFA model was found to be high as 0.81. This might be because of the restriction on cross-loadings. Failure to detect very small cross-loadings by CFA may cause bias in correlation values between factors (Jung, 2019). On the other hand, in the 5-factor ESEM model, the cross-loading values of 2 items move slightly away from 0 and give rise to doubts on the discriminant validity of these items. After these two items were removed from the analysis, it was concluded that ESEM was compatible with the data. Polat (2019) stated that the TIMSS 8th grade mathematics affective feature model fits CFA with 4 factors and 34 items. However, in order to be able to decide this in his study, he first applied EFA to the data and decided to exclude 3 items from the analysis according to EFA results. He stated that the model was confirmed with 4 factors and 34 items by applying CFA after the items were removed. In the model he established, he reported the correlation between the dimensions of "students like learning mathematics" and "students are confident in mathematics" as .85. Although the removal of items contributes to the model, the level of correlation between dimensions remains high. This situation can be considered as an indication that the factor loadings are affected by the cross-loadings being forced to zero. Guay et al. (2015) stated that the correlations between factors in the motivation scale are lower than they are in CFA. In the TIMSS 8th grade science attitude questionnaire, which Jung (2019) examined as 3 factors, the correlation values between the dimensions of 'students' like learning science', 'students' confidence in science' and 'students' value science' remained at low levels compared to CFA (the highest correlation with ESEM is .762; the highest correlation is .823 with CFA). The correlation results obtained in the current study are consistent with the related studies in the literature. On the other hand, when cross-loadings were examined with the ESEM model, it was concluded that the cross-loadings of the items "mathematics will help me" and

"job involving mathematics" in the factor called students value mathematics were high. Therefore, these two items were excluded from the analysis. The item "mathematics will help me" was excluded from the analysis because it was in a different dimension according to EFA results in the study conducted by Polat (2019). In addition, Polat (2019) concluded that two items in the dimension of "self-confidence in mathematics" should be removed from the analysis as a result of EFA. In the current study conducted with ESEM, the dimension of "students' self-confidence in mathematics" was divided into two different dimensions according to ESEM results and was named as "students' self-confidence in mathematics" and "students' mathematics anxiety". For this reason, different from the study of Polat (2019), there was no need for item exclusion.

In the current study, since it was decided that ESEM was the more suitable model, measurement invariance was carried out with ESEM. The TIMSS 8th grade mathematics attitude questionnaire achieved configural, weak, strong, strict, variance/covariance, and latent mean invariance. According to the results obtained, it can be said that the factor structure, factor loadings, factor means, errors, variance/covariance matrix, and latent means of the questionnaire are similar in male and female groups, so it can be used safely in studies related to gender (Guay et al., 2015). TIMSS questionnaires have been the subject of many measurement invariance studies. Polat (2019) examined the TIMSS 8th grade mathematics affective model in the Turkish data as 4 factors and 34 items with MGCFA. He stated that all the invariance stages were satisfied between the genders in the study. Jung (2019) examined the invariance of the TIMSS 2015 8th grade science attitude questionnaire between genders with three-factor ESEM in American data. In the study, configural, weak, and strong invariance was examined and it was stated that these three stages were satisfied. Ertürk and Erdinç-Akan (2018) examined three of the affective characteristics affecting mathematics separately in the TIMSS 2015 4th grade Turkish data; namely, "like learning mathematics", "interest in mathematics" and "self-confidence in mathematics", and the invariance of each variable depending on gender. As a result of the study, strict invariance condition was met for only the variable of "like learning mathematics". It was observed that the variables of "interest in mathematics" and "self-confidence in mathematics" achieved configural invariance. Bofah and Hannula (2015) stated that the TIMSS 2011 scale, which consisted of items such as "like learning mathematics", "value mathematics", "self-confidence in mathematics", "teacher responses", and "parent participation" achieved configural, weak and strong invariance in ten different countries in male and female groups. Marsh et al. (2013) examined the factor structure of the TIMSS 2007 mathematics and science motivation scale in the Arab countries of Saudi Arabia, Oman, Egypt, and Jordan, England, Scotland, Australia, and the U.S. with CFA. They also examined the invariance of the questionnaire between genders for each country and stated that the scale fulfils the requirement of complete invariance. On the other hand, it was stated that in the Turkish data in PISA 2012, another international exam, the learning model fulfilled all the conditions of invariance (Kıbrıslıoğlu, 2015). Unlike these results, according to PISA 2015 Turkish data, Güngör and Atalay Kabasakal (2019) stated that the science motivation and self-efficacy model and Uyar and Kaya-Uyanık (2019) stated that the science learning model did not achieve invariance in relation to gender. In the studies conducted, it is stated that although it is not clear whether the questionnaires used in international studies achieve the invariance in terms of gender, the measurement invariance for gender in mathematics questionnaires has been achieved in general. It is seen that the results obtained in the current study are consistent with the studies in the related literature.

Failure to achieve measurement invariance for a measurement tool may prevent a healthy comparison of scores. For this reason, researchers are recommended to examine the invariance of the measurement tools used between the groups they will compare. In addition, it may be suggested that they take into consideration the advantages of ESEM and use ESEM in studies where EFA and CFA will be used together. In the current study, invariance only in terms of

gender was examined. In future studies, invariance in sub-groups such as country, socio-economic level or geographical region can be examined with the ESEM model. This study focused solely on the questionnaire applied to 8th grade students. Since TIMSS is an evaluation study that is also applied to 4th graders, it may be suggested to examine the invariance of questionnaires related to mathematics and science among younger students with ESEM. Researchers need to apply EFA followed by CFA in scale development or adaptation studies. ESEM can be recommended to researchers as it can provide information about the model at once instead of analyzing it twice.

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

Seyma UYAR  <https://orcid.org/0000-0002-8315-2637>

5. REFERENCES

- Asparouhov, T., & Muthén, B. (2006). Robust chi square difference testing with mean and variance adjusted test statistics. *Mplus Web Notes*, 10.
- Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, 16, 397–438. <https://doi.org/10.1080/10705510903008204>
- Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, 112(4), 545. <https://doi.org/10.1037/0021-843X.112.4.545>
- Asparouhov, T; Muthén, B. (2010). Computing the strictly positive Satorra-Bentler chi-square test in Mplus. *Mplus Web Notes*, 12, 1-12.
- Başusta, N. B., & Gelbal, S. (2015). Gruplararası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği [Examination of measurement invariance at groups' comparisons: A study on PISA student questionnaire]. *Hacettepe University Journal of Education*, 30(4), 80-90. <http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/1773-published.pdf>
- Bofah, E. A. T., & Hannula, M. S. (2015). TIMSS data in an African comparative perspective: Investigating the factors influencing achievement in mathematics and their psychometric properties. *Large-Scale Assessments in Education*, 3(1), 1-36. <http://dx.doi.org/10.1186/s40536-015-0014-y>
- Booth, T., & Hughes, D. J. (2014). Exploratory structural equation modeling of personality data. *Assessment*, 21(3), 260-271. <https://doi.org/10.1177/1073191114528029>
- Bornstein, M. H. (1995). Form and function: Implications for studies of culture and human development. *Culture & Psychology*, 1(1), 123-137. <https://doi.org/10.1177/1354067X9511009>
- Brown, T. (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Sage.
- Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(3), 372-398. <https://doi.org/10.1080/10705511.2012.687671>
- Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı [Factor analysis: Basic concepts and using to development scale]. *Educational Administration in Theory and Practice*, 8(4), 470-483. <https://dergipark.org.tr/tr/pub/kuey/issue/10365/126871>
- Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2014). Cultural, social, and economic

- capital constructs in international assessments: An evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, 25(3), 433-450. <https://doi.org/10.1080/09243453.2013.812568>
- 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>
- 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. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 25(10), 2559–2564. <https://doi.org/10.1007/s11136-016-1300-z>
- Cudeck, R., & MacCallum, R. C. (Eds.). (2007). *Factor analysis at 100: Historical developments and future directions*. Lawrence Erlbaum
- Çokluk, Ö., Şekercioglu, G. & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. Pegem Akademi.
- Ertürk, Z., & Erdinç-Akan, O. (2018). TIMSS 2015 matematik başarıları ile ilgili bazı değişkenlerin cinsiyete göre ölçme değişmezliğinin incelenmesi [The investigation of measurement invariance of the variables related to TIMSS 2015 mathematics achievement in terms of gender]. *Journal of Theoretical Educational Science*, 204-226. <https://dergipark.org.tr/pub/akukeg/issue/40520/412604>
- Gomes, C., Almeida, L., & Nunez, J. (2017). Rationale and Applicability of Exploratory Structural Equation Modeling (ESEM) in psychoeducational contexts. *Psicothema*, 29(3), 396-401.
- Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.
- Guay, F., Morin, A., Litalien, D., Valois, P., & Vallerand, R. (2015). Application of Exploratory Structural Equation Modeling to Evaluate the Academic Motivation Scale. *The Journal of Experimental Education*, 83(1), 51-82. <https://doi.org/10.1080/00220973.2013.876231>
- Guo, J. M. (2019). A Systematic evaluation and comparison between exploratory structural equation modeling and bayesian structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 26, 529-556. <https://doi.org/10.1080/10705511.2018.1554999>
- Guo, J., Parker, H., Dicke, P., Lüdtke, T., & Diallo, T. (2019). A systematic evaluation and comparison between exploratory structural equation modeling and bayesian structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 26 (4), 529-556. <https://doi.org/10.1080/10705511.2018.1554999>
- Güngör, M & Atalay Kabasakal, K. (2020). Investigation of measurement invariance of science motivation and self-efficacy model: PISA 2015 Turkey sample. *International Journal of Assessment Tools in Education*, 7(2), 207-222. <https://doi.org/10.21449/ijate.730481>
- 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, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Sage.
- Jennrich, R. I. and Sampson, P. F. (1966). Rotation to simple loadings. *Psychometrika*, 31(3), 313–323. <https://link.springer.com/article/10.1007/BF02289465>
- Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.
- Joshanloo, M., & Lamers, S. M. (2016). Reinvestigation of the factor structure of the MHC-SF in the Netherlands: Contributions of exploratory structural equation modeling. *Personality and Individual Differences*, 97, 8-12. <https://doi.org/10.1016/j.paid.2016.02>

089

- Jung, J. Y. (2019): *A Comparison of CFA and ESEM approaches using TIMSS science attitudes items: evidence from factor structure and measurement invariance*. [Master's Thesis, Purdue University]. Purdue University Graduate School, Department of Educational Studies, <https://doi.org/10.25394/PGS.7995890.v1>
- Kıbrıslıoğlu, N. (2015). *The investigation of measurement invariance PISA 2012 mathematics learning model according to culture and gender: Turkey-China (Shanghai)-Indonesia* [Master's Thesis]. Hacettepe University.
- Kline, R. B. (2005). *Methodology in the social sciences. Principles and practice of structural equation modeling* (2nd ed.). Guilford Press.
- Kristjansson, S. D., Pergadia, M. L., Agrawal, A., Lessov-Schlaggar, C. N., McCarthy, D. M., Piasecki, T. M. & Heath, A. C. (2011). Smoking outcome expectancies in young adult female smokers: Individual differences and associations with nicotine dependence in a genetically informative sample. *Drug and Alcohol Dependence*, 116, 37-44. <https://doi.org/10.1016/j.drugalcdep.2010.11.017>
- Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., & Kramer, M. D. (2007). Linking antisocial behavior, substance use, and personality: an integrative quantitative model of the adult externalizing spectrum. *Journal of Abnormal Psychology*, 116(4), 645. <https://doi.org/10.1037/0021-843X.116.4.645>
- Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford press.
- Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.
- Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J. S., Abdelfattah, F., Leung, K. C., Xu, M. K., Nagengast, B., & Parker, P. (2013). Factorial, convergent, and discriminant validity of timss math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology*, 105(1), 108–128. <https://doi.org/10.1037/a0029907>
- Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of Fit in Structural Equation Models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (p. 275–340). Lawrence Erlbaum Associates Publishers.
- Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J., & Nagengast, B. (2011). Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment*, 29(4), 322-346. <https://doi.org/10.1177/0734282911406657>
- Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, 22(3), 471. <https://doi.org/10.1037/a0019227>
- Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85-110. <https://doi.org/10.1146/annurev-clinpsy-032813-153700>
- Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 439-476. <https://doi.org/10.1080/10705510903008220>
- Mattsson, M. (2012). Investigating the factorial invariance of the 28-item DBQ across genders and age groups: an exploratory structural equation modeling study. *Accident Analysis &*

- Prevention*, 48, 379-396. <https://doi.org/10.1016/j.aap.2012.02.009>
- 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(4), 525-543. <https://link.springer.com/article/10.1007/BF02294825>
- Morin, A. J. S., & Maïano, C. (2011). Cross-validation of the short form of the physical self-inventory (PSI-S) using exploratory structural equation modeling (ESEM). *Psychology of Sport and Exercise*, 12, 540–554. <https://doi.org/10.1016/j.psychsport.2011.04.003>
- Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill
- Özdamar, K. (2004). *Paket Programlar ile istatistiksel veri analizi* (Çok değişkenli analizler). Kaan Kitabevi.
- Polat, M. (2019). *The investigation of measurement invariance of TIMSS-2015 mathematics and science affective characteristics models according to culture, gender and statistical region* [Master's Thesis], Hacettepe University.
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71-90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Rencher, A. (2002). *Methods of multivariate analysis*. John Wiley & Sons, Inc.
- Raykov, T., & Marcoulides, G. (2008). *Introduction to applied multivariate analysis*. Routledge Taylor & Francis Group.
- Rosellini, A. J., & Brown, T. A. (2011). The NEO Five-Factor Inventory: Latent structure and relationships with dimensions of anxiety and depressive disorders in a large clinical sample. *Assessment*, 18(1), 27-38. <http://dx.doi.org/10.1177/1073191110382848>
- Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data with a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling*, 21, 167-180. <https://doi.org/10.1080/10705511.2014.882658>
- Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, 75(2), 243–248. <http://dx.doi.org/10.1007/s11336-009-9135-y>
- 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. <https://psycnet.apa.org/record/2003-08119-003>
- Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, 29(4), 304-321. <https://doi.org/10.1177/0734282911406653>
- Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2005). Parents' socializing behavior and children's participation in math, science, and computer out-of-school activities. *Applied Developmental Science*, 9(1), 14-30. https://doi.org/10.1207/s1532480xads0901_3
- Stark, S., Chernshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91(6), 1292–1306.
- Stapleton, C. (1997). *Basic concepts and procedures of confirmatory factor analysis*. Paper presented at the annual meeting of the Southwest Educational Research Association January 23-25. Austin.
- Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25(1), 78-90.

- Stromeyer, W. R., Miller, J. W., Sriramachandramurthy, R., & DeMartino, R. (2015). The prowess and pitfalls of Bayesian structural equation modeling: Important considerations for management research. *Journal of Management*, 41(2), 491-520. <https://doi.org/10.1177/0149206314551962>
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5, pp. 481-498). Pearson.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association.
- Tomás, I., Marsh, H. W., González-Romá, V., Valls, V., & Nagengast, B. (2014). Testing measurement invariance across Spanish and English versions of the Physical Self-Description Questionnaire: An application of exploratory structural equation modeling. *Journal of Sport and Exercise Psychology*, 36(2), 179-188. <https://doi.org/10.1123/jsep.2013-0070>
- Uyar, Ş, Kaya Uyanık, G. (2019). Fen bilimlerine yönelik öğrenme modelinin ölçme değişmezliğinin incelenmesi: PISA 2015 örneği [Investigating measurement invariance of learning model towards science: PISA 2015 sample] *Kastamonu Education Journal*, 27(2), 497-507. <https://doi.org/10.24106/kefdergi.2570>
- Watt, H. M. (2004). Development of adolescents' self-perceptions, values, and task perceptions according to gender and domain in 7th-through 11th-grade Australian students. *Child Development*, 75(5), 1556-1574.
- Whitaker, B. G., & McKinney, J. L. (2007). Assessing the measurement invariance of latent job satisfaction ratings across survey administration modes for respondent subgroups: A MIMIC modelling approach. *Behavior Research Methods*, 39(3), 502-509. <https://doi.org/10.3758/BF03193019>

Validation of a new State Test Anxiety Scale (STAS)

Alper Sahin ^{1,*}

¹Atılım University, School of Foreign Languages, Department of Basic English, Ankara, Turkey

ARTICLE HISTORY

Received: Dec. 10, 2020

Revised: Aug. 23, 2021

Accepted: Sep. 29, 2021

Keywords:

Test Anxiety,
State Test Anxiety Scale,
Bio-psychosocial model,
STAS English.

Abstract: This study aimed to validate the English version of Sahin's (2019) State Test Anxiety Scale (STAS) which was originally developed in Turkish based on the latest bio-psychosocial model of test anxiety. For this purpose, the Turkish version of STAS was translated into English and administered to 360 (123 females, 237 males) students from 22 countries. The data were subjected to the confirmatory factor analysis to confirm the three-factor structure as in the Turkish version. Most of the fit indices examined ($\chi^2/df=1.94$, CFI=.98, NFI=.96, NNFI=.98, IFI=.98, RMSEA=.05, SRMR=.05) indicated that the data of the STAS had a good fit to the three-factor structure as its Turkish version. The Alpha internal consistency coefficients were found to be .81, .77, .91, and .92, for Physiological effects, Psychosocial effects, Cognitive effects subscales, and the total scale respectively. Stratified Alpha was also calculated and was found to be .93. All in all, evidence collected in this study indicated that the English version of STAS was a valid and reliable scale as its Turkish version.

1. INTRODUCTION

Test anxiety is one of the common types of anxiety. It affects 25% to 40% of the individuals at a certain level (McDonald, 2001). Test anxious individuals can experience some symptoms that prevent their brain and body function properly. This can decrease their test performance dramatically. Especially in a high-stakes exam situation, it is nearly impossible for some students not to get anxious before or after the exam. The numbers reveal the truth that test anxiety is very common even since the early days of schooling. According to Gençdoğan (2006), 66% of the students experience exam anxiety. Moreover, it is found in recent studies that test anxiety starts in primary school (Tornio, 2019; Popa et al., 2019) and elevates as the number of failures increases (Karataş et al., 2013). It is also said that test anxiety experienced in early childhood can be transferred to adulthood if necessary precautions are not taken (Delvecchio et al., 2017). The reason for the test anxiety can be the test-oriented nature of educational systems in different countries. For example, children in the United States take tests even before being accepted to kindergarten. Individuals may experience test anxiety in variegated levels due to the pressure of the expectations to be successful in tests starting in the early years of their life, the uncertainty about the result of the tests, and the possibility of losing face among family members.

*CONTACT: Alper ŞAHİN ✉ alpersahin2@yahoo.com 📧 Atılım University, School of Foreign Languages, Department of Basic English, Ankara, Turkey

The first studies on test anxiety started after the mid-20th century with some pioneering studies from a couple of scholars (Mandler & Sarason, 1952; Sarason & Stoops, 1978; Suinn, 1969; Osterhouse, 1970). According to Spielberger (1972) anxiety has two types: Trait-level anxiety and state-level anxiety. The former is a serious mood disorder in which individuals take their experience as a threat. The latter is a version of trait-level anxiety that triggers some problematic physiological and cognitive responses and it emerges in some specific situations like tests. Test anxiety is a kind of state-level anxiety as it has observable symptoms in only testing situations (Sapp, Farrel, & Durand, 1995). It is commonly accepted that if an individual has high trait level anxiety, they will experience high test anxiety as well (Spielberger & Vagg, 1995).

Test anxiety was first accepted as a phenomenon manifesting itself with a couple of bodily symptoms such as nausea, sweating, faster heartbeats, trembling, dizziness, dry mouth, tense muscles, headache, stomachache (Hagtvet et al., 2001). Later, it was noticed that anxiety had some deeper symptoms which could not be attributed to bodily reactions only. They were thought of as the emotional and cognitive reactions that the body produces against an uncertain situation like a test. Therefore, test anxiety was, then, considered to have two commonly accepted main sub-dimensions (Liebert & Morris, 1967). They are “worry” and “emotionality”. Worry constitutes the cognitive reactions that the body produces. They can be considered as the self-deprecative thoughts that may lead the test taker to failure (Bozkurt et al., 2017). Emotionality constitutes bodily reactions such as nausea, sweating, and faster heartbeats (Spielberger, 1972). Although this two-factor model was commonly accepted, there was another model of test anxiety taking the thoughts about social derogation of the individuals into account put forward by Friedman and Bendas-Jacob (1997) and it was then called the biopsychosocial model by Lowe et al. (2008). The Bio-psychosocial model takes social and environmental factors into account and considers the environment and the social interaction of the individuals with others as the sources of anxiety. These social factors may be observed as being teased by friends, getting negative reactions from parents, teachers and losing face with them.

There are plenty of scales developed to measure test anxiety. The test anxiety questionnaire (Mandler & Sarason, 1952) is one of the first instruments developed to measure test anxiety. It has 42 items accumulated under three subscales: “anxiety about group intelligence tests”, “anxiety about individual intelligence tests”, and “anxiety about course examinations”. Responders reflect their anxiety by pinpointing their level of endorsement to the items on a 15-centimeter line. The state-trait anxiety scale (STAI) developed by Spielberger et al. (1970) is also one of the initial scales to measure test anxiety. It has 40 items with 20 items in state and 20 others in trait anxiety subscales. Driscoll (2007) has another test anxiety scale named Westside Test Anxiety Scale. It has ten items and it takes around five minutes to administer. Another test anxiety scale is the Revised Test Anxiety Scale (RTAS) developed by Benson and El-Zahhar (1994). It has subscales called Tension, Worry, Test-irrelevant thoughts, and Bodily symptoms. Test Anxiety Inventory (TAI) which was developed by Spielberger (1980) is another test anxiety scale and it is one of the most used test anxiety scales (Chapell et al., 2005). It has 20 items under two subscales of Worry, and Emotionality.

There are two trait-level test anxiety scales recently developed. One of them is the cognitive test anxiety scale revised (Cassady & Finch, 2014). It has 25 items in one common factor related to the cognitive dimension of test anxiety. The other one is the IDA test anxiety scale (Başol, 2017) which was developed in Turkish and has nine items in the cognitive and physiological reactions subscale and six items in a subscale targeting social and environmental triggers of test anxiety.

Test anxiety has the potential to cause individuals to experience undesired drawbacks before, during, and after test-taking situations. It has some drawbacks from the point of view of

educational measurement as well. The current reliability coefficients calculate the reliability from the point of view of the items on a scale or a test based on the responses by the test takers. However, when it is test anxiety in action during a test, it is not the test items threatening the reliability of the scores. It is the test anxiety the individuals experience threatening it. That is to say, the test items may be statistically sound items and 5% or 10% of the test takers experiencing test anxiety in severe levels may not be adequate to indicate a problem in the reliability coefficients via statistical means. As their total scores are affected negatively by test anxiety, highly test anxious students are seen as low or mid-ability students who are not being able to answer the items correctly (as all other students with similar scores) without taking the anxiety they experience into consideration. As a result, students who get 60 out of 100 points without experiencing test anxiety and students who get 60 out of 100 due to the anxiety they experience are considered as equal by statistical formulas and everyone in the society. However, the students experiencing high test anxiety get fewer points than they deserve due to the negative effects of test anxiety. According to Hembree (1988) highly anxious students get around 12 percentile points lower than their peers who experience lower anxiety levels. What is worse is that it is nearly impossible to measure the exact impact of test anxiety on their test scores as they experience test anxiety in varying degrees. Therefore, the score of the students experiencing test anxiety does not reflect the real level of their latent ability on the measured construct due to test anxiety. As this hidden test anxiety effect on test scores changes based on the severity level of the anxiety individuals experience during the test, it is currently not possible to devise a correction formula to eliminate the error in the test scores of these students caused by test anxiety. As a result, it can be said that until the effects of test anxiety are lifted from the test environment, the reliability of the test scores obtained via statistical means may be much less than calculated via legitimate formulas.

The accurate and early diagnosis of test anxiety is highly critical, and such an early and accurate diagnosis may yield more reliable results if scales developed considering the latest model of test anxiety (bio-psychosocial model) are used. As studies conducted on test anxiety in the early 1950s suggest that anxiety has a connection to different situations, it would be more valid to diagnose anxiety in specific situations which is called the "Specificity Theory of Anxiety" (Harper, 1971). There are some scales developed specifically measuring state test anxiety (e.g. Mandler & Sarason, 1952; Spielberger et al., 1970). It would not be wrong to state that the youngest of them is a half-century old. The recent scales developed to measure trait level test anxiety may be limited as they are administered in the relaxed classroom environment for which the test takers should consider their overall test anxiety experience in all tests cumulatively while responding. Such scales are useful as it is practical to collect data with them. Moreover, as it is commonly accepted that there is a high correlation between state and trait test anxiety, that is, individuals experiencing high trait level anxiety will probably experience high state-level test anxiety (Spielberger & Vagg, 1995), the scale scores obtained from these scales can be used to diagnose and support the individuals experiencing test anxiety. However, the author of this paper believes that there may be some specific tests (as mentioned in the Specificity Theory of Anxiety) that the individuals with trait level anxiety may experience state-level test anxiety before, during, and after some particular tests much more than some others. Moreover, there may be some tests that they do not experience test anxiety at all or experience it to a lower stance. The current trait level test anxiety scales administered in a classroom environment away from the threat or the perception of the threat influencing the emotional state of the individuals and getting data from individuals collectively for all tests may not always distinguish between test anxiety experienced before, during, or after specific test-taking scenarios. Moreover, it would not be wrong to claim that the existing state test anxiety scales developed half a century ago may not be accurate enough to be used for such a purpose. Therefore, there may be an immediate need for a state test anxiety scale that would include the latest bio-psychosocial

model of the test anxiety, and which would enable the researchers to collect data for each specific test situation individually. For this purpose, the Turkish version of the State Test Anxiety Scale (Sahin, 2019) was developed. However, the need for such a scale in the English language to close the aforementioned gap in the literature still exists. Its translation to English and validation study was conducted to provide the international literature with the English version of this up-to-date state test anxiety scale. Therefore, the purpose of this study was to validate the English version of the State Test Anxiety Scale (STAS) which was developed originally in Turkish by Sahin (2019).

2. METHOD

In this part of the paper, the data collection instruments, the participants, the data collection procedures, and the statistical analyses used to collect data for the validity, reliability, and linguistic equivalence of the scale with its Turkish version will be detailed.

2.1. Data Collection Instruments

This study was conducted at the Northern Cyprus Campus of a reputable Turkish university with the participation of the students taking an academic English course during their freshmen year. A couple of data collection tools were used due to the nature of this study. They are the Turkish version of the *State Test Anxiety Scale* (STAS-TR; Sahin, 2019), the English version of the *State Test Anxiety Scale* (STAS-EN), and the Turkish version of the *Revised Test Anxiety Scale* (RTAS; Akin et al., 2012).

2.1.1. STAS-TR

The Turkish version of STAS was developed by Sahin (2019) using 312 participants (129 Females, 183 males) who were students at the Preparatory School Program (PSP) of an English medium private university located in the Turkish Republic of Northern Cyprus. Students attend PSP to develop their English language skills so that they can easily follow their departmental courses which will be held in English. STAS-TR has 22 items accumulated under three subscales called Physiological Effects (PE), Psychosocial Effects (PSE), and Cognitive Effects (CE). Cronbach's Alpha internal consistency coefficient for each subscale was found to be .85, .84, .93 respectively, and .94 for the total scale (Sahin, 2019). The participants respond to STAS-TR through a 4-point Likert scale with the Turkish versions of "1-Not At all, 2-Slightly, 3-Moderately, 4-A lot". The PE subscale of STAS-TR has eight items covering physiological effects of test anxiety such as headache, tighter muscles, difficulty breathing, fever, nausea, and stomachache. Two of the items under this subscale are "I have difficulty breathing" and "I have a dry mouth".

The PSE subscale has five items covering the effects of test anxiety pushing the individuals to think over the reactions of the people in their social environment to their possible incompetence. This subscale has items about negative reactions from the family members, feeling of losing face with the teacher, fear of being teased by the classmates in case of failure. One of the items in this subscale is "I worry about my classmates making fun of me for getting a low grade on the exam".

The CE subscale has nine items like the fear of not being able to come up with the right answers, not being able to understand what is asked in the exam, getting a low grade, failure, not being able to finish the exam on time. One of the items under this subscale is "I am afraid of not being able to finish the exam on time".

As the purpose of this study was to validate the English version of STAS-TR, STAS-TR was used to compare the reliability, validity, and linguistic equivalence of STAS-EN in this study.

2.1.2. STAS-EN

In order to ensure the linguistic equivalence of STAS-EN and STAS-TR, STAS-TR, which was found to be a valid and reliable state test anxiety scale in its original study (Sahin, 2019), was given to two bilingual translators who were both highly competent in Turkish and English and it was translated to English by them. When the translations were finalized, the translations of each item were reviewed and the most appropriate translations for each item were selected by the author who is also bilingual and who has B.A. and M.A. degrees in English language teaching. After this stage, the initial form of STAS-EN was given to two separate competent bilingual translators and was translated back to Turkish. The back translations were reviewed by the author and amendments in the items of STAS-EN were done based on the back translations. Then, the draft version of the STAS-EN was obtained. Two separate bilingual language experts with backgrounds in linguistic studies reviewed the draft version of STAS-EN. Some statements were amended at this stage based on their feedback as well.

At the final stage, the draft version of STAS-EN was reviewed by a British, a Canadian, and two American native speakers one of which was a scholar engaged in anxiety research and two of which were psychologists and psychometricians at the same time. STAS-EN was edited one more time based on the native speakers' suggestions. Finally, STAS-EN was administered to a group of students with diverse nationality and English proficiency levels in order to get feedback about the clarity and comprehensibility of the items in the scale. The draft version was edited one last time based on the student feedback and questions.

2.1.3. RTAS

RTAS, developed by Benson and El-Zahhar (1994) and adapted to Turkish by Akin et al. (2012), was used to collect evidence on the concurrent validity of the STAS-EN as was done while STAS-TR was being developed. The first reason why RTAS was selected as the scale to be used for concurrent validity was that it had many subscales with a close relationship with the subscales of STAS-TR and STAS-EN. Moreover, it was a widely used scale with separate validity studies done by multiple samples (Egyptian, American, Spanish) and authors (Bados & Sanz, 2005; Benson & Al-Zahhar, 1994; Hagtvet & Benson, 1995). More importantly, RTAS (Turkish version) was used in the development of STAS-TR and it was thought that using it while developing STAS-EN would yield comparable results with STAS-TR.

RTAS is a trait level test anxiety scale that has 20 items under four subscales entitled "Tension", "Bodily Symptoms (BS)", "Worry", and "Test-irrelevant Thoughts (TiT)" with a 4-point Likert scale. The Tension subscale has five items covering anxious thoughts of the individuals about the test like "I worry a great deal before taking an important exam", BS subscale has five items on the bodily reactions such as "I get a headache during an important test", Worry subscale has six items covering the anxiety about failure and self-deprecating thoughts like "During tests, I find myself thinking about the consequences of failing" and TiT subscale has 4 items about test-irrelevant thoughts like "While taking tests, I sometimes think about being somewhere else".

Benson and El-Zahhar (1994) reported alpha coefficients ranging between .76 and .91 with the American sample (as cited in Akin et al. (2012), p.107). Hagtvet and Benson (1995) found that the alpha reliability coefficients were .89, .81, .81, .85, and .91 for the Tension, Worry, BS, TiT, and the whole scale respectively. Akin et al. (2012) also reported .78, .71, .78, .80, and .88 for the Tension, Worry, BS, TiT, and the whole scale.

2.2. The Participants, Data Collection Procedures, and Statistical Analyses

STAS-EN was administered to students a couple of times just before different exams to collect adequate data as evidence of validity, reliability, and indirect evidence of linguistic equivalence of the scale. Apart from STAS-EN, STAS-TR, and RTAS (Turkish version) were also

administered to participants at different stages of the study. The data collection started with the administration of RTAS to 196 students (77 Females, 119 Males) in the classroom environment and an informed consent form detailing the purpose of the study, how the data collected will be used and how participants can leave the study anytime they wish was signed by the voluntary participants. STAS-EN was administered to these students right before their midterm exam for an academic English course and the Pearson product-moment correlation coefficient between the same 196 students' scale and subscale scores obtained from both RTAS and STAS-EN were calculated using SPSS 23 (IBM Corp, 2015) as evidence of concurrent validity of STAS-EN.

The main group of participants of the study (n=360) was selected from a population of students with which STAS-TR was developed. STAS-TR was developed with the participation of only students with Turkish origins due to the language of the scale. The students with Turkish origins constituted a large part of this study as well (282 including the 20 students who were citizens of the Turkish Republic of Northern Cyprus). Moreover, 78 international students from 21 different countries such as Pakistan, Bangladesh, Egypt, Jordan, Kenya, Syria, Kazakhstan, Azerbaijan, Ruanda, Nigeria, and Uganda also participated in the study reaching a total number of 360 students (123 females, 237 Males) from 22 countries. 257 (56 Females, 201 Males) of the 360 students were from Engineering programs and 103 (67 Females, 36 Males) of them were students registered to Social Sciences programs. The more detailed descriptive statistics about the gender and fields of the participants can be found in [Table 1](#).

Data collected from the main participants of the study was also used to confirm the three-factor structure of STAS-EN that it inherited from STAS-TR. That is, the factorial structure of STAS-EN was already known and it was expected to maintain the three-factor structure of STAS-TR. Therefore, confirmatory factor analysis (CFA) was utilized to confirm this structure. Before the CFA, to detect multivariate outliers, Mahalanobis Distance (D^2) was calculated using SPSS 23 (IBM Corp, 2015) and their p values were examined. Moreover, the Q-Q plots of the Mahalanobis Distances were examined. In this examination, no multivariate outliers were detected in the data. Following this examination, data was taken into confirmatory factor analysis (CFA) based on the default options of LISREL 8.51 (Jöreskog & Sörbom, 2001) in order to confirm and analyze whether the three-factor model of STAS-TR fit the data collected using STAS-EN.

Table 1. *Descriptive statistics of the main participants of the study (n=360).*

Department		Female	Male	Frequency	Percent
Social Sciences Programs	Economy	6	5	11	3.1
	Guidance and Psychological Counseling	14	1	15	4.2
	Business Administration	5	14	19	5.3
	Political Science and International Relations	6	11	17	4.7
	Psychology	36	5	41	11.4
Social Sciences Total		67	36	103	28.7
Engineering Programs	Aerospace Engineering	16	27	43	11.9
	Electrical and Electronics Engineering	7	38	45	12.5
	Chemical Engineering	7	6	13	3.6
	Mechanical Engineering	4	47	51	14.2
	Petroleum and Natural Gas Engineering	6	8	14	3.9
	Computer Engineering	9	47	56	15.6
	Civil Engineering	7	28	35	9.7
Engineering Total		56	201	257	71.3
Grand Total		123	237	360	100.0

Evidence about the reliability and internal consistency of STAS-EN was assessed utilizing different analyses. For this purpose, Cronbach's Alpha, stratified Alpha, McDonald's Omega, the Spearman-Brown split-half reliability coefficients were calculated. Apart from these, test-retest reliability analysis was also conducted. While calculating the reliability coefficients, data collected from the main participants ($n=360$) of the study were used. For test-retest reliability analysis, a separate group of 98 students (32 Females, 66 Males) than the main participants of the study enrolled in an academic English course at the same institution where the study was held were asked to respond to STAS-EN before a quiz and a midterm exam which had around 4-week interval in between. The data for test-retest reliability analysis was collected from a different group of students because it was not possible to find an adequate number of exams within one single course to collect the necessary evidence to properly assess psychometric properties of the STAS-EN from a single group of students. The Pearson-product moment correlation between scale and subscale scores obtained from both administrations of the STAS-EN were used as evidence of test-retest reliability. Moreover, item-by-item correlation coefficients between the two administrations of STAS-EN were also calculated.

The test-retest reliability analysis was done one more time to evaluate whether there was a change in correlations when two language versions, STAS-EN and STAS-TR, were administered to a common group of bilingual Turkish students who were competent in both languages. For this purpose, first, STAS-EN was administered to a group of students ($n=90$), and then STAS-TR was administered to the same group of students before another exam three weeks later. Then, the correlations between the scale and subscale scores obtained from these administrations were calculated. This analysis was taken as an indirect proof of linguistic equivalence.

3. RESULTS

The results of the study will be presented here under three headings: Results as evidence of the validity of STAS-EN, Results as evidence of the reliability of STAS-EN, and Results as evidence of linguistic equivalence of STAS-TR and STAS-EN.

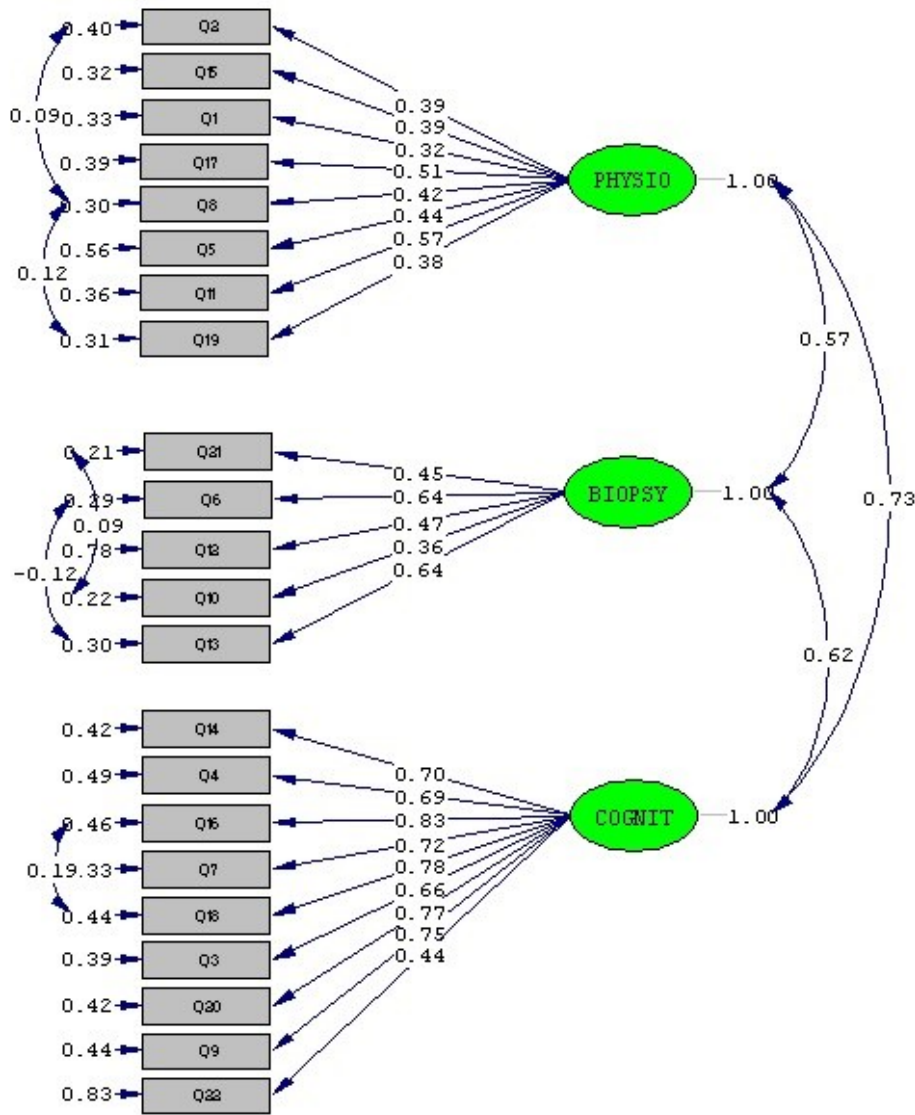
3.1. Results as Evidence of the Validity of STAS-EN

The first validity evidence collected for STAS-EN was out of the analysis conducted to evaluate its construct validity. STAS-TR had a three-factor structure. In order to confirm this three-factor structure in STAS-EN and the structural equivalence of STAS-EN with STAS-TR, the data collected was used to conduct a CFA. The path diagram obtained as an output of CFA can be found in [Figure 1](#) below.

The suggested modifications were reviewed and added to the model if it was thought that the correlation between their residuals could be explained and if they improved the fit indices (Seçer, 2015). When reviewed, it was found that Q2, Q8, and Q19 were all about the stomach-related problems a test-taker experiences. Moreover, Q10 and Q21 were both items related to classmates seeing the individual's exam score, and Q16 and Q18 were both items related to anxiety experienced due to having low scores from the test. Moreover, it was observed that all these modifications contributed to the model fit indices. As a result, they were decided to be added to the model. Moreover, the model fit indices obtained out of CFA can be found in [Table 2](#).

As can be seen in [Table 2](#), out of nine fit indices calculated, STAS-TR had six of them indicating a good fit to the model; however, according to the data collected with STAS-EN, seven out of nine fit indices indicated a good fit of the data to the model. This was considered as the evidence of the construct validity of the STAS-EN. It is also important to note that STAS-EN got consistently better fit index values in many fit indices (e.g. CFI, AGFI, IFI, NFI, NNFI).

Figure 1. The path diagram of the CFA conducted over the research data.



Chi-Square=390.81, df=201, P-value=0.00000, RMSEA=0.051

Table 2. Fit values obtained from CFA.

Fit Index	Good Fit Interval*	Adequate fit Interval*	STAS-TR / EN	Degree
χ^2/df	$0 \leq \chi^2/df \leq 2$	$2 \leq \chi^2/df \leq 3$	1.72 / 1.94	Good fit
CFI	$.95 \leq CFI \leq 1.00$	$.90 \leq CFI \leq .95$.96 / .98	Good fit
SRMR	$.00 \leq SRMR \leq .05$	$.05 \leq SRMR \leq .10$.05 / .05	Good fit
GFI	$.95 \leq GFI \leq 1.00$	$.90 \leq GFI \leq .95$.91 / .91	Adequate fit
AGFI	$.90 \leq AGFI \leq 1.00$	$.85 \leq AGFI \leq .90$.88 / .89	Adequate fit
IFI	$.95 \leq IFI \leq 1.00$	$.90 \leq IFI \leq .95$.96 / .98	Good fit
NFI	$.95 \leq NFI \leq 1.00$	$.90 \leq NFI \leq .95$.92 / .96	Good fit
RMSEA	$.00 \leq RMSEA \leq .05$	$.05 \leq RMSEA \leq .08$.05 / .05	Good fit
NNFI	$.95 \leq NNFI \leq 1.00$	$.90 \leq NNFI \leq .95$.96 / .98	Good fit

*Çokluk et al. (2010), Kahraman et al. (2018), Schermelleh-Engel et al. (2003).

In order to compare the performance of STAS-EN with STAS-TR, Pearson product-moment correlation coefficients between scale and subscale scores of STAS-EN were also calculated. You can find these coefficients in [Table 3](#) with the same correlations obtained from STAS-TR in the original study (Sahin, 2019).

Table 3. *The correlations between the scale and subscale scores of STAS-TR and STAS-EN.*

Subscales	PSE TR / EN	PE TR / EN	Total TR / EN
CE-TR/EN	.65** / .66**	.65** / .59**	.94** / .92**
PSE-TR/EN		.49** / .57**	.78** / .81**
PE – TR / EN			.82** / .80**

** $p < 0.001$

As can be seen in [Table 3](#), the CE subscale of STAS-EN had correlations of .66 with PSE subscale scores, .59 with PE subscale scores, and .92 with total scale scores of STAS-EN. These correlations were .65, .65, and .94 in the original study in which STAS-TR was developed. Moreover, the PSE subscale of STAS-EN had .57 correlation with PE subscale and .81 with the total scale score of STAS-EN. These correlations were .49 and .78 in the original study in which STAS-TR was developed. Lastly, the PE subscale of the STAS-EN had .80 correlation with the total scale score of STAS-EN which was .82 in the original study with STAS-TR. These can be taken as the validity evidence of the STAS-EN and its subscales.

As mentioned earlier, to evaluate the concurrent validity of STAS-EN, and RTAS (Turkish version), another valid and reliable scale on test anxiety which was validated with multinational samples was also administered to a group of students (n=196) in the classroom environment. Then, the scale and subscale scores of RTAS and STAS-EN were correlated. These correlations can be viewed in [Table 4](#). [Table 4](#) also indicates the correlations obtained between STAS-TR and RTAS in the original study (Sahin, 2019).

Table 4. *Correlations of STAS-TR and STAS-EN scale and subscale scores with RTAS.*

	Tension	BS	Worry	TiT	RTAS Total
PE-TR/EN	.61** / .56**	.69** / .63**	.52** / .45**	.25** / .18**	.66** / .61**
PSE-TR/EN	.43** / .52**	.39** / .43**	.49** / .53**	.24** / .30**	.50** / .61**
CE- TR/EN	.73** / .70**	.47** / .49**	.63** / .65**	.22** / .19**	.67** / .70**
STAS Total-TR/EN	.73** / .73**	.60** / .61**	.73** / .66**	.27** / .25**	.73** / .76**

** $p < 0.001$

When [Table 4](#) is reviewed, it can be seen that STAS-TR and STAS-EN scale scores had very similar correlations with the subscales of RTAS. Another important finding to point out is that high correlation coefficients between STAS-EN scale and subscale scores and RTAS scale and subscale scores were obtained except the PE subscale giving hints of discriminant validity of STAS-EN. It is also important to emphasize that this was also the case in the original study of STAS-TR as can be seen in [Table 4](#).

When the correlations between the subscales were analyzed, it can be seen that the highest correlation was between the PE subscale of STAS and the BS subscale of RTAS in both versions of STAS (.69 vs .63). This was expected as the corresponding scale which contains the physiological effects of test anxiety was the BS subscale of RTAS. It is also interesting to

state here is that there was around a .05-point decrease between the PE in STAS-TR and STAS-EN with some subscales of RTAS. For example, correlation with the Tension subscale decreased from .61 to .56 and the correlations with the BS subscale of RTAS decreased from .69 to .63. This may be due to the nature of different tests triggering different bodily reactions before each at this administration of the scale.

When the correlations of the PSE subscale of STAS with subscales of RTAS are analyzed, it can be noticed that in both versions of STAS, the highest correlations were obtained with the Worry and Tension subscales. It can be considered as normal as the source of the feelings regarding the pressure from family and friends are related to the items under Tension and Worry subscales of RTAS. Apart from this, it can also be seen that correlations with the Tension subscale of RTAS with STAS-TR and STAS-EN were quite similar.

The correlations with the CE subscale of STAS (both TR and EN) and the subscales of RTAS indicated that both versions of the scale had higher correlations with the Tension subscale of RTAS. This was also expected as they had similar items covering anxious thoughts regarding the test in both scales such as “I feel nervous” and “I am afraid of not being able to finish the exam on time” in STAS-EN’s CE subscale and “I start feeling very uneasy just before getting a test paper back” and “I worry before the test because I do not know what to expect” from the Tension subscale of RTAS.

It is important to note that all subscales of STAS-TR and STAS-EN had very low correlations with TiT of RTAS. This was also an expected outcome as the TiT subscale of RTAS had items covering another dimension of test anxiety which was not covered in both STAS-TR and STAS-EN. This can be taken as evidence of the discriminant validity of STAS-TR and STAS-EN.

3.2. Results as Evidence of the Reliability of STAS-EN

Reliability evidence for STAS-EN was collected calculating Cronbach’s alpha reliability coefficient, Stratified Alpha coefficient, McDonald’s Omega, Spearman-Brown Split-half reliability coefficient, and test-retest reliability analysis (four-week interval in between). The results of these calculations can be seen in Table 5. Moreover, the reliability-related findings of STAS-TR in the original study (Sahin, 2019) were also reported in Table 5 to compare both versions of STAS.

Table 5. Reliability evidence collected for STAS-EN.

	Stratified Alpha	McDonald’s Omega (ω)	C’s Alpha (STAS-TR/EN)	Spearman-Brown Split Half (STAS-TR/EN)	Test-Retest STAS-TR (n=108) / STAS-EN (n=98)
PE		.81	.85 / .81	.89 / .80	.74** / .76**
PSE		.80	.84 / .77	.86 / .81	.80** / .76**
CE		.91	.93 / .91	.94 / .89	.78** / .76**
Total	.93	.92	.94 / .92	.96 / .91	.81** / .78**

** $p < 0.001$

As can be seen in Table 5, it can be said that Stratified Alpha and McDonald’s Omega (ω) indicated high reliability. Moreover, it can also be said that the Cronbach’s alpha reliability coefficients of STAS-EN were consistently lower compared to STAS-TR. However, they still indicated consistently high reliability. A similar decrease is also evident in the reliability coefficients obtained from Spearman-Brown’s Split-half reliability. However, the Split-half reliability coefficients obtained still indicate high stability between the scores from both halves as well.

The negative shift observed in alpha reliability coefficients of STAS-EN compared to STAS-TR can be attributed to the multinational sample which was used to collect the data for STAS-EN. As mentioned earlier, the sample used in the development of STAS-TR was composed of only Turkish nationals; however, the sample used in this study was of a multinational one. Therefore, the highly diverse test-taker sample may have yielded some inconsistent response patterns in the data. However, STAS-EN's capability to accommodate high cultural diversity in return for a slight decrease in reliability coefficients should be praised and could even be taken as an indicator of its being a highly reliable and stable scale of test anxiety.

Another reliability-related evidence collected was the test-retest reliability analysis of STAS-EN. As mentioned earlier, for this purpose, STAS-EN was administered to a sample of 98 (32 Females, 66 males) students before a quiz and a midterm exam in a 4-week interval. As can be seen in Table 6, the correlations between these two administrations of the STAS-EN to a common group of test-takers before different tests indicated highly consistent results. Calculating test-retest reliability between these two tests was risky as one was a quiz that constituted 10% of the total score of the students for the course and the other one was a midterm exam that was more seriously taken by most students and constituted 20% of the total scores of the students. Although there was such a drawback of evaluating the stability of STAS-EN between a quiz and a midterm, it was thought that it still yielded highly stable results by producing correlations ranging between .76 and .78 compared to STAS-TR test-retest correlations ranging between .74 and .81 which were obtained from two administrations of STAS-TR before two midterm exams with a 4-week interval.

An item-by-item correlational analysis was done to see how consistent the test taker responses were in two consecutive administrations of STAS-EN to a common group of students. The findings of this analysis can be viewed in Table 6.

Table 6. *Item-by-Item Correlations Test-Retest Reliability for STAS-EN (n=98).*

Scale	Item #	r	Scale	Item #	r	Scale	Item #	r
PE	1	.43	CE	3	.57	PSE	6	.51
	2	.30		4	.65		10	.43
	5	.47		7	.61		12	.64
	8	.70		9	.43		13	.56
	11	.66		14	.48		21	.57
	15	.28		16	.55			
	17	.31		18	.55			
	19	.74		20	.57			
				22	.42			
Ave. :		.49	Ave. :		.54	Ave. :		.54

When Table 6 is analyzed, it can be seen that there are some inconsistent items with correlations less than .40 especially in the PE subscale of STAS-EN. These items were items 2, 15, and 17. However, when these items were reviewed, it was noticed that they were about some bodily symptoms (2. I have nausea, 15. It feels like I have a fever, 17. I have a headache.) of test anxiety that may not manifest before or after each exam. Therefore, such fluctuations of correlations were expected because the symptoms experienced may alter in variety and magnitude before, or after each exam depending on the test type, the test environment, and some other factors. This item-by-item correlational analysis should be analyzed considering this variability in symptoms. However, the findings in Table 6 may be taken as an indication of high consistency for the items of STAS-EN despite this drawback.

3.3. Results as Evidence of Linguistic Equivalence of STAS-TR and STAS-EN

The findings as evidence of reliability and validity already had implications of the linguistic equivalence of both versions of STAS; however, apart from the direct evidence of forward and backward translations used during the development of STAS-EN, some indirect evidence regarding linguistic equivalence was also collected to evaluate whether the linguistic equivalence of STAS-EN and STAS-TR was maintained.

A correlational analysis was conducted between the scale and subscale scores of data obtained from the STAS-EN and STAS-TR. First, STAS-EN was administered to a new group of bilingual students ($n=90$) who were both competent in English and Turkish before a midterm exam, and then STAS-TR was administered to the same group of students before a quiz three weeks later. A comparison of correlation coefficients obtained from administration of STAS-EN to the same group of individuals ($n=98$) two times (r^{EE}) while conducting test-retest reliability analysis (last column in Table 5) and consecutive administration of STAS-EN and STAS-TR (r^{ET}) to a common group of students ($n=90$ -only Turkish ones) who are competent Turkish and English speakers are presented in Table 7. Please note that these correlations are inter-scale correlations of the same subscales in these two administrations of the scales to a common group of individuals.

Table 7. The inter-scale correlations of the subscales of STAS-EN and two administrations of STAS-EN and STAS-TR to common groups.

	STAS PE (r^{EE} / r^{TE})	STAS CE (r^{EE} / r^{TE})	STAS PSE (r^{EE} / r^{TE})	STAS-EN Total (r^{EE} / r^{TE})
STAS PE	.76 / .71			
STAS CE		.76 / .85		
STAS PSE			.76 / .78	
STAS-EN Total				.78 / .85

As can be seen in Table 7, when STAS-EN and STAS-TR were administered to a common group of students, the correlations between the scale and subscale scores of the common students for these two administrations ranged between .71 and .85. More importantly, except for the one obtained from STAS-EN PE, all other correlations obtained from scale and subscale scores of STAS-EN and STAS-TR surpassed the ones obtained from the consecutive administration of STAS-EN two times. This can be taken as an indication of the high stability of the scores between the English and Turkish versions of STAS and taken as indirect evidence of linguistic equivalence of STAS-TR and STAS-EN.

The item-by-item correlational analysis was iterated to be able to evaluate whether the items in STAS-TR and STAS-EN functioned equally well or not after their administration of these two versions to a common group of students in a three-week interval. Another comparative table, Table 8, was prepared to present the findings of item-by-item correlations obtained from test-retest reliability analysis of STAS-EN (Table 6) and the same analysis done using student responses after the consecutive administration of STAS-EN and STAS-TR to a common group of students.

When Table 8 is analyzed, it can be seen that the item-by-item correlations between the two versions of the STAS range between .35 and .74. The lowest correlation obtained was from item 1 (I have difficulty breathing.) while the highest correlation was obtained from Item 21 (I worry about my classmates seeing my exam score). Moreover, it can also be seen that the correlations between STAS-EN and STAS-TR were quite like the ones obtained after the administration of STAS-EN two times during the test re-test reliability analysis. In most cases,

the correlations obtained from scale and subscale scores of STAS-TR and STAS-EN were higher than the ones obtained while conducting test-retest reliability analysis for STAS-EN. This was also taken as the indirect evidence of linguistic equivalence of STAS-EN and STAS-TR.

When all findings are put together, it would not be wrong to state that the STAS-TR and STAS-EN can be considered as linguistically equivalent forms of the same construct in different languages.

Table 8. Comparison of Item-by-Item correlations.

Scale	Item #	r^{EE}	r^{TE}	Scale	Item #	r^{EE}	r^{TE}	Scale	Item #	r^{EE}	r^{TE}
PE	1	.43	.35	CE	3	.57	.65	PSE	6	.51	.69
	2	.30	.48		4	.65	.59		10	.43	.71
	5	.47	.49		7	.61	.65		12	.64	.66
	8	.70	.40		9	.43	.48		13	.56	.54
	11	.66	.67		14	.48	.55		21	.57	.74
	15	.28	.55		16	.55	.68				
	17	.31	.36		18	.55	.58				
	19	.74	.48		20	.57	.61				
				22	.42	.49					
Ave. :		.49	.47	Ave. :		.54	.59	Ave. :		.54	.67

3.4. Calculation of the Scale and Subscale Scores of STAS-EN

The scale and subscale scores of the STAS-EN can be calculated by simply adding the responses given to each item in each subscale. The items numbered 1, 2, 5, 8, 11, 15, 17, and 19 can be collected to obtain the subscale score for PE. The lowest score that can be obtained from this subscale is 8 and the highest score is 32. The score obtained from adding up the responses given to items numbered 3, 4, 7, 9, 14, 16, 18, 20, 22 will yield the CE subscale score. The lowest score that can be obtained from this subscale is 9 and the highest score is 36. The score obtained from items numbered 6, 10, 12, 13, 21 yields the subscale score for PSE. The lowest score for this subscale is 5 and the highest score is 20. Apart from all these subscale scores, all item scores can be collected to obtain the scale score. 22 is the lowest score and 88 is the highest score that can be obtained from the scale.

When the scale or subscale scores reach 1/2 of the highest score (e.g. 18/36 for CE) that can be obtained from that subscale, it may be considered that the subscale score is at its medium level. When it reaches 2/3 (e.g. 24/36 for CE) of the total scale score, the score may be considered as high. However, these figures should be approached with caution as there were no validation studies done to confirm them.

4. DISCUSSION & CONCLUSION

The need for a scale that can be used to measure and pinpoint the test anxiety levels of the individuals right before each exam triggered the development of STAS-TR (Sahin, 2019). The existing need for such an up-to-date scale in English to close the gap in the international literature constituted the motivation of this study. Therefore, the purpose of this study was to validate the English version of STAS to make it available to international researchers. For this reason, multiple pieces of evidence from a multinational group of individuals were collected. First, STAS-EN was administered to a multinational sample consisted of 360 freshman students who constituted the main participants of the study. Evidence of validity was collected at this stage. For the construct validity of the scale, data collected from the main sample of 360

students were subjected to a CFA in order to check whether STAS-EN had the same factorial structure with STAS-TR. Most of the fit indices for CFA indicated good fit ($\chi^2/sd=1.94$, CFI=.98, NFI=.96, NNFI=.98, IFI=.98, RMSEA=.05, SRMR=.05) of the data to the three-factor model. The findings confirmed that STAS-EN had a three-factor structure as was the case in STAS-TR. Apart from construct validity, more data was collected to compare the performance of the STAS-EN to that of another valid scale, RTAS, as evidence of concurrent validity. The findings of this analysis also confirmed that STAS-EN was a valid scale as RTAS to measure test anxiety.

The evidence regarding the reliability of STAS-EN was also collected through Cronbach's Alpha, Stratified Alpha, McDonald's Omega, test-retest (n=98), and Spearman Brown's split-half formula. Data obtained from these analyses ($\alpha^{PE}=.81$, $\alpha^{PSE}=.77$, $\alpha^{CE}=.91$, $\alpha^{Total}=.92$) supported each other and confirmed that STAS-EN was a highly reliable and stable scale.

For the indirect evidence of linguistic equivalence, STAS-TR and STAS-EN were administered to a new common group of students (n=90). The correlations between the scale and subscale scores indicated that STAS-EN had similar linguistic properties. All in all, the evidence collected for reliability, validity, and linguistic equivalence of STAS-TR and STAS-EN indicated that STAS-EN was both valid and reliable equivalent of STAS-TR.

With the STAS-EN available in the literature and as it is a scale that can be administered before each exam separately, it will be possible for the international researchers to gain deeper insights into whether some factors such as the type of the course, the instructor, or even the proctor can leverage or soothe the test anxiety levels of individuals. Moreover, it will also be possible for the researchers to do research on some exam room arrangements like the use of soothing music, meditation, or different lighting before the exam and test their effects on the anxiety levels of test anxious individuals. Such research is missing in the literature and the researchers are kindly invited to research such topics. Moreover, international researchers are also invited to design studies to identify the cut scores of the STAS-EN scale and subscale scores and validate STAS-EN in different cultures.

Acknowledgments

The author of this study thanks to the reviewers wholeheartedly for their meticulous work and contribution to the manuscript.

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). The ethical committee approval dated 06.05.2019 numbered BAYEK 2019 / 04-06 was obtained from the Scientific Research and Publication Ethics Committee of the Middle East Technical University Northern Cyprus Campus.

ORCID

Alper Şahin  <https://orcid.org/0000-0001-7750-4408>

5. REFERENCES

- Aydın, U., Bulgan, G. (2017). Adaptation of children's test anxiety scale to Turkish. *Elementary Education Online*, 16(2), 887-899. <https://doi.org/10.17051/ilkonline.2017.304742>
- Akın, A., Demirci, İ., & Arslan, S. (2012). Revize edilmiş sınav kaygısı ölçeği: Geçerlik ve güvenilirlik çalışması [Revised Test Anxiety Scale: Validity & Reliability Study]. *Educational Sciences and Practice*, 11(21), 103-118.
- Bados, A. & Sanz, P. (2005). Validation of the revised test anxiety scale and the Friedben test anxiety scale in a Spanish sample. *Ansiedad y Estrés*, 11 (2/3), 163-174.

- Başol, G. (2017). IDA test anxiety scale: Validity and reliability study. *The Journal International Education Science*, 4(13), 173-193. <https://doi.org/10.16991/INESJOURNAL.1506>
- 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*, (7)1, 1-9. <https://doi.org/10.1177/2158244016669549>
- Cassady, J. C., & Finch, W. H. (2014). Confirming the factor structure of the Cognitive Test Anxiety Scale: Comparing the utility of three solutions. *Educational Assessment Journal*, 19(3), 229-242. <https://doi.org/10.1080/10627197.2014.934604>
- Chapell, M. S., Blanding, B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology*, 97(2), 268-274. <https://doi.org/10.1037/0022-0663.97.2.268>
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010). *Sosyal Bilimler için Çok değişkenli istatistik : SPSS ve LISREL uygulamaları [Multivariate Statistics for Social Sciences: SPSS and LISREL Applications]*. Pegem Academy Press.
- Delvecchio, E., Cavallina, C., Di Riso, D., & Mazzeschi, C. (2017). Early evidence of Italian validation of the trait anxiety scale of the state trait anxiety inventory for children. *European Journal of Developmental Psychology*, 15(2), 214-223. <https://doi.org/10.1080/17405629.2017.1297227>
- Driscoll, R. (2007). *Westside test anxiety scale validation*. Education Research Information Center. <https://files.eric.ed.gov/fulltext/ED495968.pdf>
- Friedman, I. A., & Bendas-Jacob, O. (1997). Measuring perceived test anxiety in adolescents: A self-report scale. *Educational and Psychological Measurement*, 57(6), 1035-1046. <https://doi.org/10.1177/0013164497057006012>
- Gençdoğan, B. (2006). Lise öğrencilerinin sınav kaygısı ile boyuneğicilik düzeyleri ve sosyal destek algısı arasındaki ilişkiler [The relationship between the test anxiety, submissive behavior level and perception of social support of high school students]. *Atatürk University Journal of Social Sciences Institute*, 7(1), 153-164.
- Hagtvet, K. A., & Benson, J. (1997). The motive to avoid failure and test anxiety responses: Empirical support for integration of two research traditions. *Anxiety, Stress and Coping*, 10(1), 35-57. <https://doi.org/10.1080/10615809708249294>
- Hagtvet, K.A., Man, F., & Sharma, S. (2001). Generalizability of self-related cognitions in test anxiety. *Personality and Individual Differences*, 31(7), 1147-1171. [https://dx.doi.org/10.1016/S0191-8869\(00\)00212-9](https://dx.doi.org/10.1016/S0191-8869(00)00212-9)
- Harper, F. (1971). Specific Anxiety Theory and the Mandler-Sarason Test Anxiety Questionnaire. *Educational and Psychological Measurement*, 31(4), 1011-1014. <https://doi.org/10.1177/001316447103100431>
- Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research, 58 (1), 47-77. <https://doi.org/10.3102/00346543058001047>
- IBM Corp. (2015). *IBM SPSS Statistics for Windows, Version 23.0*. IBM Corp.
- Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8.5 for Windows* [Computer software]. Scientific Software International, Inc.
- Kahraman, S., Özbaşı, D., & Özdemir, M. (2018). A Study on the Development of an Attitude Scale Towards the Use of PowerPoint in Classroom. *Kastamonu Education Journal*, 26(4), 1237-1246. <https://doi.org/10.24106/kefdergi.434168>

- Karataş, H., Alici, B., & Aydın, H. (2013). Correlation among high school senior students' test anxiety, academic performance and points of university entrance exam. *Educational Research and Reviews*, 8(13), 919-926. <https://doi.org/10.5897/ERR2013.1462>
- Liebert, R. M., & Morris, L.W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, 20(3), 975-978. <https://doi.org/10.2466/pr0.1967.20.3.975>
- Lowe, P. A., Lee, S. W., Witteborg, K. M., Prichard, K. W., Luhr, M. E., Cullinan, C. M. ... Janik, M. (2008). The test anxiety inventory for children and adolescents (TAICA): Examination of the psychometric properties of a new multidimensional measure of test anxiety among elementary and secondary school students. *Journal of Psychoeducational Measurement*, 26(3), 215-230.
- Mandler, G., & Sarason, S. B. (1952). Some correlates of test anxiety. *Journal of Abnormal and Social Psychology*, 47(4), 810-817. <https://doi.org/10.1037/h0060009>
- 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>
- Osterhouse, R. A. (1970). Desensitization and study skills as treatment for two types of test-anxious students. *Journal of Counseling Psychology*, 19(4), 301-307. <https://doi.org/10.1037/h0034177>
- Popa, C., Bochis, L., & Clipa, O. (2019, July 3-5). *School assessment and test anxiety at primary school pupils* [Paper Presentation]. 4th International Conference on Lifelong Education and Leadership for All, Wroclaw, Poland.
- Sahin, A. (2019). State Test Anxiety Scale (STAS): Validity and Reliability Study. *Trakya Journal of Education*, 9(1), 78-90. <https://doi.org/10.24315/tred.450423>
- Sapp, M., Farrel, W., & Durand, H. (1995). The effects of mathematics, reading and writing tests in producing worry and emotionality test anxiety with economically and educationally disadvantaged college students. *College Students Journal*, 29(1), 122- 125.
- Sarason, I. G., & Stoops, R. (1978). Test anxiety and the passage of time. *Journal of Consulting and Clinical Psychology*, 46(1), 102- 109. <https://doi.org/10.1037//0022-006x.46.1.102>
- 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.
- Seçer, İ. (2015). Process of psychological test development and adaptation: SPSS and LISREL applications. Anı Publishing.
- Spielberger, C. D. (1972). *Theory and research in anxiety*. Academic Press.
- Spielberger, C. D. (1980). *Test anxiety inventory: Preliminary professional manual*. Consulting Psychologists Press.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for state-trait anxiety inventory*. Consulting Psychologists Press.
- Spielberger, C. D., & Vagg, P. R. (1995). Test anxiety: A transactional process model. In C. D. Spielberger & P. R. Vagg (Eds.), *Test anxiety: Theory, assessment, and treatment* (pp. 3-14). Taylor & Francis.
- Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Routledge/Taylor & Francis Group.
- Suinn, R. M. (1969). The STABS, a measure of test anxiety for behavior therapy: Normative data. *Behavior Research and Therapy*, 7(3), 335-339. [https://doi.org/10.1016/0005-7967\(69\)90018-7](https://doi.org/10.1016/0005-7967(69)90018-7)
- Tornio, S. (2019, March 14). *More kids than ever are dealing with test anxiety, and we need to help*. We Are Teachers. <https://www.weareteachers.com/test-anxiety/>

The Study of Developing and Validating the Union Bias Scale

Ender Kazak ^{1,*}

¹Düzce University, Faculty of Education, Department of Educational Sciences, Düzce, Turkey

ARTICLE HISTORY

Received: Oct. 20, 2020

Revised: July 21, 2021

Accepted: Oct. 07, 2021

Keywords:

Union bias,
Ingroup bias,
Ingroup favoritism,
Outgroup discrimination,
Social distance between
groups.

Abstract: Observation of different levels and types of organizational problems, such as school principals and teachers suing each other and conflicts between teachers, caused by union bias in schools in Turkey today, is the starting point of this study. This study, therefore, aimed to develop and validate a scale that helps determine union bias in schools. Participants of the study included teachers being a union member and working at a primary, middle or high school in the 2017-2018 and 2018-2019 academic years in Düzce, Turkey. During the data analysis, firstly, the first data set was examined in terms of the assumptions of the factor analysis and then Exploratory Factor Analysis (EFA) was applied. Confirmatory Factor Analysis (CFA) was performed on the second data set. Convergent validity of the scale was examined with item load values, mean-variance, and composite reliability coefficients. Discriminant validity was examined by the Fornell-Lacker criteria. Also, measurement invariance in gender groups was examined. The Cronbach Alpha and combined reliability coefficients were calculated to determine the scale's reliability. The Union Bias scale consists of 27 items and six dimensions. The explained total variance was 64%. As a result of the first and second order confirmatory factor analyses, it was revealed that the six-dimensional structure predicted Union Bias and the scale's structure did not differ in gender groups. The Cronbach's alpha value was .90 and the composite reliability was .96. As a result, the structure of the Union Bias Scale was concluded to be valid and reliable.

1. INTRODUCTION

The starting point of the study has been the observation of different levels and types of organizational problems between teacher-principal and teacher-teacher led by union bias in some schools in Turkey nowadays. These problems have resulted in several conflicts among teachers, affected school climate negatively, harmed communication climate, and even caused the teachers and principals to sue each other. The reflection of ingroup bias based on the social identity theory on a school comes into existence through union commitment and acts as union bias. Since ingroup bias is a concept generally identified with collectivist cultures, union bias is likely to appear in such cultures. The study may make sense in countries with collectivist cultures such as Turkey, but also the point that dimensions of union bias created due to different reasons, under recent conditions, in countries with individualist cultures may be an object of interest. The study is, therefore, crucial and necessary in terms of finding out the dimensions of

*CONTACT: Ender KAZAK ✉ enderkazak81@hotmail.com 📍 Düzce University, Faculty of Education, Department of Educational Sciences, Düzce, Turkey

union bias, observing its reflection on organizational environments, and understanding its relationship with organizational variables.

The word “union” is defined as “the unity created together by employees and employers in order to protect and develop their benefits with regard to work, income, social and cultural issues much more” in the Current Turkish Dictionary of the Turkish Language Association (TDK.gov.tr.). Unions started to show up in democratic western countries where the industrial revolution appeared during the first half of the 18th century. Urbanization that came out together with industrialization in the 17th century led to the rise of a working class having a poor standard of living, low wages, and bad working conditions as well as the emergence of an upper class. A search of voicing their demands in an organized manner to improve working conditions of labors and to provide better life standards generated unions (Güneş, 2013). Union movements in the Western world took place in the wake of a difficult and long period on a social base with the industrial revolution and formed its present-day condition. However, in Turkey, the process of industrialization started late. In the formation of unions in Turkey, a social base or social realities union movements in the West were based on did not appear and the union movements emerged in a factitious way under the government’s control (Özgiraz & Talu, 2008).

In the literature, there have been many theories regarding workers’ goals for being a union member. Those theories have been collected under two distinct titles as structural approaches and approaches on individual union membership manner. The former explains changes on the rates of labor union membership based on environmental factors. The latter explains it based on demographic, social, and attitudinal variables and on variables being peculiar to industry and business. Social psychological theories regarding union membership as a process focus on how individuals decide to be a union member. The social psychological theories clarifying the process of being a union member are classified as the frustration-aggression hypothesis, the rational choice theory, the social identity theory, the attribution theory, and interactionist theory (Seçer, 2009). This study is based on the social identity theory and ingroup bias among the social psychological theories clarifying the process of being a union member while union bias is discussed in the context of the social identity theory and ingroup bias.

The aim of education unions is to protect and improve common economic, social, professional, and union rights and benefits of their members and to provide a more prestigious standard of living (Eraslan, 2012). Nonetheless, unfortunately, the primary goal of today’s unions is not to protect and improve workers’ benefits, but to be a reflection of political and ideological opinions on business life (Özgiraz & Talu, 2008). In fact, it is a known fact that four confederations organized in the public sector in Turkey and the affiliated unions are mostly close to different long-established political views (Karaman & Erdoğan, 2016; Kayıkçı, 2013). The reasons why workers become a union member are based not only on economic factors, but also on ideological rationales. Especially, the ideological view is essential in choosing a union to be a member (Bayar, 2015). The ideological dimension implies a unity of values on union commitment between a union and its members. Commitment to union values and principals and reliance on the union are the matters of the ideological dimension. In the dimension, nonutilitarian union ideology is used to achieve union goals being close to their own values and to move collectively. In the utilitarian dimension, a union member makes gain-loss evaluation. There is a relationship in which short-term benefits such as wages, job security, and work safety are prioritized between the union and its members (Süreklı, 1998). On the other hand, teacher unions in ideological conflicts, because of membership and commitment based on ideological dimensions, push their duties on the protection of teachers’ social and economic rights and the attainment of new rights into the background (Mert, 2013). That unions have a political view is an intelligible phenomenon, but the perception of education unions as a political organization

rather than a professional teacher union is a significant obstacle in teacher organizations (Baysal & Yücel, 2010). Therefore, because of this political perception, teachers stay out of being a union member due to their concerns on alienation (Berkant & Gül, 2017), grouping, giving a negative impression to school principals, and being treated unfairly in consequence of being a different union member (Arslan, 2015; Demir, 2013; Karaman & Erdoğan, 2016). As a result, unions' claims to protect their members' rights that affect politics and policymakers could be accepted; nevertheless, union politicization is not a desired situation. This situation may lead to increase in social distance, harming workers' rights as well.

Since ideologically union commitment meets a requirement of a certain collective identity (Sürekli, 1998), being a union member helps workers develop feelings of sense of belonging into a group, collaborating, acting in unison and feeling that they are not alone in their professional lives (Baydar, 2016). The social identity theory advocates that people tend to perceive themselves and others as belonging to several groups because joining a group meets important psychological and social requirements such as belonging, attracting attention, overcoming much more difficult situations, feeling secure, protecting themselves from an outgroup, and having a positive social identity (Kağıtçıbaşı, 2008). According to this theory, people's needs of boosting their self-esteem lead them to better evaluate and glorify the group they get involved in than other groups, while making other groups less significant (Çimendağ, 2013). Obviously, a natural consequence of an understanding on regarding the group they are involved with as precious and the other group as worthless is that people show biased behaviors to the group they get involved with while they demonstrate discriminatory behaviors to the other group, namely the outgroup.

An ingroup is defined as a group of people with a sense of belonging and a shared identity, in other words, a community of "us"; an outgroup is identified as a group of people perceived as distinct and basically different from an ingroup, in other words, a community of "them". The definition of who we are describes who we are not. The circle including "us" (ingroup) excludes "them" (outgroup) (Myers, 2015). Other is an identity of "he/she/it" (outgroup) against "I" personally, and an identity of "they" (outgroup) is against "we" (ingroup) socially. Other is an entity that does not have the characteristics we have. To put it simply, anybody who is not himself/herself is the other (Yurdigül & İspir, 2015). Ingroup bias is identified as evaluating ingroup members more positively than outgroup members to improve one's self-esteem or stay ahead of the curve in intergroup relations (Çoksın, 2019), shortly as favoring one's own group (Myers, 2015). Ingroup bias is mostly linked with collectivist cultures. In such countries as Turkey, where a collectivist culture is dominant, making a distinction between "we" and "others", unethically, in social and administrative relations leads people in these groups to neglect principals of law and social values, and have an absolute bias against people in their own group or in any events or circumstances. This problem first results in polarization, then in discrimination, and finally in hatred, hostility and conflicts between different labeled identities (Akyürek, 2016). Hostility between groups occurs when an ingroup member shows negative attitudes to members of another group called as outgroup. There are three interrelated but distinguishable components of this kind of group hostility. The first component is that stereotypes, beliefs related to the most common features of group members, are cognitive. The second component is that prejudices, negative emotions towards the target group, are affective. In fact, both stereotypes and prejudices reflect cognitive and affective moods at the same time. The last component is that discrimination, making people at a disadvantage just because they are the other group's members and act upon, is behavioral (Tajfel et al., 1971; Taylor et al., 2007). Evaluating an outgroup based on stereotypes pioneers prejudices as stereotypes enhance specifically the feeling of sympathy within ingroups and discrimination in outgroups (Göregenli, 2012).

Considering all these ongoing intergroup problems today, the social contact theories offering a solution to these problems become more important (Küçükkömürlü & Sakallı-Uğurlu, 2017). Most studies on intergroup contact have been based on Allport's (1954) study named as "Nature of Prejudice." This theory focuses on the idea that the way to overcome the prejudice is "communication." The intergroup contact theory claims that interpersonal communication between different social group members is one of the most effective ways to promote positive intergroup attitudes (Dovidio et al., 2017; Pettigrew, 2016; Pettigrew & Tropp, 2006; Seat et al., 2015). The intergroup contact is an essential technique to overcome prejudices, but it requires common goals, equal status, institutional support (support of authority), and mutual close and ongoing contacts based on collaboration in order to be beneficial (Pettigrew, 2016; Taylor et al., 2007). No contact between social groups fosters prejudices, segregation, and social distance but supports discrimination (Çuhadar Gürkaynak, 2012). Discrimination against outgroup members reinforces their commitment to their own group (Keskinılıç Kara, 2016) and broadens the social distance.

1.1. The Purpose of the Study

It is indispensable that workers in the school environment have different beliefs, goals, cultures, and personality characteristics as can be seen in any working environment. The fact that teachers, even if they have a political/social identity, cannot be a member of a political party is stated in the Constitution of the Republic of Turkey. Unions, however, are one of the settings in which this function is partly carried out. Teachers in a union setting are able to discuss on daily political subjects as well as seek the rights of their members; however, the process of politicization has somewhat been continuing in unions. Then, such effects are necessarily reflected on emotions, thoughts, attitudes, and behaviors in organizational environments. These reflections may negatively influence personal relations in these environments and organizational variables such as organizational climate, organizational communication, motivation, etc. by sometimes reinforcing groupings and polarizations. Considering the studies on workers being a union member in Turkey, it is clear that the scales used are about the reasons why workers are a union member, their union commitment and their expectations from a union, and that these subjects have been studied mostly in the context of workers being a union member. However, this paper aims to develop a scale to measure union bias level in the context of ingroup bias unlike the studies in the literature. The developed "union bias scale" is to provide the researchers with an interdisciplinary study enabling them to consider the organizational, educational, and personal effects of union bias thereby helping them to explore the relationship of union bias with organization climate, organizational conflict, organizational cynicism, organizational trust, organizational justice, and communication climate. The scale is designed in order to apply it in all organizational settings including schools as areas of its application. It is also possible to carry out studies in school environments on such issues as political discrimination (Keskinılıç-Kara, 2016; Keskinılıç-Kara & Oguz, 2016), discrimination (Çelik, 2011; Polat & Hiçyılmaz, 2017), and favoritism (Erdem & Meriç, 2012; Erdem & Meriç, 2013; Polat & Kazak, 2014). It is likely to encounter studies on out-of-school environment like ingroup bias (Çimendağ, 2013; Hasta & Arslantürk, 2013; Kostakoğlu, 2010; Akyürek, 2016); however, the related literature shows no previous research conducted on investigating ingroup bias and union bias in the context of a school, especially in the national and international literature using the union bias scale. Therefore, the main objective of this study is to fill this gap in the literature by proposing a union bias scale.

2. METHOD

2.1. Research context and participants

The study aimed to develop and conduct the “Union Bias Scale” in the context of ingroup bias with its tested reliability and validity. Observation of different levels and types of organizational problems, such as school principals and teachers suing each other and conflicts between teachers, caused by union bias in schools in Turkey was the starting point of the study. Participants included teachers being a union member and working at a primary, middle or high school in Düzce.

The research sample included teachers working at a primary, middle or high school in Düzce Province, Turkey and its seven districts in the 2017-2018 (summer seminar) and 2018-2019 (spring term) academic years. The first implementation of the study was in the summer seminar term in the 2017-2018 academic year, while the second and third implementations were during the fall seminar term and the 2018-2019 academic year. To determine the research sample, the methods of convenience sampling and criterion sampling among non-random sampling methods were used. The criterion was to be a member of any education unions. For this reason, easily accessible schools in the central district and the seven districts of Düzce on the official website of Düzce Provincial Directorate of National Education were listed. The scale was conducted with teachers being a union member, working at those schools and volunteering in filling in the scale. Data was collected from teachers who were the members of the four major unions [Educators’ Trade Union (Eğitim Bir Sen), Turkish Education Union (Türk Eğitim Sen), Education and Science Workers’ Union (Eğitim İş), and Education and Science Workers’ Union (Eğitim Sen)] which had the maximum number of members in Düzce, Turkey. The scale was administered to 272 teachers in the first implementation EFA and to 243 teachers in the second implementation CFA by applying the rule of “being at least five times of the number of item” (Kline, 1994; Tavşancıl, 2014) for determining the sample size. The participants for conducting the scale consisted of a total of 329 teachers including 107 (35.52%) primary school; 145 (44.07%) middle school; and 77 (23.40%) high school teachers. Of all these participants 192 (58.35%) were male and 137 (41.64%) were female teachers. These participants were also composed of 160 (48.63%) teachers being a member of Eğitim Bir Sen; 106 (32.21%) teachers being a member of Türk Eğitim Sen; 32 (9.72%) teachers being a member of Eğitim İş; and 31 (9.42%) teachers being a member of Eğitim Sen. Also, 164 (49.84%) of the participants were the members of the same union with their school principals, while 165 (50.15%) of them were not. In terms of their professional seniority, 65 (19.75%) participants had 1-5 years of professional seniority; 76 (23.10%) participants were with 6-10 years of professional seniority; 66 (20.06%) participants with 11-15 years of professional seniority; 65 (19.75%) participants with 16-20 years of professional seniority, and 57 (20.06%) participants with 21 and above years of professional seniority.

2.2. The Process of Developing the Scale

In the process of the development of the Union Bias Scale (UBS), firstly, the literature on ingroup bias and unions/union members were reviewed and then an item pool including 59 items to represent the scale ideally was composed. Significant concepts constituting and determining ingroup bias (ingroup favoritism, outgroup discrimination; ingroup glorification, outgroup disdain; prejudices; stereotypes and social distance between groups) contributed to the development of the dimensions of the UBS. In addition to the literature review, the draft scale was revised by asking for opinions from one active union member teacher from each three different unions and two union representatives in the province, and an item pool was reconstituted with 65 items by adding six more items related to stereotypes between unions. Content validity of the scale was ensured by obtaining expert opinions. Later, opinions of three academicians studying on the subjects of “political discrimination” and “favoritism” and on

educational science were asked. In the light of the opinions, necessary corrections were made: 12 items that were seen as not being associated with statements, as being unsuitable in terms of meaning and expression, or as being interpreted differently were dropped from the scale. The final draft of the scale consisted of 53 items, which are thought to reflect all of the sub-dimensions, with a 5-point Likert-type, ranging from “*strongly disagree*”, “*disagree*”, “*slightly agree*”, “*quite agree*”, and “*strongly agree*”. When each grade is expressed verbally, the reliability coefficient is higher when compared to that of the numerical expression (Uyumaz & Çokluk, 2016). Before conducting the scale, necessary permissions were obtained from the Ministry of National Education, and the scales that the researcher distributed to teachers were gathered by the same researcher within the same day.

2.3. Data Analysis

Exploratory factor analysis was applied in order to reveal the factor structure of the scale within the validity study of the Union Bias Scale. First, EFA's assumptions were tested; then, factor analysis was conducted. The parameter estimation bias obtains the lowest value at the maximum likelihood method when the sample is larger than 200 (Uyumaz & Sırgancı, 2020). Therefore, in this study, maximum likelihood method was preferred for factor extraction because such assumptions were met. Due to the theoretical background of the Union Bias Scale and since its dimensions were considered as related, the direct oblimin method was preferred among oblique rotation methods. In addition, the cut-off value for factor loadings was determined as 0.50 for both AFA and CFA (Hair et al., 2009).

The accuracy of the factor structure of the Union Bias Scale, whose factor structure was revealed with EFA, was tested with CFA over a second data set. The sample used for the development study and the sample used for verification of the scale were different from each other. In other words, a second data set was used to make CFA. This data set consisted of 243 volunteer teachers working in Düzce. Before the CFA, the assumptions of the second data set were tested. To reveal whether the factor structure of the Union Bias Scale was provided or not in the first order; a second-order confirmatory factor analysis was applied to show that the dimensions of the scale came together and represented the variable of Union Bias as a supreme concept. Confirmatory factor analysis was calculated from the covariance matrix and based on the marginal maximum likelihood estimation (MLM) method (Joreskog, 1999). CFA model fit was examined with the factor load values of the items, the variance values they explained, and the model data fit index values. The cut-off value for the factor load value is .50, and the items with a factor load below this value are recommended to be excluded from the scale (Hair et al., 2009). R^2 is the square of the standardized factor load value of the items and gives the variance ratios explained in the factor of the variable and it is suggested that it should not be less than 0.40. Model data fit was examined by chi-square (χ^2), Standardized Root Mean Square Residual (SRMSR), Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Tucker Lewis Fit Index (TLI) (Brown, 2006).

Convergent validity and discriminant validity were examined after the construct validity studies. Convergent validity is the evaluation made to measure the correlation level of more than one indicator/item of the same structure that is in harmony. To ensure convergent validity, item factor load values should be ≥ 0.5 (Hair, et al., 2009), average variance extracted (AVE) value should be 0.5, and construct reliability values should be 0.7 (Fornell & Larcker, 1981). Discriminant validity refers to the degree to which the structure is empirically different from each other. It also measures the degree of differences between overlapping structures (Hair et al., 2014). In this study, the discriminant validity was examined with the Fornell-Lacker criterion (Fornell & Larcker, 1981). This method compares the square root of the average variance extracted (AVE) with the correlation of latent structures. A latent structure should better explain the variance of its own indicator rather than the variance of other latent structures.

Therefore, the square root of the AVE of each structure must have a greater value than the correlations with other latent structures (Hair et al., 2014).

In this study, measurement invariance in gender groups was examined as another validity proof. Measurement invariance is that the relationship between observed variables (items) and latent variables (measured structure) is the same between the examined subgroups (Widaman & Reise, 1997). In this study, measurement invariance between gender groups was tested with multi-group CFA. Besides, the statistical significance of the difference between the loads and interceptors of the items estimated according to gender groups with the alignment method was examined (Asparouhov & Muthén, 2014).

The most common measurements used for internal consistency are Cronbach alpha and composite reliability; they measure reliability based on the interrelationship of observed item variables. The values range from 0 to 1. A higher value indicates a higher reliability level. In exploratory research, the values of composite reliability/Cronbach alpha between 0.60 to 0.70 are acceptable, while in a higher stage the value has to be higher than 0.70 (Hair et al., 2014). However, the value that is more than 0.90 is not desirable and the value that is 0.95 or above is undesirable (Nunnally & Bernstein, 1994). Indicator reliability is the proportion of indicator variance explained by the latent variable. The values range from 0 to 1. The outer loadings value should be higher than 0.70 and it should be considered for deletion if the removal of the indicator with outer loadings which is between 0.40 and 0.70 and if it contributes to an increase in composite reliability and average variance extracted (Hair et al., 2014).

Hypothesis tests and exploratory factor analysis were performed with SPSS 20.0 and confirmatory factor analysis and multi-group confirmatory factor analysis were performed with Mplus 7.3. The explained average variance (AVE) and composite reliability (CR) were calculated in an Excel program using the formulas as suggested by Fornell and Larcker (1981).

3. FINDINGS

3.1. Findings Regarding the Validity of the Union Bias Scale

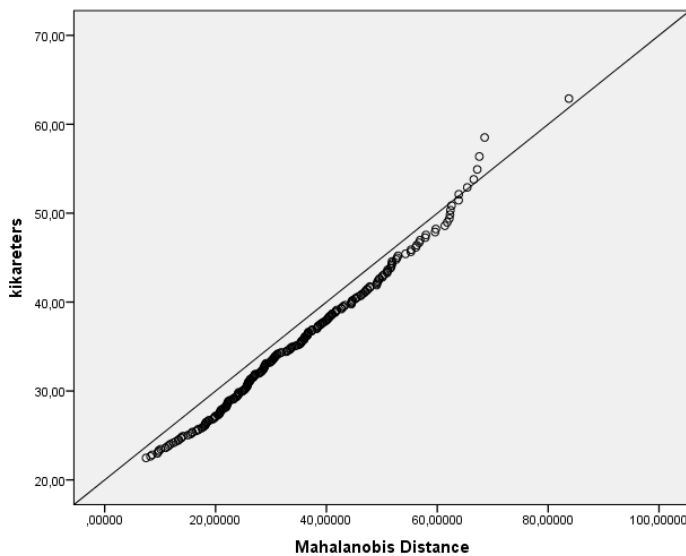
3.1.1. Exploratory factor analysis (EFA)

The factor structure of the Union Bias Scale was determined by the exploratory factor analysis applied to the data set collected from the first sample. Before the exploratory factor analysis was applied, the data set was tested in terms of the assumptions of the factor analysis, such as missing value, one-way and multivariate extreme value, univariate and multivariate normality, multicollinearity, and singularity. One-sided extreme values were examined by converting the item scores of the scale to the standard z score (Tabachnick & Fidel, 2007) and all of the standard scores that were outside the score range were removed from the 4 observation data sets of $\pm 4 z$ (Mertler & Vannata, 2005). Mahalanobis Distances (MU) were calculated for multivariate extreme value analysis and 19 MU values were extracted from observation data sets that exceeded $\alpha = 0.001$ and critical = 90.57 in 53 degrees of freedom (Tabachnick & Fidell, 2007). The skewness coefficients of the items varied between -1.816 and 2.658 and the kurtosis coefficients varied between -1.592 and 6.593. Chou and Bentler (1995) stated that the assumption of univariate normality is fulfilled as long as the coefficient of skewness is 3 and Kline (2005) stated that the assumption of univariate normality is fulfilled as long as the kurtosis coefficient does not exceed 10. Therefore, it is seen that the assumption of univariate normality is provided. Since the scatter plot (Figure 1) formed by squared Mahalanobis distance values (m_i^2) and inverse cumulative chi-square values show a linear structure, the assumption of multivariate normality is achieved (Alpar, 2011).

For multicollinearity, the dual correlations of the items were examined and no correlation value exceeding the critical value of $r = 0.85$ was found (Kline, 2005). The factor analysis was

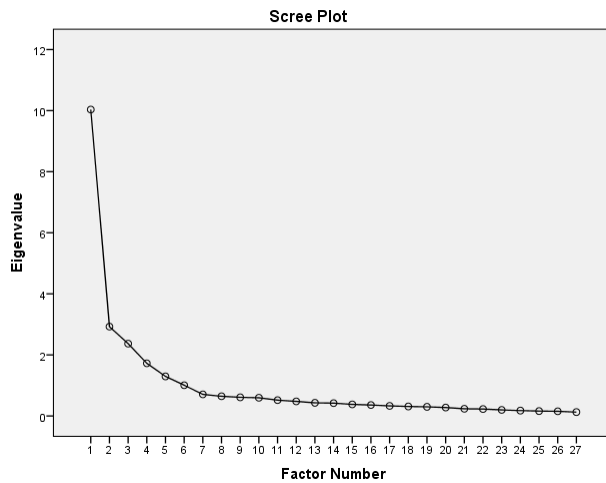
performed by removing 5 items (1-33-35-36-40) with item-total correlations below .30 from the data set (Nunnally & Bersntein, 1994).

Figure 1. *Multivariate Normality.*



As a result of testing the assumptions, 23 observations were extracted from the first sample consisting of 272 observations and EFA was applied to the data set of 249 people consisting of 48 items. Kline (2005) stated that a sample of 200 people is sufficient for factor analysis. The suitability of the data set related to the Union Bias Scale to the exploratory factor analysis was examined by Kaiser-Meyer-Olkin (KMO) and Bartlett tests. The KMO value approaching 1 means that each variable in the scale can be predicted by other variables and 0.60 and above is sufficient for social sciences (Kline, 2005). In this study, the KMO value was calculated as 0.92. When Bartlett test results are examined, it is seen that the value obtained as $\chi^2 = 8889.412$; $sd=1128$ ($p=0.000$) is significant at the 0.01 level. Therefore, it was concluded that the correlation matrix is different from the identity matrix. According to the KMO value and Bartlett test results, it was concluded that the data matrix of the Union Bias Scale consisting of 48 items is suitable for factor analysis.

In factor analysis, factor extraction was performed by using Direct Oblimin, maximum likelihood method, rotation, and oblique rotation techniques. To decide whether the items would be removed in EFA, the minimum level of factor loading was accepted as .30 (Tabachnick & Fidell, 2001). As a result of EFA, 27 out of 48 items in the item pool were grouped under six factors/dimensions, whose eigenvalues were greater than 1.0. 21 items (items 2, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 23, 24, 31, 32, 34, 39, 41, 43, 45, 53) that did not load on any factor, whose item factor loading was smaller than .30 and loaded under more than one factor and the difference between the load values less than 0.10 were eliminated from the analysis. The removal of items was performed one by one and the analysis was repeated after each removed item. As a result of the repeated factor analysis, a 6-factor/dimensional structure with an eigenvalue above 1.00 was observed. The scree plot given in [Figure 2](#) also shows that the items can be collected under six dimensions.

Figure 2. Scree Plot.

In [Table 1](#), factor load values of 27 items remaining after EFA are given. The factor loadings of the remaining 27 items before being subjected to rotation were found to be 0.349 and 0.737. After the oblique rotation technique was applied, it was observed that factor load values varied between 0.472 and 0.891. [Table 1](#) shows that the scale was composed of the first dimension with 3 items (factor loadings between 0.56 and 0.88); the second dimension with 5 items (factor loadings between 0.52 and 0.89); the third dimension with 6 items (factor loadings between 0.62 and 0.86); the fourth dimension with 5 items (factor loadings between 0.47 and 0.87); the fifth dimension with 4 items (factor loadings between 0.51 and 0.83); and the sixth dimension with 4 items (factor loadings between 0.52 and 0.89). It was found out that all factors explained 64.30% of the total variance: the first factor explained 35.60%; the second explained 8.64%; the third explained 7.90%; the fourth explained 5.47%; the fifth explained 3.87%; and the sixth explained 2.73% of the total variance. The explained variance ratios were found enough to be 30% in scales with one factor and to have ranged between 40% and 60% in scales with multi-factors (Büyüköztürk, 2006; Tavşancıl, 2014). Accordingly, the explained variance ratio was adequate.

Table 1. Item Load Values of the Union Bias Scale.

Item/Dimension	Social Distance	Stereotypes	In Group Glorification	Out Group Disdain	In Group Favoritism	Prejudices
Item 38	.879	.040	-.084	.054	.000	.021
Item 37	.868	.042	-.115	-.005	-.023	.026
Item 30	.565	.057	.013	.059	.195	.087
Item 50	-.075	.890	-.020	.016	.007	.023
Item 49	.054	.777	.025	.049	-.046	-.022
Item 52	-.048	.777	-.019	.063	-.045	.037
Item 51	.182	.653	.079	-.095	.197	.000
Item 48	-.036	.525	-.208	.061	-.092	.082
Item 26	-.003	.008	-.862	.018	-.038	.055
Item 27	-.010	-.015	-.807	.119	.022	.039
Item 25	.164	.009	-.802	.056	-.035	.004
Item 22	.198	.045	-.692	.016	.046	.010
Item 29	-.081	.073	-.660	.013	.067	-.030
Item 28	-.015	-.041	-.622	-.058	.147	-.007
Item 19	.011	.008	-.048	.875	-.032	.009
Item 18	-.027	.027	.022	.839	.093	-.037

Table 1. *Continues.*

Item 21	-.044	.162	-.046	.613	.074	.115
Item 17	.088	-.080	-.212	.495	.015	.160
Item 20	.323	-.022	.049	.472	.143	.062
Item 4	-.033	-.008	-.073	.028	.834	-.019
Item 5	-.017	.017	-.001	.085	.832	-.011
Item 3	.085	-.005	-.073	-.043	.678	.101
Item 8	.028	.017	-.132	.148	.515	.056
Item 46	.031	-.004	-.013	-.030	.010	.891
Item 44	-.079	.014	-.095	-.017	.025	.793
Item 42	.260	-.108	.066	.033	.091	.532
Item 47	-.029	.130	.057	.109	-.028	.519
Eigenvalue	4.690	4.004	6.304	6.202	5.712	5.537
Explained Variance	35.60	8.64	7.99	5.48	3.89	2.73
Cumulative Variance	35.60	44.24	52.22	57.70	61.57	64.30

Note. Factor load values of 0.20 and above are presented in the table.

When the items in the dimensions are examined, it is seen that there is a factorization consistent with the literature. Accordingly, since the items in the first dimension are Social Distance (SD), the items in the second dimension are Stereotypes (S), the items in the third dimension are In Group Glorification (IGG), the items in the fourth dimension are Out Group Disdain (OGD), the items in the fifth dimension are In Group Favoritism (IGF), and the items in the sixth dimension are related to Prejudices (P); the factors are named accordingly. The correlation coefficients between the dimensions of the scale are given in Table 2. It is seen that the correlation coefficients between dimensions change between 0.20 and 0.60. It has been revealed that the dimensions are in a positive, meaningful, and moderate relationship with each other.

Table 2. *Correlation Coefficients between Factors.*

	SD	S	IGG	OGD	IGF	P
SD	1	.569**	.568**	.600**	.399**	.196**
S		1	.531**	.465**	.545**	.352**
IGG			1	.370**	.275**	.251**
OGD				1	.453**	.244**
IGF					1	.327**
P						1

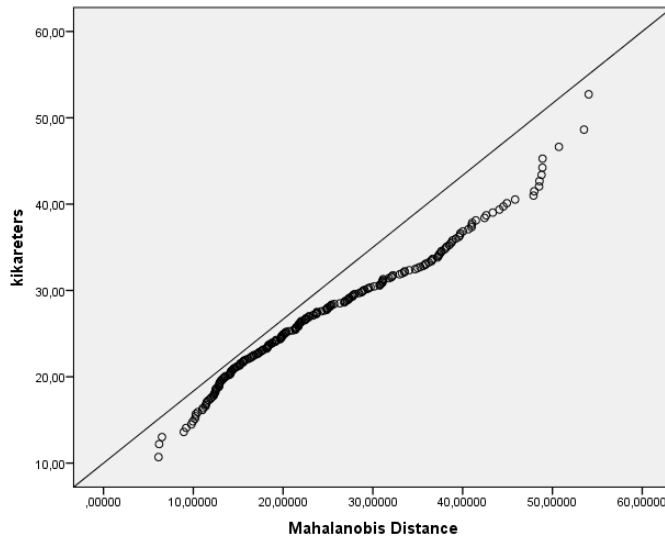
** $p < .01$

3.1.2. Confirmatory factor analysis (CFA)

A second-order confirmatory factor analysis was conducted because the important concepts associated with ingroup bias (ingroup favoritism, outgroup discrimination; in-group glorification, outgroup disdain; prejudices; stereotypes and social distance between groups) contributed to forming the dimensions of union bias (Allport, 1954; Myers, 2015; Pettigrew & Tropp, 2006; Tajfel et al., 1971; Taylor et al., 2007). For the one-way outlier analysis, the items of the scale were converted to the standard z score, and 10 observations outside the ± 4 z score range (Mertler & Vannata, 2005) were removed from the data set. For the versatile extreme value analysis, Mahalanobis Distances (MD) were calculated and 13 observations whose MD values exceeded $\alpha = 0.001$ and 27 degrees of freedom exceeding the critical value of $\chi^2 = 55.48$ were removed from the data set (Tabachnick & Fidell, 2007). The skewness coefficients of the items varied between 0.609 and 1.336 and the kurtosis coefficients varied between -1.090 and

5.840. Since the coefficient of skewness $|3|$ (Chou & Bentler, 1995) does not exceed the kurtosis coefficient $|10|$ (Kline, 2005), it has been determined that the assumption of univariate normality is provided. The scatter plot (Figure 3) formed by squared Mahalanobis distance values (M_i^2) and inverse cumulative chi-square values show a structure close to the linear. Therefore, it can be said that the assumption of multivariate normality is also provided.

Figure 3. *Multivariate Normality.*



For multicollinearity, the binary correlations of the items were examined and no correlation value exceeding the critical value of $r = 0.85$ was found (Kline, 2005). In Table 3, the standardized factor load (λ_i) obtained as a result of the first and second-order confirmatory factor analysis of the six-dimensional structure of the Union Bias Scale and the variance (R^2) explained by the items and the goodness of fit values are given. Besides, the diagrams for first and second-order factor analysis are presented in Figures 4a and 4b, respectively. Standardized factor loadings show the contribution of the item/indicator to the relevant factor. Accordingly, the factor loads of the items in the "In Group Favoritism" dimension ranged between 0.61 and 0.87; between 0.58 and 0.84 in the "Out Group Disdain" dimension; between 0.60 and 0.84 in the "In Group Glorification" dimension; between 0.62 and 0.83 in "Social Distance" dimension; between 0.64 and 0.81 in the "Prejudices" dimension; between 0.55 and 0.90 in "Stereotypes" dimension, and these values are higher than 0.5 specified as the acceptable factor load (Hair et al., 2009). When the variance values explained by the items were examined, it was seen that the acceptance value of six items was below 0.40. Since the variance values explained by these items were very close to the limit value, the model was examined together with the goodness of fit and item reliability index values and it was decided to keep the items in the scale. When the goodness of fit indexes regarding the first order CFA were evaluated, the rate of χ^2/sd was found to be 1.60 ($\chi^2/sd=493.870/309$). When this value is $0 < \chi^2/sd < 3$, it shows a perfect consistency (Schermelleh-Engel, Moosbrugger & Müller, 2003). It was found that the value of RMSEA was .052, Comparative Fit Index (CFI) was .97, Tucker Lewis Index (TLI) was 0.91, and Standardized Root Mean Square Residual (SRMR) was 0.063. In the literature, acceptable limit values for the goodness of fit values are in the range of 0.90-1.00 for CFI and TLI values (Bentler & Bonnet, 1980; Tucker & Lewis, 1973); for RMSEA and SRMR values, it is reported that the lower limit should be 0 and the upper limit should be 0.08 (Hooper, Coughlan, & Mullen, 2008). When the findings are evaluated together, it is seen that the six-factor structure of the Union Bias Scale revealed by EFA is confirmed by CFA.

A second-order confirmatory factor analysis was conducted to show that the dimensions of "In Group Favoritism", "Out Group Disdain", "In Group Glorification", "Social Distance", "Prejudices", and "Stereotypes", which were obtained with the first-order confirmatory factor analysis of the Union Bias Scale, come together, and represent the higher dimension of the Union Bias variable (Büyüköztürk, 2002). The relationships between the latent variables obtained in the first-order factor analysis were used as the basis for the model examined. The variances explained by the bias variable in the first-order variables were revealed by the analysis. The factorial model of the second-order CFA result is presented in Figure 4b and the standardized factor load values and explained variance values regarding the factor-item relationship are presented in Table 3. The results of testing the second-order factor model by adding the second-order "bias" latent variable to the first-order confirmatory structure tested with six latent and 27 indicator variables showed that the goodness of fit values were: $\chi^2/df = 1.65$ (523.589/318), CFI = 0.91, TLI = 0.90, RMSEA = 0.056, and SRMR = 0.076. These values reveal that the data show an acceptable fit.

Table 3. Standardized Factor Loads (λ_i) of the Items of Union Bias Scale and Explained Variance (R^2) Values.

Factor	Item	(λ_i)	First Order	Second Order				
			R^2	(λ_i)	R^2			
IGF	Item 3 (y1)	0.630	0.40	0.630	0.40			
	Item 4 (y2)	0.874	0.76	0.873	0.76			
	Item 5 (y3)	0.800	0.64	0.803	0.64			
	Item 8 (y4)	0.606	0.37	0.605	0.37			
OGD	Item 17 (y5)	0.582	0.34	0.582	0.34			
	Item 18 (y6)	0.784	0.61	0.783	0.61			
	Item 19 (y7)	0.839	0.70	0.839	0.70			
	Item 20 (y8)	0.666	0.44	0.670	0.45			
	Item 21 (y9)	0.806	0.65	0.805	0.65			
IGY	Item 22 (y10)	0.753	0.57	0.756	0.57			
	Item 25 (y11)	0.843	0.71	0.849	0.72			
	Item 26 (y12)	0.772	0.59	0.775	0.60			
	Item 27 (y13)	0.775	0.60	0.770	0.59			
	Item 28 (y14)	0.596	0.36	0.585	0.34			
	Item 29 (y15)	0.609	0.37	0.603	0.36			
SD	Item 30 (y16)	0.622	0.39	0.627	0.39			
	Item 37 (y17)	0.834	0.70	0.824	0.68			
	Item 38 (y18)	0.786	0.62	0.793	0.63			
P	Item 42 (y19)	0.636	0.40	0.622	0.39			
	Item 44 (y20)	0.777	0.60	0.782	0.61			
	Item 46 (y21)	0.813	0.66	0.821	0.67			
	Item 47 (y22)	0.583	0.34	0.577	0.33			
KYG	Item 48 (y23)	0.588	0.35	0.584	0.34			
	Item 49 (y24)	0.667	0.44	0.667	0.44			
	Item 50 (y25)	0.905	0.82	0.901	0.81			
	Item 51 (y26)	0.553	0.30	0.559	0.31			
	Item 52 (y27)	0.702	0.49	0.707	0.500			
Goodness of Fit Values		χ^2	sd	χ^2/sd	CFI	TLI	RMSEA	SRMR
First Order		493.870	309	1.60	0.92	0.91	0.052	0.063
Second Order		523.589	318	1.65	0.91	0.90	0.054	0.076

The factor loadings between the first-order latent variables in the model and the higher level (second-order) variable and the explanation ratios of the second-order variable in the first-order variables (R^2) are presented in Table 4.

Table 4. Second Order CFA Standardized Factor Load and Explained Variance Values.

Second Order Variable	First Order Variable	λ	R^2
Union Bias	IGF	0.524	0.275
	OGD	0.853	0.727
	IGG	0.582	0.339
	SD	0.412	0.169
	P	0.848	0.719
	S	0.359	0.129

Figure 4a. First Order CFA

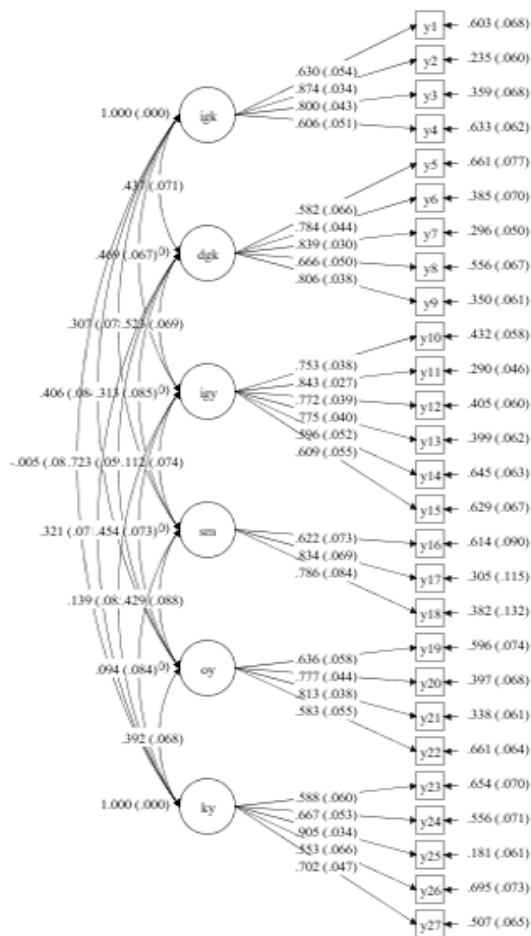
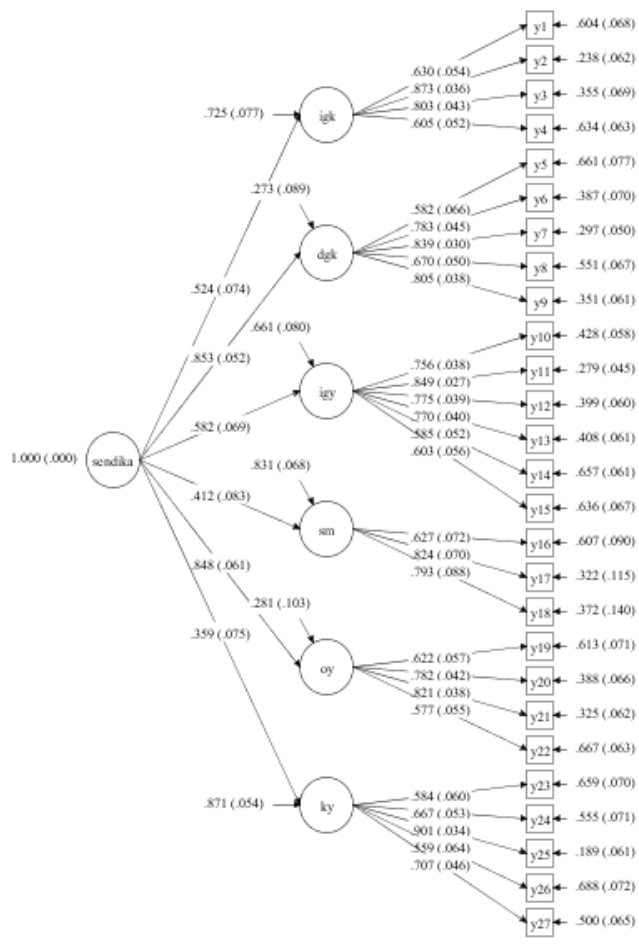


Figure 4b. Second-Order CFA



When Figure 4b and Table 4 are evaluated together, the strongest relationship between the latent variable "union bias" and the first-order latent variables is seen in union bias and the Out Group Disdain and Prejudice factors, and the weakest relationship is seen with the factors of Social Distance and Stereotypes. Looking at the variances explained by the second-order variable in the first-order variables, among the first-order variables, the most variability was explained in the Out Group Disdain and Prejudice factors, and the least variability was explained in the Social Distance and Stereotypes factors.

As is displayed in Table 5, the correlation coefficients between the scores obtained from all the items and the scores obtained from the factors and the scale were calculated and the

discrimination rate of each item was determined to reveal the extent to which each item served the general purpose of the factor. For this, item subscale correlations, item-test correlations, and alpha reliability of the scale were reported when each item was deleted.

Item-total correlations ranged between 0.52-0.76 in In Group Favoritism; between 0.54-0.77 in Out Group Disdain; between 0.57-0.75 in In Group Glorification; between 0.55-0.69 In Social Distance; between 0.52-0.70 in Prejudices; and between 0.49-0.77 in Stereotypes. When the item-test correlation coefficients for the whole scale were examined, the lowest correlation value was found to be 0.25 and the highest correlation was found to be 0.67. Each item had a significant and positive relationship with the overall scale ($p < 0.001$). The acceptable value for item-total correlations is around 0.20 (Kalaycı, 2010). Since there was no significant increase in alpha reliability when the item with the lowest correlation value was removed from the scale, it was decided to keep this item in the scale. These coefficients are validity coefficients for the discrimination of all items, and they show consistency of the items both with their dimensions and with the whole scale. When the alpha reliability values given in the last column are examined, it can be said that each item contributes at similar levels to the whole scale.

Table 5. *Item-Total Correlations on the Basis of Dimensions and Scales.*

Dimensions	Items	Item-Subscale Correlation	Alpha if Item Deleted	Item-Test Correlation	Alpha if Item Deleted
IGF	Item 3 (y1)	.549	.793	.408	.888
	Item 4 (y2)	.757	.688	.462	.887
	Item 5 (y3)	.694	.724	.446	.887
	Item 8 (y4)	.525	.813	.482	.887
OGD	Item 17 (y5)	.538	.851	.526	.886
	Item 18 (y6)	.707	.811	.669	.882
	Item 19 (y7)	.773	.790	.619	.883
	Item 20 (y8)	.614	.837	.525	.886
	Item 21 (y9)	.717	.806	.625	.883
IGG	Item 22 (y10)	.672	.844	.530	.886
	Item 25 (y11)	.753	.829	.548	.885
	Item 26 (y12)	.724	.834	.478	.887
	Item 27 (y13)	.695	.839	.619	.883
	Item 28 (y14)	.566	.864	.394	.889
	Item 29 (y15)	.596	.856	.380	.889
SD	Item 30 (y16)	.553	.778	.287	.890
	Item 37 (y17)	.688	.645	.297	.890
	Item 38 (y18)	.655	.685	.338	.889
P	Item 42 (y19)	.520	.768	.508	.886
	Item 44 (y20)	.655	.698	.598	.884
	Item 46 (y21)	.703	.676	.588	.884
	Item 47 (y22)	.517	.777	.476	.887
S	Item 48 (y23)	.491	.802	.253	.893
	Item 49 (y24)	.597	.770	.280	.892
	Item 50 (y25)	.770	.718	.385	.889
	Item 51 (y26)	.514	.795	.287	.892
	Item 52 (y27)	.621	.763	.402	.888

3.1.3. Convergent Validity

The convergent validity of the scale was examined by considering item factor loads, inferred mean-variance, and combined reliability. It is seen that item factor loads presented in Table 3 have a cut-off value above 0.5 in both first and second order CFA (Hair et al., 2009). The average variance and combined reliability values obtained are presented in Table 9. Accordingly, the average variance value extracted from the whole scale and its dimensions are above 0.5. The average variance extracted (AVE) value for only the stereotypes dimension was 0.48. The average variance extracted value extracted only in the stereotypes dimension was found to be 0.48. Fornell and Larcker (1981) stated that the convergent validity of the construct is still sufficient if the mean-variance is less than 0.5, but the composite reliability is higher than 0.6. Therefore, it can be said that this dimension also has convergent validity. Besides, it is seen that the structural reliability values (combined reliability and Cronbach alpha) are higher than 0.7 (Fornell & Larcker, 1981) in all dimensions and the entire scale. When all findings are evaluated together, it is seen that the convergent validity of the Union Bias Scale is provided.

3.1.4. Discriminant Validity

In this study, the differential validity was examined by comparing the square root of the mean-variance (AVE) and the correlation of latent structures. Factors are considered to be discriminatory when the square root of AVE values is greater than the correlations between latent variables (Fornell & Larcker, 1981). In Table 6, it is seen that the average variance inferred by each dimension is higher than the relationships between dimensions. Therefore, it was revealed that the discriminative validity of the scale was also provided.

Table 6. Discriminant Validity Findings.

	IGF	OGD	IGG	SD	P	S
IGF	0.735	.569**	.568**	.600**	.399**	.196**
OGD		0.742	.531**	.465**	.545**	.352**
IGG			0.728	.370**	.275**	.251**
SD				0.756	.453**	.244**
P					0.707	.327**
S						0.693

** $p < .01$

3.1.5. Measurement Invariance

In this study, the multi-group CFA (WG-CFA) analysis was conducted to test whether the factor structure of the Union Bias Scale differentiated in gender groups. Besides, with the alignment method, it was examined whether the difference between the factor load and intercept values of each item was statistically significant or not. Table 7 includes the findings of WG-CFA. Sokolov (2019) suggested considering the CFI value of measurement invariance with WG-DFA and stated that the relative goodness of fit should be CFI < 0.01 to ensure weak invariance and strong invariance as cut-off values. Accordingly, when Table 7 is examined, it is seen that strong invariance with weak invariance in all dimensions is very close to the limit value. Also, gradually, it is seen that the chi-square difference between models is not statistically significant. Weak invariance is based on the assumption that factor loads between groups are equal. Thus, factor variances and structural relationships between groups are comparable. When the results are evaluated together, it can be said that the factor loads of the Union Bias Scale are equal between gender groups. In the Social Distance dimension, fit index values for the structural model were not produced. This is thought to be due to the number of items.

Table 8 includes the findings regarding the factor load and intercept of each item and the statistical significance of their differences in gender groups.

Table 7. Findings of Multi-Group CFA Analysis.

	Model	χ^2	df	p	CFI*	Δ CFI
IGF	Weak-Structural	3.391	3	0.3352	0.974	-0.001
	Strong-Structural	11.344	6	0.0783	0.973	-0.018
	Strong-Weak	7.954	3	0.0470	0.956	-0.017
OGD	Weak-Structural	6.327	4	0.1760	0.950	-0.006
	Strong-Structural	13.386	8	0.0992	0.944	-0.013
	Strong-Weak	7.059	4	0.1338	0.937	-0.007
IGG	Weak-Structural	6.029	5	0.3034	0.942	-0.001
	Strong-Structural	12.675	10	0.2424	0.941	-0.003
	Strong-Weak	6.646	5	0.2483	0.939	-0.002
P	Weak-Structural	4.843	3	0.1836	0.945	-0.005
	Strong-Structural	12.068	6	0.0605	0.940	-0.017
	Strong-Weak	7.225	3	0.0651	0.928	-0.012
S	Weak-Structural	4.495	4	0.3431	0.982	-0.001
	Strong-Structural	14.650	8	0.0663	0.981	-0.018
	Strong-Weak	10.154	4	0.0379	0.964	-0.017

*CFI values are presented in order of structural, weak and strong invariance.

The results of the alignment analysis show that there is no statistically significant difference between the factor loads and intercepts of all items except the y19 coded item. Therefore, this result shows that strong invariance is provided, which assumes that both factor loadings and intercepts are invariant between gender groups. Therefore, it is possible to compare factor averages and intercepts between gender groups. The cut-off value of the y19 coded item in gender groups did not differ significantly, but the load value varied. Accordingly, it can be said that while weak invariance is provided in this item, strong invariance is not provided. As a result, it can be said that all the items of the Union Bias Scale are invariant for men and women.

Table 8. Test of Significance of Item Loadings and Intercepts Between Gender Groups.

		Group		Value 1	Value 2	Difference	SE	P-value	
IGF	Intercept	Item 3 (y1)	2	1	1.455	1.462	-0.007	0.043	0.873
		Item 4 (y2)	2	1	1.887	1.808	0.079	0.185	0.672
		Item 5 (y3)	2	1	1.727	1.713	0.014	0.055	0.798
		Item 8 (y4)	2	1	2.270	2.434	-0.164	0.184	0.373
	Loading	Item 3 (y1)	2	1	0.542	0.529	0.013	0.075	0.866
		Item 4 (y2)	2	1	0.998	1.090	-0.092	0.162	0.569
		Item 5 (y3)	2	1	1.041	0.820	0.222	0.174	0.203
		Item 8 (y4)	2	1	0.711	0.729	-0.018	0.103	0.864
OGD	Intercept	Item 17 (y5)	2	1	1.849	1.884	-0.036	0.093	0.700
		Item 18 (y6)	2	1	2.207	2.336	-0.129	0.168	0.443
		Item 19 (y7)	2	1	2.234	2.044	0.190	0.142	0.180
		Item 20 (y8)	2	1	1.544	1.517	0.027	0.048	0.574
		Item 21 (y9)	2	1	2.105	2.102	0.003	0.035	0.924
	Loading	Item 17 (y5)	2	1	0.578	0.587	-0.009	0.082	0.915
		Item 18 (y6)	2	1	0.990	1.008	-0.018	0.116	0.876
		Item 19 (y7)	2	1	1.117	1.062	0.055	0.136	0.687
		Item 20 (y8)	2	1	0.441	0.503	-0.062	0.112	0.579
		Item 21 (y9)	2	1	1.082	0.754	0.328	0.167	0.050

Table 8. *Continues.*

IGG	Intercept	Item 22 (y10)	2	1	2.306	2.267	0.040	0.071	0.573
		Item 25 (y11)	2	1	2.552	2.561	-0.009	0.057	0.876
		Item 26 (y12)	2	1	3.014	3.096	-0.081	0.097	0.400
		Item 27 (y13)	2	1	2.855	3.057	-0.202	0.183	0.267
		Item 28 (y14)	2	1	3.020	2.864	0.157	0.157	0.320
	Item 29 (y15)	2	1	3.481	3.466	0.015	0.044	0.733	
	Loading	Item 22 (y10)	2	1	0.706	0.789	-0.083	0.104	0.422
		Item 25 (y11)	2	1	0.879	0.969	-0.090	0.105	0.390
		Item 26 (y12)	2	1	0.874	0.951	-0.077	0.098	0.432
		Item 27 (y13)	2	1	0.962	0.810	0.151	0.137	0.271
Item 28 (y14)		2	1	0.716	0.661	0.055	0.087	0.529	
Item 29 (y15)	2	1	0.684	0.564	0.120	0.131	0.360		
SD	Intercept	Item 30 (y16)	2	1	0.526	0.482	0.044	0.433	0.919
		Item 37 (y17)	2	1	0.272	0.331	-0.059	0.306	0.848
		Item 38 (y18)	2	1	0.265	0.265	0.000	0.074	0.999
	Loading	Item 30 (y16)	2	1	0.417	0.405	0.012	0.051	0.810
		Item 37 (y17)	2	1	0.542	0.526	0.016	0.066	0.811
		Item 38 (y18)	2	1	0.566	0.614	-0.048	0.098	0.623
P	Intercept	Item 42 (y19)	2	1	1.205	1.208	-0.004	0.009	0.671
		Item 44 (y20)	2	1	1.660	1.677	-0.017	0.025	0.512
		Item 46 (y21)	2	1	1.521	1.376	0.145	0.221	0.512
		Item 47 (y22)	2	1	2.006	2.356	-0.350	0.217	0.106
	Loading	Item 42 (y19)	2	1	0.243	0.542	-0.299	0.116	0.010*
		Item 44 (y20)	2	1	0.882	0.835	0.047	0.174	0.788
		Item 46 (y21)	2	1	0.805	0.862	-0.057	0.173	0.740
		Item 47 (y22)	2	1	0.416	0.395	0.021	0.102	0.841
S	Intercept	Item 48 (y23)	2	1	4.546	4.499	0.048	0.164	0.772
		Item 49 (y24)	2	1	4.142	4.191	-0.049	0.190	0.796
		Item 50 (y25)	2	1	4.635	4.564	0.071	0.230	0.757
		Item 51 (y26)	2	1	3.343	3.388	-0.044	0.125	0.723
		Item 52 (y27)	2	1	3.917	3.920	-0.003	0.129	0.980
	Loading	Item 48 (y23)	2	1	0.681	0.594	0.087	0.180	0.628
		Item 49 (y24)	2	1	0.848	0.697	0.151	0.209	0.470
		Item 50 (y25)	2	1	0.955	1.104	-0.149	0.213	0.484
		Item 51 (y26)	2	1	0.582	0.702	-0.120	0.179	0.503
		Item 52 (y27)	2	1	0.711	0.710	0.001	0.019	0.958

* $p < .05$

3.2. Reliability

The internal consistency reliability of the Union Bias Scale was examined with the Cronbach Alpha and the indicator reliability and the composite reliability coefficients. Table 9 shows the Cronbach's alpha, average variance, and composite reliability coefficients of the scale. When Table 9 is examined, it can be seen that the Cronbach alpha value of the scale is .81 for the in group favoritism dimension; .85 for the out group disdain dimension; .87 for the in group glorification dimension; .78 for the social distance dimension; .78 for the prejudices dimension, and .81 for the stereotypes dimension. Cronbach's alpha reliability of the whole scale is .89.

Composite reliability is above the limit value of .70 in each dimension and the whole scale (Hair et al., 2014). When the findings are evaluated together, it is seen that the reliability of the scale is also provided.

Table 9. *Cronbach's Alpha, Mean Variance and Composite Reliability Coefficients of the Scale.*

		Cronbach Alpha	Mean Variance	Composite Reliability
Dimensions	IGF	0.81	0.54	0.82
	DGK	0.85	0.55	0.86
	IGG	0.87	0.53	0.87
	SD	0.78	0.57	0.79
	P	0.78	0.50	0.80
	S	0.81	0.48	0.82
Whole Scale	Union Bias	0.89	0.53	0.96

4. DISCUSSION and CONCLUSION

This study aimed to develop a reliable and valid scale on union bias and apply it in order to measure union bias. The union bias scale was constructed in the context of ingroup bias. The 27-items scale (see Appendix Table A) including only positive wordings was a 5-point Likert-type scale (ranging between extremely disagree, disagree, slightly agree, quite agree, and extremely agree). As a result of EFA, 27 items were grouped under six factors, whose eigenvalues were greater than 1.0. The scale's seven-factor structure accounted for 64.30% of the total variance. To determine the accuracy of the six-dimensional structure of the Union Bias Scale determined by EFA, first and second-order CFA was applied to the 27-item structure of the scale. It was also revealed that the scale has convergent validity, discrimination validity, and measurement invariance. Also, measurement invariance in gender groups was examined. The Cronbach Alpha and combined reliability coefficients were calculated to determine the scale's reliability. As a result of the first and second order confirmatory factor analysis, the scale's structure did not differ in gender groups. The Cronbach's alpha value was .90 and the composite reliability was .96. The Cronbach alpha reliability and composite reliability of the union bias scale were found above 0.70, which is the limit value, both in each dimension and in the entire scale. When the findings were evaluated together, it was demonstrated with different validity and reliability determination methods that the six-dimensional structure of the Union bias Scale consisting of 27 items is valid and reliable.

If unions enhance productivity, and if this productivity is economically crucial, the spread of unionism is essential (Doucouliagos et al., 2005). Understanding the unique identities of teachers' unions is possible by understanding the contexts in which the unions operate, understanding the relevant literature analysis (the concepts they refer to), and also understanding the dynamics within the union (Charlie, 2002). The effect of teacher unions on school output depends on the definition of unionization and its goals (Guthery, 2018). Therefore, it is possible to hear and see that different actions are taken even if the unions' objectives are the same or similar in theory. It can be said that teacher unions should focus on what they can do for teachers rather than what they can give to the government (or the political parties they are associated with) (Bağcı, 2009). Unions rely on active members to achieve their goals by participating in public meetings, strikes, civil disobedience, and political actions and defending union positions and policies in their schools (Popiel, 2013). Since an active member has a high union affiliation, the level of identification with his/her union is also high. People with a high level of identification with their groups are expected to exhibit higher levels of ingroup bias than others (Çoksan, 2016). This situation increases the impact of union bias in

organizational environments. Those unions' engagement in political efforts rather than in the improvement of teachers' rights (Yalçın Durmuş, 2018) increases the grouping among teachers (Kara, 2016). The reflection of the unions on the school climate is felt negatively in inner group bias. One way to decrease the level of union bias at schools may be to have common goals of unions, organize common activities, and therefore encourage teachers to communicate with each other. As long as the tendency on politicization of unions goes down, it is expected that union bias as its reflection on the organizational environments also decreases. Politicians, practitioners, and researchers are also expected to find solutions to reduce such negative effects as well.

There are no scales for trade union bias in the literature. Although union bias is a stronger form of union affiliation, it does not contain a harmless sense of belonging such as union affiliation. Union bias includes meanings such as factionalism and partiality in Turkish and it refers to fanaticism that has more destructive consequences. Especially in organizational environments, it can have effects that deeply influence the organizational climate, communication climate, organizational justice, and organizational trust feelings. One way of understanding these negative effects caused by union bias was thought to require the presence of a measurement tool that measures union bias and this study was thought to overcome this deficiency.

Based on these comments, the following recommendations can be made:

- 1- The scale can be applied in schools, enterprises, and organizations employing workers due to its design as it is not limited to a certain institution and it includes general expressions; it can also be applied to employees who are members of different unions such as civil servant unions or trade unions.
- 2- The relationship of trade union bias with social identity or variables such as organizational climate, communication climate, organizational justice, organizational trust, and organizational cynicism can be investigated.
- 3- The scale can be applied to different union member teachers and their views can be compared.
- 4- Comparisons can be made by applying the scale at different levels of education.
- 5- Since union bias is a phenomenon that is affected by the effectiveness of the union, the scale can be applied to employees with too many and too few members in certain regions or cities and comparisons can be made.

Acknowledgments

The study was presented as an abstract of paper in VI. International Eurasian Educational Research Congress / EJERCongress 2019.

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 **Ethics Committee Number:** Düzce University, 17/09/2020 - 2020/181.

Authorship Contribution Statement

Ender Kazak: Investigation, Resources, Introduction, Methodology, Software, Analysis, Findings, Discussion, Supervision, and Validation, Writing original draft.

ORCID

Ender KAZAK  <https://orcid.org/0000-0001-5761-6330>

5. REFERENCES

Akyürek, S. (2016). Türkiye'de iç grup yanlılığının toplumsal adalet ve güvenliğe etkisi [The effect of ingroup bias on social justice and security in Turkey]. *The Journal of Europe-*

- Middle East Social Science Studies*, 2(2), 161-179. <https://dergipark.org.tr/tr/download/article-file/174247>
- Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison Wesley.
- Alpar, R. (2011). *Çok değişkenli istatistiksel yöntemler [Multivariate statistical methods]*. Detay Yayıncılık.
- Arslan, C. (2015). *Öğretmenlerin sendikal örgütlenmeye ilişkin tutumları ve sendikal örgütlenme nedenleri [Teachers' attitudes towards teachers unions and the reasons of their unionization]*. [Unpublished master's thesis]. Cumhuriyet University.
- Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), 495-508.
- Bağcı, A. (2009). *Application of total quality management in teacher unions from the perspectives of union members* [Unpublished master's thesis]. Middle East Technical University.
- Bayar, L. S. (2015). *The effect of belonging to a labour union on work values and work behaviours*. [Unpublished Doctoral Dissertation]. Dokuz Eylül University.
- Baydar, F. (2016). *Analysis of teachers' opinions on the role of unions for constitution of educational policy*. [Unpublished master's thesis]. Marmara University.
- Baysal, Ö., & Yücel, C. (2010). Elementary school teachers' attitudes toward unions: sample of Uşak Province. *Educational Administration: Theory and Practice*, 16(3), 329-352. <https://dergipark.org.tr/tr/download/article-file/108220>
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588-606. <https://doi.org/10.1037/0033-2909.88.3.588>
- Berkant, H. G., & Gül, M. (2017). Union member teachers' perceptions and expectationstowards unions. *Journal of the Human and Social Sciences Researches*, 6(1), 419-442.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.
- Büyüköztürk, Ş. (2002). Factor analysis: basic concepts and using to development scale. *Educational Administration in Theory & Practice*, 8(4), 470-483. <https://dergipark.org.tr/tr/download/article-file/108451>
- Büyüköztürk, S., Kılıç Çakmak, E., Akgün, O. E., Karadeniz, S., & Demirel, F. (2011). *Bilimsel araştırma yöntemleri [Scientific research methods]*. 8. Baskı, Pegem Akademi Yayınları.
- Büyüköztürk, Ş. (2006). *Sosyal bilimler için veri analizi elkitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum [Data analysis handbook for social sciences: Statistics, research design, SPSS applications and interpretation]*. 6. Baskı, Pegem A. Yayıncılık.
- Charlie, N. (2002). *Reconciling teacher unionism's disparate identities: A View from the Field*. BCTF Research Report. British Columbia Teachers' Federation, Vancouver.
- Chou, C. P., & Bentler, P. M. (1995). Estimation and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37-55). Sage.
- Çelik, G. (2011). *İlköğretim okullarında örgütsel ayrımcılık ve öğretmenlerin tükenmişlik 654 düzeylerine etkisi [Organizational discrimination in primary schools and its effects on teachers' burnout levels]*. [Unpublished master's thesis]. Sakarya University.
- Çimendağ, F. Ş. (2013). *Yüksek ve düşük statülü gruplarda iç grup ve dış grup yanlılığı [In-group and out-group favouritism in high and low status groups]*. [Unpublished master's thesis]. Mersin University.

- Çokluk, Ö., Şekercioglu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları [Multivariate Statistics for Social Sciences: SPSS and LISREL Applications]*. Pegem Academy Press.
- Çoksan, S. (2016). *Sosyal kimlik özdeşimi ve iç grup normunun iç grup yanlılığı ile ilişkisi [Social identity identification and the relationship of ingroup norm to ingroup bias]* [Unpublished master's thesis]. Mersin University.
- Çoksan, S. (2019). Causal attributions of in-group favoritism and equal allocation between in and out-group. *Nesne*, 7(14), 83-101. <https://doi.org/10.7816/nesne-07-14-06>
- Çuhadar Gürkaynak, E. Ç. (2012). *Toplumsal temas: Önyargı ve ayrımcılığı önlemek için bir sosyal değişim aracı olarak kullanılabilir mi? [Social contact: Can it be used as a tool for social change to avoid prejudice and discrimination?]* Çayır, K. & Ceyhan, M. A. (derl.), *Ayrımcılık çok boyutlu yaklaşımlar [Multidimensional approaches to discrimination]* (255-265). İstanbul Bilgi Üniversitesi Yayınları.
- Demir, F. (2013). Sendikaların kuruluşu ve işleyişi. [Establishment and functioning of trade unions]. *Çalışma ve Toplum*, 4, 17-42.
- Doucoulagos, H., Laroche, P., & Stanley, T. (2005). Publication Bias in Union-Productivity Research? *Relations industrielles/Industrial Relations*, 60(2), 320-347. <https://doi.org/10.7202/011724ar>
- Dovidio, J. F., Love, A., Schellhaas, F. M. H., & Hewstone, M. (2017). Reducing intergroup bias through intergroup contact: Twenty years of progress and future directions. *Group Processes & Intergroup Relations*, 20(5), 606-620.
- Eraslan, L. (2012). Evaluation of today's teacher unionism. *21. Yüzyılda Eğitim ve Toplum*, 1(1), 59-72. <https://dergipark.org.tr/tr/download/article-file/59603>
- Erdem M., & Meriç, E. (2012). Study on scale development about favoritism at school administration. *Journal of Educational Sciences Research*, 2(2), 141-154. <https://dergipark.org.tr/tr/download/article-file/697379>
- Erdem M., & Meriç, E. (2013). According to the perceptions of primary school teachers favoritism on school management. *Educational Administration: Theory and Practice*, 19(3), 467-498. <https://dergipark.org.tr/tr/download/article-file/108150>
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18, 39-50.
- Göregenli, M. (2012). *Temel kavramlar: Önyargı, kalıpyargı ve ayrımcılık [Basic concepts: Prejudice, stereotype and discrimination]*. Çayır, K. & Ceyhan, M. A. (derl.), *Ayrımcılık çok boyutlu yaklaşımlar [Multidimensional approaches to discrimination]*, İstanbul Bilgi Üniversitesi Yayınları.
- Guthery, S. (2018). The influence of teacher unionization on educational outcomes: A summarization of the research, Popular methodologies and gaps in the literature. *The William & Mary Educational Review*, 5(1), 124-136.
- Güneş, H. (2013). Sendikal haklar ve ülkemizde kamu görevlileri sendikacılığının gelişimi [Right to unionization and development of public servants trade unionism in Turkey]. *ÇSGB Çalışma Dünyası Dergisi [Labour World]*, 1(1), 62-79.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis (7th ed.)*. Pearson Prentice Hall.
- Hair, J. Hult, GTM. Ringle, C., & Sarstedt M. (2014). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)* (SAGE Publications, Incorporated).
- Hasta, D., & Arslantürk, G. (2013). Polislik mesleğine yönelik iç-grup yanlılığı ve tutumlar [Ingroup bias and attitudes towards the policing profession]. *Türk Psikoloji Yazıları*, 16(31), 60-70. <https://www.psikolog.org.tr/tr/yayinlar/dergiler/1031828/tpy1301996120130000m000094.pdf>

- Hooper, D., Coughlan, J., & Mullen, M. (2008). Evaluating model fit: A synthesis of the structural equation modelling literature. In *7th European Conference on research methodology for business and management studies* (pp. 195-200).
- Joreskog, K.G. (1999). How large can a standardized coefficient be. Unpublished report. SSI Central, Inc. <http://www.statmodel.com/download/Joreskog.pdf>
- Kağıtçıbaşı, Ç. (2008). *Günümüzde insan ve İnsanlar Sosyal Psikolojiye Giriş [Man and People Today. Introduction to Social Psychology]*. 11. Basım, Evrim Yayınevi.
- Kalaycı, Ş. (2010). *SPSS uygulamalı çok değişkenli istatistik teknikleri [SPSS applied multivariate statistical techniques]* (Vol. 5). Ankara, Turkey: Asil Yayın Dağıtım.
- Kara, M. (2016). The reasons why teachers do not affiliate to unions and their expectations from them. *The Journal of Academic Social Science*, 4(22), 423-440.
- Karaman, H. G., & Erdoğan, Ç. (2016). An investigation of the education unions in Turkey: goals, expectations and problems. *Sakarya University Journal of Education*, 6(2), 123-140. <https://dergipark.org.tr/tr/download/article-file/227590>
- Kayıkçı, K. (2013). Unionization in the public and education sector in Turkey, and Expectations of School administrators and teachers expectations from unions. *Amme İdaresi Dergisi*, 46(1), 99-126.
- Keskinkılıç-Kara, (2016). Individual and organizational effects of political orientation discrimination on teachers in schools. *Kastamonu Eğitim Dergisi*, 24(3), 1371-1384. <https://dergipark.org.tr/tr/download/article-file/210080>
- Keskinkılıç-Kara, S. B., & Oğuz, E. (2016). Relationship between political discrimination level perceived by teachers and teachers' organizational cynicism levels. *Eurasian Journal of Educational Research*, 63, 55-70.
- Kline, P. (1994). *An Easy Guide to Factor Analysis*. Routledge.
- Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage Publications.
- Kostakoğlu, G. (2010). *Grup kimliğine yönelik tehdit ile iç-grup yanlılığının benlik değeri üzerindeki etkileri [The Effects of Threat to Group Identity and In-Group Bias on Self-Esteem]* [Unpublished master's thesis]. Hacettepe University.
- Küçükkömürler, S., & Sakallı-Uğurlu, N. (2017). Social contact theories to regulate intergroup relations: intergroup, extended, imagined contact. *Nesne Psikoloji Dergisi (NPD)*, 5(9), 1-31. <https://www.nesnedergisi.com/makale/pdf/1466592418.pdf>
- Mert, Ö. (2013). *Organization activities and intellectual actions of teachers in Turkey(1960-1980)* [Unpublished Doctoral Dissertation]. Süleyman Demirel University.
- Mertler, C. A., & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation*, (3rd ed.). Edition Taylor & Francis.
- Myers, D. G. (2015). *Sosyal Psikoloji [Social Psychology]*. Onuncu basımdan çeviri, Nobel.
- Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory*. (3rd ed.). McGraw Hill.
- Özkiraz, A., & Talu, N. (2008). Emergence of trade-unions: comparison between Turkey and Western European Countries. *Sosyal Bilimler Araştırmaları Dergisi*. 2, 108-126. <https://dergipark.org.tr/tr/download/article-file/801854>
- Pettigrew, T. F., & Tropp, L. R. (2006). Interpersonal relations and group processes A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, 90(5), 751-783.
- Pettigrew, T. F. (2016). In pursuit of three theories: Authoritarianism, relative deprivation, and intergroup contact. *Annual Review of Psychology*, 67, 1-21.
- Polat, S., & Kazak, E. (2014). The correlation between school principals' favoritist behaviors and attitudes and teachers' perception of organizational justice. *Educational Administration: Theory and Practice*, 20(1), 71-92. <https://dergipark.org.tr/tr/download/article-file/108137>

- Polat, S., & Hiçyılmaz, G. (2017). Discrimination behaviors that classroom teachers are exposed and the causes of these behaviors. *Journal of Qualitative Research in Education*, 5(2), 47-66. <http://enadonline.com/public/assets/catalogs/0838411001544185376.pdf>
- Popiel, K. (2013). Teacher union legitimacy: Shifting the moral center for member engagement. *J Educ Change*, 14, 465–500.
- 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.
- Seat, A. A., Joyce, N., Harwood, J., & Arroyo, A. (2015). Necessary and sufficient conditions for positive intergroup contact: A fuzzy set qualitative comparative analysis approach to understanding intergroup attitudes. *Communication Quarterly*, 63(2), 135–155.
- Seçer, B. (2009). The effect of women's union attitudes and perceived gender discrimination on the willingness to join a union. *Çalışma ve Toplum*, 4, 27-60. <https://calismatoplum.org/Content/pdf/calisma-toplum-1317-fcee87fe.pdf>
- Sokolov, B. (2019). *Sensitivity of Goodness of Fit Indices to Lack of Measurement Invariance with Categorical Indicators and Many Groups*. Higher School of Economics Research Paper No. WP BRP, 86.
- Sürekli, D. (1998). *Kimya sektörüne bağlı bir sendikada sendikaya bağlılık ve sendikal faaliyetlere katılım arasındaki ilişkinin incelenmesi [Investigation of the relationship between union commitment and participation in union activities in a union affiliated to the chemical industry]* [Unpublished Doctoral Dissertation]. Marmara University.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics*. Person Education Inc.
- Tajfel, H., Billig, M., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behavior. *European Journal of Social Psychology*, 1, 149-177. <https://doi.org/10.1002/ej.sp.2420010202>
- Tavşancıl, E. (2014). *Tutumların ölçülmesi ve SPSS ile veri analizi [Measuring attitudes and data analysis with SPSS]*. (5.basım). Nobel Yayıncılık.
- Taylor, S. E., Peplau, L. A., & Sears, D. O. (2007). *Sosyal Psikolojiye Giriş [Introduction to Social Psychology]*. (1. Baskı). Çev.: Ali Dönmez, İmge Kitabevi Yayınları.
- TDK. *Güncel Türkçe Sözlük [Current Turkish Dictionary]*. www.tdk.gov.tr (Erişim tarihi: 20 Eylül 2019).
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1-10. <https://doi.org/10.1007/BF02291170>
- Uyumaz, G., & Çokluk, Ö. (2016). An investigation of item order and rating differences in likert-type scales in terms of psychometric properties and attitudes of respondents. *Journal of Theoretical Educational Science*, 9(3), 400-425. <https://dergipark.org.tr/tr/download/article-file/304330>
- Uyumaz, G., & Sırgancı, G. (2020). What is the required sample size for confirmatory factor analysis?: Bayesian Approach and maximum likelihood estimation. *International Journal of Society Researches*, 16(32), 5302-5340. <https://dergipark.org.tr/tr/download/article-file/1400546>
- Widaman, K. F., & Reise, S. P. (1997). *Exploring the measurement invariance of psychological instruments: Applications in the substance use domain*. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (p. 281–324). American Psychological Association. <https://doi.org/10.1037/10222-009>
- Yalçın Durmuş, G. (2018). *Öğretmenlerin sendikal örgütlenmelerinin okulda karar verme ve iletişim süreçleriyle ilişkisinin analizi [Analysis of the correlation between unionization of teachers and decision making and communication processes at the school]* [Unpublished Doctoral Dissertation]. İnönü University.

Yurdigül, Y., & İspir, N. (2015). *Ötekinin inşa edildiği sorunlu bir alan olarak Oscar ödül törenleri (85. Akademi Ödülleri ve “Argo” Filmi Örneği)* [Other being built, as a problem area: Oscar award ceremonies (85th Academy Awards and the “Argo” Movie Example)]. 1. Ulusal Toplumsal ve Kurumsal Çatışmalar/Çözümler Kongresi. Bildiriler Kitabı, Düzce Üniversitesi Gölyaka Meslek Yüksekokulu, Düzce.

6. APPENDIX

Table A displays the 27 items of standardized English and Turkish version of Union Bias Scale.

Table A. Union Bias Scale (Standardized English and Turkish version)

Bu ölçek, sendikal yanlılığı belirlemeye yönelik maddelerden oluşmaktadır. Maddelerin her birini okuyarak, “Hiç Katılmıyorum, Katılmıyorum, Çok Az Katılıyorum, Oldukça Katılıyorum, Tamamen Katılıyorum” seçeneklerinden birini işaretleyiniz. Maddelerin tümünü içtenlikle işaretlemeniz araştırma sonuçları açısından oldukça önemlidir. <i>Teşekkür ederim.</i>		Strongly disagree	Disagree	Slightly agree	Quite agree	Strongly agree
In Group Favoritism/ İç Grubu Kayırma	1. People from my own union always take precedence over those from other unions. 1. Kendi sendikamdan olanlar, her zaman diğer sendikalardan olanlardan önce gelir.	①	②	③	④	⑤
	2. It makes me happy that someone from my union is chosen to distribute the tasks. 2. Görevlerin dağıtımında kendi sendikamdan olan birinin seçilmesi beni mutlu eder.	①	②	③	④	⑤
	3. It makes me happy that someone from my union is chosen to distribute the awards. 3. Ödüllerin dağıtımında kendi sendikamdan olan birinin seçilmesi beni mutlu eder.	①	②	③	④	⑤
	4. I prefer to work with a manager from my own union. 4. Kendi sendikamdan olan bir yöneticiyle çalışmayı tercih ederim.	①	②	③	④	⑤
Out Group Disdain Dış Grubu Küçümseme	5. The activities of other unions are very ineffective and weak. 5. Diğer sendikalardan olanların faaliyetleri oldukça etkisiz ve zayıftır.	①	②	③	④	⑤
	6. The ideological approaches of other unions cause an atmosphere of unrest. 6. Diğer sendikaların ideolojik yaklaşımları huzursuzluk ortamına sebep olmaktadır.	①	②	③	④	⑤
	7. The activities of other unions are segregating for employees. 7. Diğer sendikaların faaliyetleri, çalışanları ayırıştırıcıdır.	①	②	③	④	⑤
	8. Those from other unions cause conflicts in our institution. 8. Diğer sendikalardan olanlar kurumumuzda çatışma ortamına sebep olmaktadır.	①	②	③	④	⑤
	9. Those from other unions serve ideological purposes rather than union goals. 9. Diğer sendikalardan olanlar, sendikal amaçlardan çok ideolojik amaçlara hizmet etmektedirler.	①	②	③	④	⑤
In Group Glorification İç Grubu Yüceltme	10. The union I am a member of is unique. 10. Üyesi olduğum sendika eşsizdir.	①	②	③	④	⑤
	11. The union I am a member of always makes the best decisions. 11. Üyesi olduğum sendika her zaman en iyi kararları alır.	①	②	③	④	⑤
	12. The union I am a member of is the one that defends our rights most effectively. 12. Üyesi olduğum sendika, haklarımızı en etkili savunan sendikadır.	①	②	③	④	⑤
	13. The aims of the union I am a member of are higher than the aims of other unions. 13. Üyesi olduğum sendikanın amaçları diğer sendikaların amaçlarından yüksektir.	①	②	③	④	⑤
	14. The union I am a member of is not a supporter of a political party.	①	②	③	④	⑤

	14. Üyesi olduğum sendika bir siyasi partinin arka bahçesi değildir.					
	15. The activities of the union that I am a member of are unifying the employees.	①	②	③	④	⑤
	15. Üyesi olduğum sendikanın faaliyetleri çalışanları birleştiricidir.					
Social Distance Sosyal Mesafe	16. At my institution, my relations with employees from other unions are not good.	①	②	③	④	⑤
	16. Kurumumda, diğer sendikalardan olan çalışanlarla ilişkilerim soğuktur.					
	17. In my institution, my professional communication with employees from other unions is poor.	①	②	③	④	⑤
	17. Kurumumda, diğer sendikalardan olan çalışanlarla mesleki iletişimim zayıftır.					
	18. In my institution, I have poor social communication with employees from other unions.	①	②	③	④	⑤
	18. Kurumumda, diğer sendikalardan olan çalışanlarla sosyal iletişimim zayıftır.					
Prejudices Önyargılar	19. Those from other unions have a grudge and hatred towards me/us.	①	②	③	④	⑤
	19. Diğer sendikalardan olanlar bana/bize karşı kin ve nefret içindedirler.					
	20. Although it is in favor of the employees, those from the other union do not support our union activities.	①	②	③	④	⑤
	20. Çalışanların lehine olsa da diğer sendikadan olanlar sendikal faaliyetlerimizi desteklemezler.					
	21. It is not possible for those from other unions to cooperate with us on union matters.	①	②	③	④	⑤
	21. Diğer sendikalardan olanların sendikal konularda bizimle işbirliği yapması mümkün değildir.					
	22. It is not possible for all unions to unite around the same union goals.	①	②	③	④	⑤
	22. Tüm sendikaların aynı sendikal amaçlar etrafında birleşmesi mümkün değildir.					
Stereotypes Kalıpyargılar	23. Among other unions there are unions that support the government in power.	①	②	③	④	⑤
	23. Diğer sendikalar arasında “İktidar yanlısı sendika” vardır.					
	24. Among other unions there are pro-terror unions.	①	②	③	④	⑤
	24. Diğer sendikalar arasında “Terör yanlısı sendika” vardır.					
	25. Among other unions there are nationalist unions.	①	②	③	④	⑤
	25. Diğer sendikalar arasında “Ulusalcı sendika” vardır.					
	26. Among other unions there are fascist unions.	①	②	③	④	⑤
	26. Diğer sendikalar arasında “Faşist sendika” vardır.					
	27. Among other unions there are collaborator unions.	①	②	③	④	⑤
	27. Diğer sendikalar arasında “İşbirlikçi sendika” vardır.					

Adaptation of the Adlerian Personality Priority Assessment into Turkish

Abdi Gungor ^{1,*}, Dalena Dillman Taylor ²

¹Guidance and Psychological Counseling Program, Düzce University, Düzce, Turkey

²Department of Counselor Education and School Psychology, University of Central Florida, Orlando, FL, USA

ARTICLE HISTORY

Received: Jan. 18, 2021

Revised: Sep. 14, 2021

Accepted: Oct. 25, 2021

Keywords:

Adlerian therapy,

The Adlerian personality
priority assessment,

Adaptation,

Psychometric properties,

Reliability,

Validity.

Abstract: Personality priorities are important concepts in Adlerian theory, contributing to understanding and conceptualizing clients' lifestyles. Even though Adlerian psychology promises multicultural applications and has been interested in Turkey, no instrument measuring personality priorities has been developed or adapted into Turkish. Therefore, the purpose of this study was to adapt the Adlerian Personality Priority Assessment (APPA) into Turkish and examine its psychometric properties with a sample of Turkish undergraduate students. This study was conducted in three steps. In the first step, a linguistic equivalence test was performed with a sample of 73 students enrolled at the Department of English Language Education. In the second step, the structure of the APPA was examined using exploratory and confirmatory factor analyses with a sample of 1,279 undergraduate students. In the final step, test-retest reliability was tested with a sample of 93 undergraduate students within 4-week interval. The results of the linguistic equivalency study revealed that translations were linguistically and culturally proper. According to the exploratory and confirmatory factor analyses, the Turkish form of the APPA consisted of 24 items loaded with four factors consistent with the original form. The results also revealed good levels of internal and test-retest reliabilities. The findings of this study showed that the Turkish form of the APPA is a valid and reliable instrument, and it can be used in research and practice with Turkish populations. The results and limitations were discussed, along with implications for future research and practice.

1. INTRODUCTION

Adlerian theorists strongly consider individuals' social and cultural contexts when conceptualizing people within their social environments (Carlson & Carlson, 2000). Individual psychology, created by Alfred Adler (1931), appears to be suitable for the characteristics of Turkish culture. For instance, Turkish culture values cooperation, connection with others, and social life, and the roles of family in early childhood are also highly emphasized in Turkish traditions (Sümer & Rasmussen, 2012). In this regard, Adlerian psychology receives attention in Turkey for both practice and research. However, research within Turkish psychology practice appears to be in its infancy. For example, Adlerian concepts were conceptually discussed and reviewed in Turkish literature such as Adlerian encouragement within the counseling relationship (Ergün-Başak & Ceyhan, 2011) and the use of Adlerian family counseling

*CONTACT: Abdi Gungor ✉ abdigungor@duzce.edu.tr 📍 Düzce University, Faculty of Education, No: 311, Merkez/Duzce 81620

(Akçabozan & Sümer, 2016). Those studies discussed the Adlerian concepts and provided suggestions for Adlerian practice in Turkish culture. Given the heightened interest in adopting Adlerian psychology in the Turkish culture (Sümer & Rasmussen, 2012), it is crucial to develop valid and reliable instruments that measure the essential concepts of Adlerian theory, such as personality priorities.

1.1. Individual Psychology

Across all theoretical approaches, therapists work to conceptualize clients' presenting concerns and develop a treatment plan to intervene to best help clients overcome challenges. Within the Adlerian framework, the concept of understanding a client's presenting issue is positioned within their lifestyle: their view of self, others, and the world as influenced by genetic and environmental factors (Carlson et al., 2006). Although Adler (1929) stated, "we do not consider human beings types because every human being has an individual style of life" (p.102), he also noted that lifestyles serve as an intellectual device to understand similarities and differences between people. Even though Adlerian therapists prize each client's uniqueness and personal assets, understanding an overarching framework of individuals' lifestyles can be clinically useful (Kefir & Corsini, 1974). In this regard, Kefir (1971) developed four personality priorities that capture broad categories of individuals' worldviews that can quickly provide the therapist insight into their lifestyle, mistaken beliefs/cognitive distortions, and how that worldview can impact their progress in therapy.

1.2. Personality Priorities

Kefir (1971) once described personality priorities as a "window into one's lifestyle," indicating this construct can be used as a snapshot into an individual's view of self, others, and the world. The brief snapshot can provide therapists with opportunities to present tentative hypotheses early on in the counseling process to begin deconstructing mistaken beliefs and developing more positive coping skills. However, Kefir and Corsini (1974) clarified that priorities are not considered fixed, providing fluidity for modifications in different situations (e.g., personal versus professional life). Priorities instead give insights on one's general dispositional set and central tendency towards life (Kefir & Corsini, 1974). Each person functions from a primary priority, although they can constantly access all four as needed to strive for belonging and significance. Dependent upon early childhood experiences, individuals may choose to operate from their primary priority on the socially useful or useless side of life. This functionality allows the therapist to understand the level of discouragement or distress in which the client is currently presenting. Therefore, there is no hierarchy of priorities; each priority serves a purpose to enable the client to achieve their primary goals best.

Kefir (1981) originally defined four types of personality priorities: (a) avoider, (b) pleaser, (c) controller, and (d) morally superior. However, Pew (1976) modified the priorities as follows: control, pleasing, superiority, and comfort, removing the person's focus and replacing it with an action. This slight modification aligned more with the purpose of fluidity and appeared less trait-like. Due to page limitations, see Dillman Taylor et al. (2015) for complete definitions of the four priorities.

1.3. Measuring Personality Priorities

Dillman Taylor et al. (2015) developed the Adlerian Personality Priority Assessment (APPA) in response to the lack of validation of previous personality priority instruments (e.g., Allen Assessment for Adlerian Personality Priorities [AAAPP]; Langenfeld Inventory for Personality Priorities [LIPP]). The APPA was created to assess individuals' priority of achieving significance and belonging in their lives. The original study confirmed Kefir's (1971) four initial priorities with 393 undergraduates, concluding 30 items represented four personality priorities: control (six items), pleasing (nine items), superiority (seven items), and comfort

(seven items). Dillman Taylor, Bratton, and Henson (2019) conducted another study to examine the psychometric properties of the APPA with a sample of 1201 undergraduate students. Results supported the four-factor structure of the APPA and provided preliminary results for the usefulness of the APPA for research and practice. For example, in a study utilizing the APPA to measure personality priorities, Dillman Taylor et al. (2018) examined the relationship between personality priorities and wellness in counselors-in-training and found that pleasing and comfort negatively predicted wellness even though superiority and control were not found significantly related to wellness. In addition, Dillman Taylor and Mullen (2019) modified the 30-item APPA to 22 items, resulting in a four-factor model: control (four items), pleasing (seven items), superiority (seven items), and comfort (four items). All studies examining the factorial structure of the APPA confirm four factors, which align with the original theoretical structure of personality priorities. Therefore, preliminary evidence of the internal structure of the APPA, using EFAs and confirmatory factor analyses across studies, was demonstrated in addition to evidence of relationship to other variables (e.g., wellness, Dillman Taylor et al., 2018).

Even though the studies on Adlerian concepts in Turkish culture were limited, there is a growing interest (e.g., Akçabozan & Sümer, 2016; Ergün-Başak & Ceyhan, 2011; Sümer & Rasmussen, 2012). As discussed, understanding personality priorities contribute to conceptualizing the person's lifestyle (Kefir, 1971; Kefir & Corsini, 1974). To date, an instrument assessing personality priority in Turkey does not exist. However, validated instruments are needed to carry out further research on Adlerian theory and practice. Thus, the purpose of this study was to adapt the APPA to Turkish culture, specifically focusing on the internal structure. More specifically, this study aimed to examine the reliability and validity of the APPA with Turkish undergraduate students.

2. METHOD

2.1. Data Collection

A convenient sampling method was used in this study. We received approval from the university ethical board in Turkey to conduct this study. Based on outlined data collection procedures, we followed university protocol to obtain permission from dean offices and class instructors to recruit participants. All data were collected in person via paper-and-pencil. We followed detailed procedures to ensure confidentiality and that participation was voluntary. For this study, we conducted a three-part process in data collection. For the first part of the study, 73 participants completed both the English and Turkish forms of the APPA for the linguistic equivalence within a 2-week interval. Based on the results of the linguistic equivalence test, translations of four items were revised. Once the translation process was finalized, we collected data from the second group ($n = 1279$). Finally, the Turkish form of the APPA was implemented twice to the last group ($n = 93$) for test-retest reliability within a 4-week interval.

2.2. Participants

A total of 1445 undergraduate students who attended a public university in the northwest of Turkey participated in this study. This study consisted of three independent samples. In the first sample, 73 students enrolled at the Department of English Language Education participated in the linguistic equivalence study. To qualify for part one of this study, the participants needed to demonstrate fluency in both Turkish and English. This qualification was to ensure the accuracy of responses to items of the English and the Turkish version of the APPA.

In the second sample, 1279 undergraduate students, whose age ranged from 17 to 38 ($M = 20.81$, $SD = 2.46$) participated. To qualify for part two of this study, the participants attended a four-year undergraduate program. The data of this group was randomly split into two equal subsamples ($n_1 = n_2 = 639$); one case was randomly removed to ensure equal sample sizes. Age

ranged from 17 to 38 ($M = 20.79$, $SD = 2.43$) in the subsample 1, and from 17 to 38 ($M = 20.83$, $SD = 2.49$) in the subsample 2. Table 1 shows demographic characteristics of the participants, such as gender, grade, and birth order for the total sample for part two, subsample 1, and subsample 2. In addition, we asked participants to rate their perceived socioeconomic status. As shown in Table 1, the total sample for part two of this study mostly reflected middle-class socioeconomic status.

Table 1. Comparison Demographics of total group subsample groups

Demographic	Part II Total Sample <i>n</i> (%)	Subsample 1 <i>n</i> (%)	Subsample 2 <i>n</i> (%)
Gender			
Male	305 (23.8)	153 (23.9)	152 (23.8)
Female	958 (74.9)	477 (74.7)	480 (75.1)
Missing	16 (1.3)	9 (1.4)	7 (1.1)
Grade			
Freshman	449 (35.1)	216 (33.8)	233 (36.5)
Sophomore	214 (16.7)	116 (18.2)	98 (15.3)
Junior	236 (18.5)	116 (18.2)	120 (18.8)
Senior	375 (29.3)	188 (29.4)	186 (29.1)
Missing	5 (.4)	3 (.4)	2 (.3)
Birth order			
First	460 (36)	214 (33.5)	245 (38.3)
Second/middle	526 (41.1)	264 (41.3)	262 (41)
Last	246 (19.2)	135 (21.1)	111 (17.4)
Only	32 (2.5)	17 (2.7)	15 (2.4)
Missing	15 (1.2)	9 (1.4)	6 (.9)
Perceived economic status			
High	58 (4.5)	33 (5.2)	25 (3.9)
Middle	1147 (89.7)	563 (88.1)	583 (91.2)
Low	59 (4.6)	33 (5.2)	26 (4.1)
Missing	15 (1.2)	10 (1.5)	5 (.8)

For the test-retest reliability study (part 3), 93 participants (17 males, 76 females) completed the APPA. Ages in this group ranged from 20 to 35 ($M = 23.17$, $SD = 2.81$). To qualify, participants in this group needed to be undergraduate students who attended a four-year program.

2.3. Measurements

2.3.1. The Adlerian Personality Priority Assessment (APPA)

Dillman Taylor et al. (2015) developed the APPA to measure Adlerian personality priorities as a mechanism to assess mistaken beliefs or cognitive distortions, which were originally proposed by Kefir (1971). The APPA has 30 items measuring four personality priorities: control (six items), pleasing (nine items), superiority (seven items), and comfort (seven items). Items are on a 5-point Likert scale ranging from not at all (1) to very much (5). A higher score of a particular personality priority indicates that a person views the world more in line with the characteristics of that priority to achieve significance and belonging in their life. Sample items include “In most situations, I prefer to be in charge” (control), “I need to know that others are pleased with me” (pleasing), “I need to be the winner in games” (superiority), and “I prefer not having a lot of work to do” (comfort).

Dillman Taylor et al. (2015) reported that the full model reproduced 47.29% of the variance. Each factor explained for each factor was 16.93% for pleasing, 10.57% for control, 10.13% for comfort, and 9.66% for superiority. In addition, the following studies confirmed the four-factor structure of the APPA with samples of 1,210 undergraduate students (Dillman Taylor et al., 2019) and 1,019 adults (Dillman Taylor & Mullen, 2019). Dillman Taylor et al. (2015) reported Cronbach Alpha coefficients for each factor as follows: .91 for pleasing, .80 for comfort, .81 for control, and .88 for superiority. The current study reported Cronbach Alpha coefficients for each factor as following .84 for superiority, .81 for pleasing, .68 for comfort, and .78 for control.

2.3.2. Translation procedure

We conducted the translation process of the APPA into Turkish using the five-step model as suggested to adapt an original instrument into another language and culture (e.g., Abubakar et al., 2013; Carlson, 2000). First, four faculty members in counseling departments who were native Turkish speakers and had fluent English separately translated the items of the APPA into Turkish. Second, all translations were compared and analyzed, and the most accurate translations were chosen for each item. Third, a faculty member in Teaching English to Speakers of Other Languages department in the United States, who was a native Turkish speaker and unfamiliar to the APPA, reverse translated the items into English. Previous researchers have suggested reverse translation as a part of the adaption procedure to ensure the correctness of the translation (Abubakar et al., 2013; Geisinger, 1994). Fourth, two faculty members in counselor education in the United States, who were highly familiar with Adlerian theory, compared the back-translated items with the original APPA in terms of correctness, clarity, and cultural relevancy. Finally, the reverse-translated version of the APPA was reviewed and approved by the developer of the original form of the APPA. In line with the original scale, the final version of the APPA-Turkish (APPA-T) contains 30 items on a 5-point Likert scale (1 = *Not at all*, 2 = *A little bit*, 3 = *Somewhat*, 4 = *Quite a bit*, and 5 = *Very much*). In the final step, a translation equivalency test was implemented with 73 undergraduate students enrolled at the Department of English Language Education in Turkey. Based on the translation equivalency test results, two faculty members in counselor education revised four items, thus finalizing the APPA-T.

2.4. Data Analysis

We conducted analyses using Statistical Package for the Social Sciences (SPSS) version 22.0 and Analysis of Moment Structures (AMOS) version 23. Before data analysis, a data screening procedure was employed. More specifically, encoding data, missing data, and outliers were checked. In addition, we examined normality, multicollinearity, and missing data. Results are presented in the following sections. In study 1, we employed a translation equivalency study to investigate the accuracy of the translation.

In study 2, we examined the structure of the original APPA using confirmatory factor analysis (CFA), but the results revealed a poor-fitting model. Thus, we tested the dimensionality of the APPA with this Turkish sample. As recommended for cross-validation (Gerbing & Hamilton, 1996), we randomly split the sample ($N = 1279$) into two subsamples ($n_1 = 639$, $n_2 = 639$) using the random sample selection procedure in SPSS 22.0, and randomly deleted one participant to obtain equal samples. Then, we conducted exploratory factor analysis (EFA) to determine the factor structure of the APPA-T with subsample 1. Next, a CFA was conducted with subsample 2 to confirm the structure acquired as the result of the EFA. Cronbach's alpha coefficients of each factor were estimated with the whole sample to examine the internal consistency reliability of the APPA-T. In study 3, we conducted a test-retest study with 93 students within a 4-week interval.

3. RESULT / FINDINGS

Regarding preliminary analyses, we checked all assumptions for the data analyses employed across the three studies. First, we employed Little's MCAR tests to examine whether the data sets were completely at random. For all three subsamples, data were found to be at random (Study 1, $\chi^2 = 857.03$, $df = 878$, $p > .05$; Study 2, $\chi^2 = 1027.99$, $df = 968$, $p > .05$; Study 3, $\chi^2 = 479.18$, $df = 472$, $p > .05$). We checked all assumptions, including z scores for potential outliers, kurtosis and skewness for normality, and multicollinearity (Tabachnick & Fidell, 2013). For all three studies, assumptions were met, thus demonstrating that the data is appropriate for the selected analyses.

3.1. Study 1: Translation Equivalency Test

In an adaption of an instrument into another language and culture, Carlson (2000) suggested three steps for the translation process: (1) one-way translation is conducted by bilingual experts translating the original instrument into the target language; (2) back-translation method is performed as an independent bilingual expert translates back the target-language version into the original language; and (3) a translation equivalency test is conducted to ensure a culturally equivalent translation. In preparation for this study, we conducted the first two steps prior to implementing with participants, as noted previously. Therefore, we asked 73 bilingual undergraduate students to complete the original 30-item APPA first then the APPA-T within a 2-week interval. The main purpose of an equivalency test is to utilize two forms of an instrument into the same group and compare the results. If the translations are deemed accurate and meaningful, no difference between the two implementations is expected (Carlson, 2000; Hambleton et al., 2004). Thus, we conducted paired-samples t-tests to compare the participants' scores on the two different versions of the APPA (Pallant, 2010). The results showed that there were no statistically significant differences in mean scores of two implementations for each of the priorities: superiority ($t(51) = .29$, $p = .77$), pleasing ($t(51) = -1.46$, $p = .15$), comfort ($t(51) = .55$, $p = .59$), and control ($t(51) = .99$, $p = .32$). Therefore, these findings supported the translation equivalency between the APPA and the APPA-T.

For further analysis, we inspected the results of both assessments (e.g., APPA and APPA-T) to ensure the accuracy of translations. Due to the ordinal nature of the items ranging from one to five, we conducted the Wilcoxon matched pairs signed ranks test for each item across the two assessments (Pallant, 2010; Tabachnick & Fidell, 2013). The results indicated only four items (APPA Item 5, 8, 18, and 24) had statistically significant differences between scores across the two instruments. These findings support the accuracy of the translation for the majority of the items. Further, we examined correlations between the items in English and Turkish using the Spearman rho formula (Pallant, 2010; Tabachnick & Fidell, 2013). All results, including mean ranks for each item, are reported in [Table 2](#). The results revealed that most correlations between English and Turkish items were statistically significant, except for three items (APPA Item 2, 17, and 28). Therefore, as suggested (Carlson, 2000), we inspected these seven items; two Turkish faculty revised these items in terms of clarity and cultural relevancy.

Table 2. Wilcoxon Matched Pairs Signed Ranks Test, Correlations, Means, and Standard Deviations for Each Item.

Items	APPA	APPA-T	<i>rho</i>	<i>Z</i>
	<i>M (SD)</i>	<i>M (SD)</i>		
APPA Item 1	2.81 (1.10)	2.72 (.98)	.31*	-.99
APPA Item 2	2.79 (1.27)	2.36 (.97)	.24	-1.44
APPA Item 3	3.33 (1.31)	3.70 (1.15)	.51**	-1.19
APPA Item 4	3.12 (1.14)	3.35 (.95)	.56**	-1.81
APPA Item 5	2.67 (1.38)	2.37 (1.33)	.69**	-2.74**
APPA Item 6	2.77 (1.16)	3.10 (1.13)	.60**	-.31
APPA Item 7	2.64 (1.23)	2.68 (.99)	.43**	-.01
APPA Item 8	2.52 (1.31)	3.02 (1.46)	.59**	-2.37*
APPA Item 9	3.32 (1.23)	3.36 (1.10)	.39**	-2.35
APPA Item 10	3.07 (1.36)	3.40 (1.06)	.46**	-1.87
APPA Item 11	2.68 (1.33)	2.53 (1.24)	.55**	-1.35
APPA Item 12	3.42 (1.09)	3.66 (.94)	.31*	-.97
APPA Item 13	3.41 (1.07)	3.47 (.87)	.36**	-.06
APPA Item 14	3.31 (1.29)	3.38 (1.14)	.74**	-.36
APPA Item 15	3.35 (1.21)	3.39 (1.18)	.38**	-.89
APPA Item 16	3.34 (1.26)	3.17 (1.21)	.55**	-.85
APPA Item 17	3.04 (1.04)	3.42 (1.24)	.20	-1.83
APPA Item 18	2.13 (1.16)	1.69 (.93)	.46**	-2.41*
APPA Item 19	3.19 (1.31)	3.33 (1.14)	.51**	-.49
APPA Item 20	3.01 (1.25)	2.98 (1.29)	.63**	-.21
APPA Item 21	2.88 (1.26)	2.95 (1.23)	.65**	-.89
APPA Item 22	3.60 (1.28)	3.65 (1.10)	.60**	-.07
APPA Item 23	3.09 (1.80)	2.85 (1.18)	.45**	-1.87
APPA Item 24	3.07 (1.24)	2.42 (1.09)	.50**	-3.74***
APPA Item 25	3.11 (1.35)	3.02 (1.33)	.54**	-.69
APPA Item 26	3.07 (1.09)	3.30 (1.06)	.57**	-1.84
APPA Item 27	2.67 (1.13)	2.63 (1.26)	.61**	-.55
APPA Item 28	2.27 (1.07)	2.13 (.03)	.19	-1.27
APPA Item 29	2.97 (1.29)	2.98 (1.14)	.61**	-.68
APPA Item 30	3.21 (1.27)	3.12 (1.25)	.68**	-1.17

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

3.2. Study 2: Structure Analyses

3.2.1. Confirmatory factor analysis

We conducted a CFA to test the original structure of the APPA developed by Dillman Taylor et al. (2015) with the entire sample ($N = 1279$). To evaluate the model, we used the following fit indices to determine the overall goodness of the model: (a) chi-square ratio ($< .4$), (b) comparative fit index (CFI, $> .94$), (c) goodness of fit (GFI, $> .90$), root mean square error of approximation (RMSEA, $< .09$), and standardized root mean square residual (SRMR, $< .06$; Brown, 2014; Hooper et al., 2008; Hu & Bentler, 1999; Kline, 2011; O'Rourke et al., 2013; Weston & Gore, 2006). The CFA results indicated a poor fit for the original APPA: χ^2 ($df = 399$, $N = 1279$) = 2,501.01, $p < .001$; $\chi^2/df = 6.27$; CFI = .82; GFI = .87; RMSEA = .06; SRMR = .08. Therefore, this model was deemed to be insufficient for the current data. Because the original factor structure of the APPA was not supported with this Turkish sample, as suggested

(Gerbing & Hamilton, 1996), we reexamined the dimensionality of the APPA with this Turkish sample. Thus, we randomly split the data into two equal subsamples: Subsample 1 ($n_1 = 639$) and subsample 2 subsample ($n_2 = 639$).

3.2.2. Exploratory factor analysis

To determine the factor structure of the Turkish version APPA, with subsample 1 ($n_1 = 639$), we conducted a principal component analysis (PCA; Pallant, 2010) with direct varimax rotation because factors were not expected to be theoretically correlated (Costello & Osborne, 2005; Pallant, 2010). The item-to-case ratio was 1:21, ideal for this analysis (Costello & Osborne, 2005). In addition, regarding the suitability of the sample for factor analysis, we met the assumptions for Bartlett’s test of sphericity ($\chi^2 = 6253.34$, $df = 435$, $p < .001$) and Kaiser-Meyer-Olkin (KMO = .86; Hair et al., 2010; Mvududu & Sink, 2013). These results indicated that the data was appropriate for the use of factor analysis.

In order to determine the number of factors to retain, we inspected Eigenvalues higher than 1, the scree plot, and conducted parallel analysis (Pallant, 2010; Tabachnick & Fidell, 2013). The results of the principal component analysis with varimax rotation found six factors with Eigenvalues greater than 1. However, when we inspected the scree plot, a significant break occurred between the fourth and fifth factors, which suggested a four-factor structure. Results from the parallel analysis also suggested retaining four factors. Thus, we retained four factors for the initial factor analysis. One item demonstrating significant cross-loading ($>.32$; Costello & Osborne, 2005; Tabachnick & Fidell, 2013) was removed from the model. We reran the model to determine model fit with 29 items. Further, we found that items 15 and 18 on the APPA-T loaded on a different factor than anticipated (Dillman Taylor et al., 2015); however, we elected to retain it based on theoretical support.

The final model (see Table 3) comprised of a four-factor structure with 29 items, which were consistent with the original model: Factor 1 (superiority, nine items), Factor 2 (pleasing, nine items), Factor 3 (comfort, seven items), and Factor 4 (control, four items). The 4-factor model explained 48.21% of the total variance, appropriate for social sciences (Hair et al., 2010; Mvududu & Sink, 2013).

Table 3. Factor Pattern/Factor Loadings for Exploratory Factor Analysis with Varimax Rotation of APPA-T

Items	Factor 1 Superiority	Factor 2 Pleasing	Factor 3 Comfort	Factor 4 Control
APPA Item 20	.84	.15	-.13	.03
APPA Item 19	.81	.12	-.09	.06
APPA Item 30	.72	.04	-.03	.14
APPA Item 16	.71	.05	.14	.02
APPA Item 25	.70	.12	-.09	.15
APPA Item 14	.60	.20	-.20	.17
APPA Item 26	.56	.04	.31	.14
APPA Item 15	.55	.26	.21	.08
APPA Item 18	.42	-.10	.21	.10
APPA Item 27	.06	.74	.09	.16
APPA Item 17	.12	.70	-.09	.03
APPA Item 11	.10	.69	-.03	.18
APPA Item 21	.21	.69	.01	.02
APPA Item 3	-.13	.68	-.06	.03

Table 3. *Continues*

APPA Item 10	.27	.67	.09	.05
APPA Item 5	.03	.58	.08	-.14
APPA Item 8	-.02	.53	.18	.06
APPA Item 13	.23	.53	.17	.14
APPA Item 6	-.02	.00	.71	-.19
APPA Item 7	.03	.05	.61	-.21
APPA Item 29	.05	.14	.58	-.25
APPA Item 28	.07	.08	.56	.05
APPA Item 4	.00	.12	.54	-.32
APPA Item 22	.12	-.13	.45	.04
APPA Item 2	-.18	.18	.43	.00
APPA Item 23	.19	.13	-.13	.82
APPA Item 24	.24	-.02	-.04	.73
APPA Item 12	.14	.16	-.17	.71
APPA Item 9	.12	.19	-.29	.70
% variance	20.78	12.47	10.13	4.83
Eigenvalue	6.03	3.62	2.94	1.40

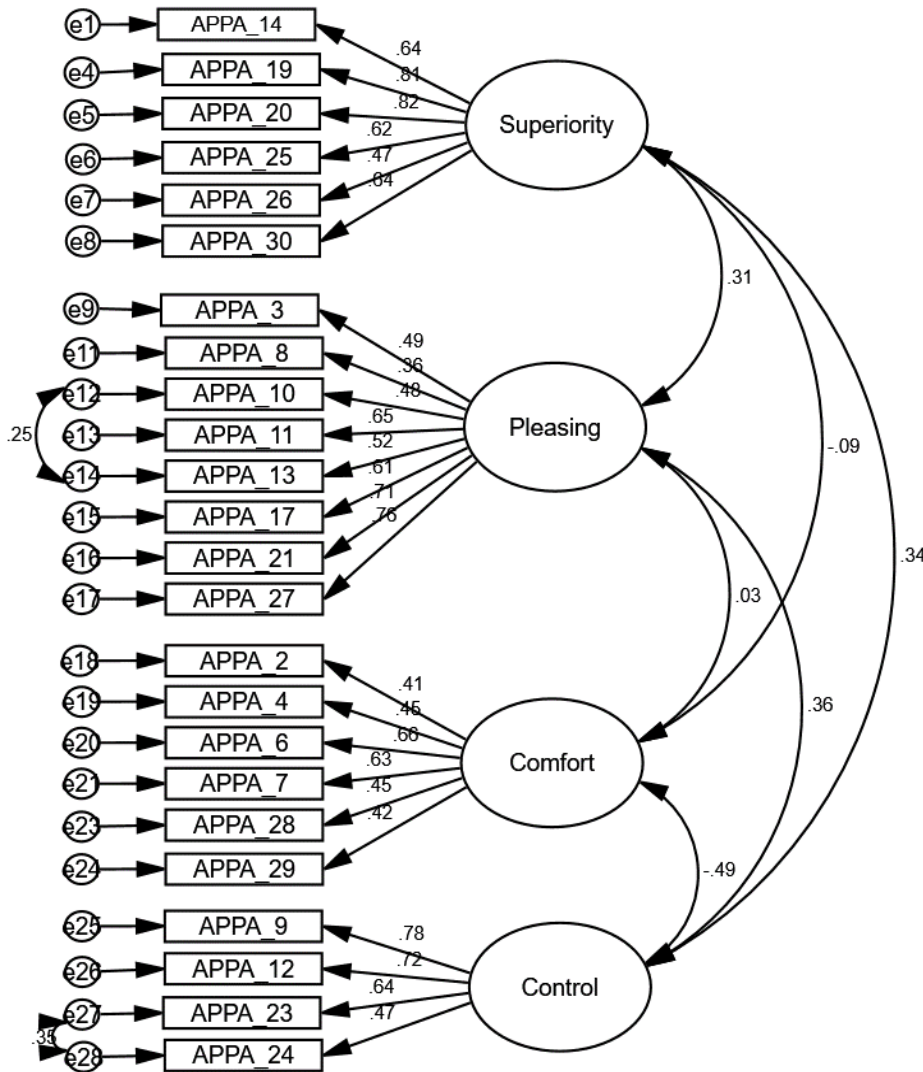
3.2.3. Second confirmatory factor analysis

The next step of the investigation of structural validity of the APPA-T was conducting a CFA with subsample 2 ($n_2 = 639$), a unique sample from subsample 1 used in the EFA. We conducted a CFA on the model derived as the result of the EFA, which indicated a four-factor structure with 29 APPA-T items. The model revealed a poor fit across some indices: $\chi^2 (df = 371, N = 639) = 1172.38, p < .001$; $\chi^2/df = 3.16$; CFI = .84; GFI = .88; RMSEA = .06; SRMR = .06. Fit indices produced mixed results, so we inspected the modification indices to improve the model.

Based on modification indices, we systematically removed APPA items 5, 15, and 16 due to covarying not being theoretically justified. In addition, we removed items 18 and 22 from the model because the loading of that items was lower than .30 (Mvududu & Sink, 2013). Due to theoretical support, we elected to free errors between APPA items 10-13 and 23-24. Items 10 and 13 relate to one's desire to know that others are pleased with them; whereas, items 23 and 24 convey an individual's preference of being in charge.

With these modifications, the final model resulted in a 24-item, four-factor structure, which also produced a mixed outcome with some indices demonstrating good fit while others show adequate fit: $\chi^2 (df = 244, N = 639) = 649.21, p < .001$; $\chi^2/df = 2.66$; CFI = .90; GFI = .92; RMSEA = .05; SRMR = .06. The standardized regression weights revealed that the loadings for all items were statistically significant, and ranged from .36 to .82, (Figure 1). In addition, the CFA revealed various relationships among priorities/factors. The superiority factor was found to be statistically correlated to pleasing ($r = .31, p < .001$) and control ($r = .34, p < .001$); and control was found to be statistically correlated to pleasing ($r = .36, p < .001$) and comfort ($r = -.49, p < .001$). Finally, we ran Cronbach's alpha internal consistency with the whole sample ($N = 1279$) for the final modified model derived from the second CFA and the found the following: .84 for superiority, .81 for pleasing, .69 for comfort, and .78 for control.

Figure 1. CFA Diagram of the Final Modified Model for the APPA-T.



3.3. Study 3: Test-Retest Reliability Analysis

To determine the test-retest reliability, we administered the APPA-T to 93 undergraduate students within 4-week interval. The correlation coefficients between two applications were all significant: ($r = .72, p < .01$) for superiority, ($r = .85, p < .01$) for pleasing, ($r = .73, p < .01$) for comfort, and ($r = .79, p < .01$) for control. Thus, those results showed that the APPA-T is a reliable instrument.

4. DISCUSSION and CONCLUSION

Personality priorities are important variables in Adlerian theory to understand individuals' worldviews and relate to others. The concept of personality priorities provides insight into individuals' goals in which they strive for significance and belonging (Dillman Taylor et al., 2015; Kefir, 1971; Kefir & Corsini, 1974). Kefir (1981) first introduced the concept of personality priorities. Shortly after that, Langenfeld and Main (1983) developed the first instrument, the LIPP, to measure personality priorities, although the items on the instrument presented with limited reliability (Ashby et al., 1998; Ashby et al., 2006; Dillman Taylor et al., 2015). Dillman Taylor et al. (2015) developed the APPA measuring personality priorities with a four-factor structure to address these concerns. Several studies demonstrate adequate to strong psychometric properties of the APPA (Dillman Taylor & Mullen 2019; Dillman Taylor et al.

2015; Dillman Taylor et al. 2019). Therefore, the researchers sought in this manuscript to adapt this instrument for the Turkish population.

We conducted three studies in order to establish the credibility of the revised instrument. First, we followed stringent guidelines for translating the APPA to the APPA-T form (Abubakar et al., 2013; Carlson, 2000) and conducted a translation equivalency test to confirm the adequate translation. Further, we employed an EFA and a CFA to examine the structural validity of the APPA-T. Regarding reliability, we estimated Cronbach's alpha coefficients to determine internal consistency and conducted a test-retest to measure the stability of the APPA-T over time. Overall, we found that the translation of APPA-T was linguistically and culturally appropriate for the undergraduate population in which we tested the instrument.

4.1. Factorial Structure

The original version of the APPA included four factors with 30 items with an undergraduate student sample (Dillman Taylor et al., 2015). Consistent with the previous studies (e.g., Dillman Taylor et al., 2019; Dillman Taylor & Mullen, 2019), the current results of the EFA produced a 4-factor structure for the APPA-T, and the CFA confirmed this structure with a sample of undergraduate students. However, unlike the original version with 30 items, the APPA-T included 24 items because we deleted APPA items 1, 5, 15, 16, 18, and 22 from the model due to poor loadings or substantial cross-loadings. These findings are consistent with previous studies that removed items as well (e.g., Dillman Taylor et al., 2019; Dillman Taylor & Mullen, 2019). However, the current study, similar to previous studies, removed items, all studies found a similar four-factor structure, inclusive of the personality priorities superiority, pleasing, control, and comfort. Hence, this study supports that the factorial structure of the APPA-T is valid for Turkish undergraduate students. This finding also indicates that the four personality priorities, proposed by Kefir (1971) as one of the concepts of Adlerian theory, can be applicable for Turkish culture. At least theoretically, this result suggests that Adlerian personality priorities are multiculturally sensitive to Turkish culture and a possible mechanism to view Turkish individuals' view of self, others, and the world.

4.2. Implication for Counseling Practice

The preliminary results of this study found that the APPA-T appears to be an applicable and valid instrument for the Turkish undergraduate sample. The APPA measures four personality priorities, which provides insight into individuals' lifestyle or worldview (Kefir, 1971; Ward, 1979). Understanding a client's lifestyle is one of the crucial goals in Adlerian therapy. More specifically, Adlerian psychotherapy identifies four steps: (1) establishing the therapeutic relationship, (2) assessing and understanding the lifestyle, (3) gaining insight, and (4) reeducation or reorientation (Oberst & Stewart, 2003; Sweeney, 2009). In the second phase, the role of a therapist is to understand and conceptualize client's lifestyle to gain insight into their presenting problems to aid in the development of a treatment plan. Adlerian theorists provided several ways to understand clients' lifestyles, such as family constellation, family values, birth order, early collections, and personality priorities (Ansbacher & Ansbacher, 1956; Sweeney, 2009). In addition, formalized lifestyle assessment tools such as Basic Adlerian Scales for Interpersonal Success-Adult Form (BASIS-A; Kern et al., 1997) and the APPA (Dillman Taylor et al., 2015) have been suggested to gather information about lifestyle (Dillman Taylor & Mullen, 2019; Oberst & Stewart, 2003). The BASIS-A is a more in-depth instrument to examine a client's lifestyle. On the other hand, the APPA help therapists briefly assess goals of significance and belonging to develop tentative hypotheses to help clients understand what behaviors, thoughts and/or feelings are keeping them stuck. In this sense, Turkish counseling professionals can use the APPA-T in their practice to guide their conceptualization and treatment of their clients' lifestyles. However, it should also be noted that Adlerian therapists

tend to avoid over-categorizing individuals into types by asserting that each person is unique in his or her context (Adler, 1929). Therefore, we suggest using the APPA-T as a conversation starter to provide tentative hypotheses regarding the clients' lifestyle in the therapeutic work. For example, practitioners in Turkey can utilize APPA-T to their clients, especially during the initial phase of the therapy process, to gather initial information of presenting problems, which can relate to clients' lifestyles. This is especially crucial when considering the lack of an instrument in Turkish to measure clients' lifestyles.

4.3. Limitations and Future Research

This study examined the psychometric properties and provided evidence of the use of the APPA-T with a sample of Turkish undergraduate students. Even though the study sample met the requirements of conducting the EFA and CFA, one of the study's limitations is that the sample consisted of only Turkish undergraduate students, representing a non-clinical population, limiting the generalizability to non-clinical samples. However, the preliminary results from this study indicate the possibility of the APPA-T as a possible option. Therefore, the researchers believe that the APPA-T should be evaluated with a clinical sample in future research. University students also represent a young and educated population when compared to the rest of society. Therefore, future studies should investigate the validity and reliability of the APPA-T with a variety of other Turkish samples (e.g., clinical populations; various educational, socioeconomic, and age groupings).

Another limitation is that the researchers confirmed the APPA-T with 24 items instead of the original 30-item APPA. Nevertheless, this study found a four-factor structure consistent with the original model (Dillman Taylor et al., 2015) and previous findings (Dillman Taylor & Mullen, 2019; Dillman Taylor et al., 2019). Future studies can be helpful to retest the factorial structure with 30-item and 24-item APPA-T with various other Turkish samples to confirm the findings of this study. Although this study investigated the construct validity of the APPA-T with Turkish undergraduate students, future studies are also recommended to examine the divergent and convergent validity of the APPA-T with other constructs such as other Adlerian concepts (e.g., social interest, feeling of inferiority, life style), concepts from other theories (e.g. cognitive schemas from cognitive theory), and/or other personality measures (e.g., NEO personality inventory- revised, MBTI).

4.4. Conclusion

Understanding one's lifestyle is crucial in Adlerian therapy, and personality priorities were developed to quickly and efficiently assess how individuals view self, others, and the world to treat clients in a more timely manner (Kefir, 1971). Dillman Taylor et al. (2015) developed the APPA to assess individuals' primary personality priority. The researchers adapted the APPA into Turkish and tested its psychometric properties with a Turkish sample in this study. The preliminary results revealed that the items on the APPA-T are reliable and valid for use in research and practice with Turkish undergraduate students. In addition, the results of this study supported that Adlerian personality priorities are appropriate for Turkish culture, which is consistent with multicultural aspects of Adlerian theory. Nonetheless, future studies would be helpful to continue testing the APPA-T with Turkish samples.

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:** Düzce University, 29/06/2018-2018/24.

Authorship Contribution Statement

Abdi GUNGOR: Research design, literature review, data collection, data analysis, methodology, software, resources, discussion, writing. **Dalena DILLMAN TAYLOR:** Research design, supervision, validation, methodology, discussion, writing.

ORCID

Abdi Gungor  <https://orcid.org/0000-0002-7945-0906>

Dalena Dillman Taylor  <https://orcid.org/0000-0002-3584-9982>

5. REFERENCES

- Abubakar, A., Dimitrova, R., Adams, B., Jordanov, V., & Stefenel, D. (2013). Procedures for translating and evaluating equivalence of questionnaires for use in cross-cultural studies. *Bulletin of the Transilvania University of Braşov*, 6 (55), 79-86.
- Adler, A. (1929). *The science of living*. Greenberg.
- Adler, A. (1931). *What life could mean to you*. Hazelden.
- Akçabozan, N. B., & Sümer, Z. H. (2016). Adler yaklaşımında aile danışmanlığı [Adlerian Family Counseling]. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, 6(46), 87-101.
- Ansbacher, H., & Ansbacher, R. (Eds.). (1956). *The individual psychology of Alfred Adler*. Basic Books, Inc.
- Ashby, J. S., Kottman, T., & Rice, K. G. (1998). Adlerian personality priorities: Psychological attitudinal differences. *Journal of Counseling and Development*, 76, 467-474. <https://doi.org/10.1002/j.1556-6676.1998.tb02706.x>
- Ashby, J. S., Kottman, T., & Stoltz, K. B. (2006). Multidimensional perfectionism and personality profiles. *The Journal of Individual Psychology*, 62, 312-323.
- Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.
- Carlson, E. D. (2000). A case study in translation methodology using the health-promotion lifestyle profile II. *Public Health Nursing*, 17(1), 61-70. <https://doi.org/10.1046/j.1525-1446.2000.00061.x>
- Carlson, J. M., & Carlson, J. D. (2000). The application of Adlerian psychotherapy with Asian-American clients. *Individual Psychology*, 56(2), 214-225.
- Carlson, J., Watts, R. E., & Maniacci, M. (2006) *Adlerian therapy: Theory and practice*. American Psychological Association.
- Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10(7). <https://doi.org/10.7275/jyj1-4868>
- Dillman Taylor, D., & Mullen, P. R. (2019). Adlerian Personality Priority Assessment: A Psychometric Evaluation. *The Journal of Individual Psychology*, 75(2), 122-144.
- Dillman Taylor, D., Bratton, S. C. & Henson, R. K. (2019): Confirming the Constructs of Adlerian Personality Priority Assessment. *Measurement and Evaluation in Counseling and Development*. <https://doi.org/10.1080/07481756.2019.1595814>
- Dillman Taylor, D., Gungor, A., Blount, A. J., & Mullen, P. R. (2018). Personality Priorities and Perceived Wellness Among Counseling Trainees. *The Journal of Individual Psychology*, 74(2), 188-208.
- Dillman Taylor, D., Ray, D. C., & Henson, R. K. (2015). Development and factor structure of the Adlerian Personality Priority Assessment. *Archives of Assessment Psychology*, 5(1), 23-36.
- Ergün-Başak, B., & Ceyhan, E. (2011). Psikolojik danışma ilişkisinde Adler yaklaşımına göre cesaretlendirme [Adlerian Encouragement in Counseling Relationship]. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, 4(35), 92-99.

- Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, 6(4), 304-312. <https://doi.org/10.1037/1040-3590.6.4.304>
- Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling*, 3, 62-72. <https://doi.org/10.1080/10705519609540030>
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data analysis*. Pearson.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Psychology Press.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods*, 6(1), 53-60.
- Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Kefir, N. (1971). *Priorities: A different approach to life style*. Paper presented at the International Committee of Adlerian Summer School and Institutes (ICASSI), Tel Aviv, Israel.
- Kefir, N. (1981). Impasse/priority therapy. In R. Corsini (Ed.), *Handbook of innovative psychotherapies*. Wiley.
- Kefir, N., & Corsini, R. J. (1974). Dispositional sets: A contribution to typology. *The Journal of Individual Psychology*, 30, 163-178.
- Kern, R. M., Wheeler, M. S., & Curlette, W. L. (1997). BASIS-A inventory interpretive manual: A psychological theory. TRT Associates.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). Guilford Press.
- Langenfeld, S., & Main, F. (1983). Personality priorities: A factor analytic study. *The Journal of Individual Psychology*, 39, 40-51.
- Mvududu, N. H., & Sink, C. A. (2013). Factor analysis in counseling research and practice. *Counseling Outcome Research and Evaluation*, 4(2), 75-98. <https://doi.org/10.1177/2150137813494766>
- Oberst, U. E. & Stewart, A. E. (2003) *Adlerian psychotherapy: An advanced approach to individual psychology*. Routledge Taylor & Francis Group.
- O'Rourke, N., Psych, R., & Hatcher, L. (2013). *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. SAS Institute.
- Pallant, J. (2010). *SPSS Survival Manual*, (4th ed.). McGraw Hill.
- Pew, W. L. (1976). *The number one priority*. John's Hospital, Marriage and Family Education Center.
- Sümer, Z. H., & Rasmussen, P. R. (2012). Individual Psychology in Turkey. *Journal of Individual Psychology*, 68(4), 411-421.
- Sweeney, T. J. (2009). *adlerian counseling and psychotherapy: A practitioner's approach* (5th ed.). Routledge Taylor & Francis Group.
- Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics* (6th ed.). Pearson.
- Ward, D. E. (1979). Implications of personality priority assessment for the counseling process. *Individual Psychologist*, 16(2), 12-16.
- Weston, R., & Gore, P. A. J. (2006). A brief guide to structural equation modeling. *Counseling Psychologist*, 34(5), 719–751. <https://doi.org/10.1177/0011000006286345>

Classification of Scale Items with Exploratory Graph Analysis and Machine Learning Methods

Ilhan Koyuncu ^{1,*}, Abdullah Faruk Kiliç ²

¹Adıyaman University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education

²Adıyaman University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education

ARTICLE HISTORY

Received: Feb. 15, 2021

Revised: Oct. 22, 2021

Accepted: Nov. 05, 2021

Keywords:

PISA,

Machine learning,

Exploratory factor analysis,

Exploratory graph analysis,

Monte Carlo simulation,

Scale development.

Abstract: In exploratory factor analysis, although the researchers decide which items belong to which factors by considering statistical results, the decisions taken sometimes can be subjective in case of having items with similar factor loadings and complex factor structures. The aim of this study was to examine the validity of classifying items into dimensions with exploratory graph analysis (EGA), which has been used in determining the number of dimensions in recent years and machine learning methods. A Monte Carlo simulation was performed with a total number of 96 simulation conditions including average factor loadings, sample size, number of items per dimension, number of dimensions, and distribution of data. Percent correct and Kappa concordance values were used in the evaluation of the methods. When the findings obtained for different conditions were evaluated together, it was seen that the machine learning methods gave results comparable to those of EGA. Machine learning methods showed high performance in terms of percent correct values, especially in small and medium-sized samples. In all conditions where the average factor loading was .70, BayesNet, Naive Bayes, RandomForest, and RselibKnn methods showed accurate classification performances above 80% like EGA method. BayesNet, Simple Logistic and RBFNetwork methods also demonstrated acceptable or high performance under many conditions. In general, Kappa concordance values also supported these results. The results revealed that machine learning methods can be used for similar conditions to examine whether the distribution of items across factors is done accurately or not.

1. INTRODUCTION

Exploratory factor analysis (EFA) is frequently used in scale development or adaptation studies (Fabrigar et al., 1999; Floyd & Widaman, 1995; Kline, 1994). There is a wide acceptance in the literature that EFA can be used in the process of searching evidence for construct validity (Nunnally & Bernstein, 1994). For this reason, the correct use of this frequently used method becomes important in terms of the correctness of decisions made by the researchers (Kılıç & Koyuncu, 2017; Koyuncu & Kılıç, 2019).

*CONTACT: Ilhan Koyuncu ✉ ilhankync@gmail.com 📍 Adıyaman University, Faculty of Education, Department of Educational Sciences, Measurement and Evaluation in Education

e-ISSN: 2148-7456 /© IJATE 2021

While performing EFA, it should be examined whether the data set meets the assumptions or not. The assumptions of EFA are the fact that the variables need to have a multivariate normal distribution, the sample size should be sufficient, there must be a linear relationship between the variables, and there need to be no extreme values in the data set, and no multicollinearity and singularity among the variables (Tabachnik & Fidell, 2012). After the data sets are analyzed in terms of assumptions, some methods are used to decide the number of factors. These methods include parallel analysis (Horn, 1965; Timmerman & Lorenzo-Seva, 2011), scree plot (Cattell, 1966), Minimum Average Partial (MAP) analysis (Guadagnoli & Velicer, 1988), and K1 (Kaiser, 1960) rule. After deciding on the number of factors by using several of these methods, the distribution of the items to the factors is examined. It is suggested that the factor loadings of the items should be above .30 (Costello & Osborne, 2005), .32 (Tabachnik & Fidell, 2012), or .40 (Howard, 2016). In this case, using different rotation methods in multi-dimensional structures, the items are to be placed in the dimensions in a meaningful way.

There are both vertical and oblique rotation methods used in multidimensional structures in placing the items to the dimensions. Oblique rotation methods are used if there is a relationship between factors, and vertical rotation methods are used if there is no relationship (Osborne, 2015). However, because there are many rotation methods and different rotation methods give different results, researchers may have difficulty in interpreting factor structures. At this stage, whether the revealed factor structure is compatible with the relevant literature or not is evaluated. On the other hand, it becomes difficult to decide which item will be included in which dimension, especially in cases with overlapping factor loadings. At this point, establishing a mainstay in placing items into dimensions will make it easier for researchers to have accurate decisions. Machine learning methods do not prevent cross loadings; however, they can be used to give researchers an idea about placing the items having cross loadings into the accurate dimension. Therefore, the primary purpose of this study is to use machine learning methods to classify items. However, in the related literature (e.g. Belvederi Murri et al., 2020; Fischer & Alfons Karl, 2020; Kjellström & Golino, 2019; Panayiotou et al., 2020), it is seen that EGA (Golino & Epskamp, 2017) is also used to reveal the relationships between the items. To this end, this study aimed to compare the results of machine learning methods, whose purpose is to make classification, with the EGA, which was developed to explain the relationships between items. Therefore, whether machine learning and EGA would give valid results in the classification of items to the dimensions was examined in this particular study.

EGA is a technique of estimating the number of dimensions and classification of items based on network psychometrics. Network psychometrics is a field that was developed to model networks in psychological data and at the same time, it has undergone advances that allow examining relationships between items (Golino & Epskamp, 2017). EGA makes estimates using the Gaussian graphic model. The Gaussian graphical model predicts the common distribution of variables using the inverse of the variance-covariance matrix. As a result of the estimation, nodes and edges connecting these nodes are obtained. In factor analysis, nodes correspond to items, while edges correspond to factor loadings (Golino & Epskamp, 2017; Golino et al., 2020). As a result of EGA, both the number of factors and those items grouped together are obtained. Golino and Epskamp (2017) and Golino et al. (2020) compared EGA with the methods of determining the number of dimensions (such as parallel analysis, K1 rule, and MAP analysis) and reported EGA as the method that gave the most accurate results when sample size was 5000, the factor structure was four-dimensional, and the correlation between dimensions was .70. Moreover, there are also many researchers who use the EGA in their studies in terms of examining individual differences (Fischer & Alfons Karl, 2020), relationship between observed and latent variables (Belvederi Murri et al., 2020), estimating the number of latent variables (Kjellström & Golino, 2019), and exploring the dimensionality of the social skills (Panayiotou et al., 2020).

Since the focus of this study was the classification of scale items, the performance of machine learning methods seeking answers to classification problems was evaluated. Machine learning methods mostly focus on classification, estimation, and clustering problems. Machine learning which is used to analyze a variety of data structures is one of the fastest developing technical areas of today. This technique, located in the center of artificial intelligence and data science, is at the intersection of computer science and statistical machine learning methods are used in many fields (Jordan & Mitchell, 2015). In this study, whether scale items were correctly classified into the dimensions or not was investigated. For this purpose, some frequently used machine learning methods (Pu et al., 2020) given under the titles Bayes, functions, lazy and trees in the Waikato Environment for Knowledge Analysis (WEKA) software (Hall et al., 2009) were compared. The reason for selecting many methods that are frequently used in the machine learning is based on the fact that different methods are effective for different data structures in the related literature (see Barker et al., 2004; Romero et al., 2013). Summary information about the algorithms used is given in Table 1.

Table 1. Machine learning algorithms used in the study.

Title	Algorithm	Explanations
Bayes	BayesNet	Classifies with the method of Bayesian networks. Outputs for network structure, conditional probability distributions, and Bayesian networks are obtained. Various search algorithms and quality measures are used (Hall et al., 2009).
	NaiveBayes	The main purpose of this algorithm that is used in supervised learning is to predict classification probabilities based on estimated class probabilities (John & Langley, 1995).
functions	RBFNetwork	Uses the normalized Gaussian Radial Basis Function network. Its main function is the k-mean clustering method, while training is performed by logistic or linear regression. Standardizes all numerical variables to 0 mean and unit variance (Hall et al., 2009).
	SimpleLogistic	It is a classifier that generates linear logistic regression models. LogitBoost, which uses simple logistic regression functions, is used to fit logistic models (see Landwehr et al., 2006; Sumner et al., 2005).
lazy	KStar	It differs from other instance-based algorithms in terms of being entropy-based. This method enables the classification of the tested objects according to their proximity to similar objects in the learning data, based on some proximity functions (Hall et al., 2009). Detailed information on the technical structure and usefulness of the method was provided by Cleary and Trigg (1995).
	RseslibKnn	This <i>k</i> -closest neighborhood classifier with many distance criteria finds fast neighborhoods in large samples and can be applied to numerical and categorical data (see Wojna & Latkowski, 2018; Wojna et al., 2019).
trees	J48Consolidated	With or without pruning, C4.5 creates a consolidated decision tree. Consolidated Tree Construction (CTC) creates a single decision tree based on subsets (see Pérez et al., 2007). A new method has been added to this algorithm to determine the number of clusters to be used in the consolidation process (see Ibarguren et al., 2015).
	RandomForest	It is a classifier based on the generation of random decision trees (see Breiman, 2001).

BayesNet and NaiveBayes given in Table 1 are algorithms based on Bayes theorem. Bayesian methods, which make inferences based on probabilistic estimates, have been an important alternative to usual methods in machine learning such as decision trees and artificial neural networks (John & Langley, 1995). Naive Bayes algorithm, which is frequently used in machine learning field as well as decision trees and neural networks, can perform in estimation and

predictions as well. The fact that the method has conditional independence assumption caused it to be described as 'naive' (Han et al., 2011). Similarly, BayesNet makes a graphical classification process, which makes estimations according to network structures obtained based on conditional probability distributions (Alpaydin, 2010; Bouckaert, 2008). Generally, a Bayes classifier assigns an instance with the highest value to the class after selecting that class with the highest probability in the model having the least error according to the Bayes rule (Alpaydin, 2010; John & Langley, 1995).

In logistic regression models, the probability of a data set belonging to the last class is estimated by subtracting the sum of the probabilities of belonging to each class from the value 1 (Landwehr et al., 2006; Sumner et al., 2005). Radial basis functions (RBF), one of the artificial neural network models, work similarly to perceptron models, but use the gauss function as the threshold function (Akpınar, 2014). RBF network (Hall et al., 2009), whose basic function is obtained with the k-mean clustering method and training with logistics or linear regression (Hall et al., 2009), is generally used in classification problems, modeling and system control fields, and time series analysis. The nearest neighborhood methods that belong to the family of instance-based classification algorithms perform analysis based on distance measures and have many types (Aha et al., 1991). In this mathematics-based method, the instances in the test data are classified according to their positions in the training data in a multidimensional space (Larose & Larose, 2014). KStar algorithm differs from other object-based algorithms in terms of using entropy-based functions (Cleary & Trigg, 1995). RseslibKnn algorithm, which can find fast neighborhoods in large samples, is a method that includes different distance metrics for different types of attributes (see Wojna & Latkowski, 2018; Wojna et al., 2019).

While the classification algorithms based on decision trees are very diverse, J48 and random forest methods are among the most frequently used machine learning methods (Pu et al., 2020). With the addition of new options to the J48 algorithm, the J48 Consolidated algorithm, which creates a single decision tree based on subsets, has been developed as a robust method for classification problems with its high performance (Ibarguren et al., 2015; Pérez et al., 2007). This algorithm generates a consolidated C4.5 decision tree (Quinlan, 1993) with or without pruning (Hall et al., 2009). The random forest classifier (Breiman, 2001) has become one of the most popular machine learning techniques used in such fields as mining, archeology, engineering and wine (Li et al., 2019) in recent years, due to its highly reliable and interpretable results in complex data and its performance comparable with other frequently used machine learning techniques (Zhang & Yang, 2020). In addition, random forests have many advantages such as high classification performance in many data types, handling dimensionality, being capable of variable importance analysis, highly adaptability and time efficiency (Li et al., 2019). The random forest method is a mixture of tree estimators in which each tree has the same distribution for every other tree in the forest and each tree is autonomously dependent on the values of the random vector sets (Breiman, 2001).

Although there are many studies on the effectiveness a wide variety of machine learning techniques on different data types in different fields such as education (e.g. Baker, 2010; Berens et al., 2019; Bulut & Yavuz, 2019; Güre et al., 2020; Hamalainen & Vinni, 2006; Koyuncu , & Gelbal, 2020; Romero & Ventura, 2013), health sciences (e.g. Beleites et al., 2013; Chu et al., 2012; Figueroa et al., 2012; Shao et al., 2013), engineering sciences (e.g. Brain & Webb, 1999; Hegde & Rokseth, 2020; Reich & Barai, 1999), economics (e.g. Azqueta-Gavaldón, 2017; Mele & Magazzino , 2020; Mullainathan & Spiess, 2017), politics (Grimmer, 2015; Guess et al., 2019), environmental sciences (e.g. Heydari & Mountrakis, 2018; Zhang, & Yang, 2020; Mele & Magazzino, 2020). However, in the relevant literature, studies on how machine learning methods can bring solutions to problems in the field of scale development are limited (e.g. Auerswald & Moshagen, 2019; Baldi & Hornik, 1989; Chattopadhyay et al., 2011; Goretzko &

Bühner, 2020; Tezbaşaran, & Gelbal, 2018). Therefore, there is a need to examine how the use of machine learning methods in scale development studies will bring solutions to existing problems. This study, in line with this need, has examined whether machine learning methods and EGA can be a solution to the problems encountered in placing the items in the dimensions.

When studies on exploratory factor analysis using machine learning methods are examined, it can be seen that such studies generally focus on factor retention (e.g. Goretzko & Bühner, 2020; Iantovics et al., 2019). As a result of these studies, it has been reported that machine learning methods can generally be used with traditional methods. In the study conducted by Goretzko and Bühner (2020), it was stated that the ranger and xgboost algorithm were the most accurate methods for 3204 conditions in determining the number of factors. However, these studies do not seek answers to the research problem of correctly classifying the scale items into the factors that the current study deals with. Therefore, it is important to examine whether machine learning methods, which provide solutions to classification problems, can be used in scale development and adaptation studies. In addition, researchers can evaluate the accuracy of their decisions by using these methods in cases where their correct classification percentages are high. For example, such methods let the researchers place the items on a two-dimensional scale as a result of their EFA. In this case, according to the characteristics of the data set, it can be checked whether the items are correctly classified to the dimensions by machine learning methods or EGA. Hence, an evidence related to decision validity can be obtained. For this reason, this study is important in terms of its contribution to the relevant literature and the convenience it will provide to researchers. This study is also important in terms of allowing practitioners to test the correct classification of the items into their dimensions by using machine learning methods. Therefore, this study seeks answers to the following research problems:

Under different simulation conditions for EGA and machine learning methods:

- 1) What are the correct classification percentage values?
- 2) How are Kappa concordance values for confusion matrices?

2. METHOD

This study is a Monte Carlo simulation since it was carried out to compare the classification performances of machine learning methods in different factor structures. In Monte Carlo simulation studies, sample data are generated in accordance with the desired distribution characteristics (Bandalos & Leite, 2013). In this study, the data sets were generated as 5-point likert type scale. The skewness of data was adjusted as left-skewed, normal, and right-skewed.

2.1. Simulation Conditions

In the present study, in order to examine the performance of different methods, a set of simulation conditions were determined. These conditions included average factor loadings (.40 and .70), sample size (100, 200, 500 and 1000), number of items per dimension (5 and 10), and number of dimensions (2 and 3). In addition, distribution of data (left-skewed, normal, and right-skewed) conditions were investigated. In the study, a total of $2 \times 4 \times 2 \times 2 \times 3 = 96$ simulation conditions were studied and 100 replications were made.

The conditions for the average factor loadings were manipulated to be .40 and .70. In addition to the researchers who state that the factor loadings of the items in the scales should be at least .30 (Costello & Osborne, 2005), there are also researchers who state that it should be at least .32 (Tabachnik & Fidell, 2012). Besides, Howard (2016) states that this value should be at least .40. For this reason, in this study, data sets were produced with an average factor loading of .40 by taking the average factor loadings. On the other hand, .70 was added as another factor loading condition in order to examine how the increase in the average factor loadings affects the performance of the methods.

The conditions for sample size were manipulated to be 100, 200, 500, and 1000. The sample size is frequently selected as 200, 500, and 1000 in factor analysis studies and is defined as small, medium, and large (Beauducel & Herzberg, 2006; Li, 2016b; West et al., 1995). In addition, Gorsuch (1974) suggested that the sample size should be at least 200. On the other hand, since this study investigated the classification performance of machine learning methods, samples with 100 instances were also added to the sample size conditions in order to examine the classification performance in smaller samples. For example, in educational data, it is possible to have data for 50 or even fewer students. Therefore, in this study, small sample sizes were preferred in order to examine the performances of methods at the same time.

The conditions for the number of items per dimension were manipulated to be 5 and 10. In classification methods, imbalanced or balanced distribution of class variable can cause different results (Sun et al., 2006). For this reason, only a balanced distribution (the same number of items per dimension) was examined in this study. Although it is suggested that a dimension should be defined with at least 3 items, it is stated that more items would increase the reliability of the dimension (Brown, 2015). For this reason, 5-item conditions for one dimension were added to the study. In addition, 10-item condition was also added to the study to examine the effect of increasing the number of items on the performance of the methods.

The conditions for the number of dimensions were manipulated to be 2 and 3. 2-dimension condition was investigated because there had to be a dependent variable with at least two categories to make the classification. In addition, 3-dimensional condition was also included to examine how the increase in the number of dimensions would affect the performance of the methods. Since the interfactor correlations in the real data sets were mostly between .20 and .40 (Li, 2016a), it was fixed to .30, the value in the middle of this interval in the present study.

The conditions for the distribution of the data were manipulated to be left-skewed, normal, and right-skewed. This condition was added to the study in order to examine how the change in the distribution of data would affect the performance of methods. Since it was stated that the skewness coefficient can be considered normal for the interval $[-2, 2]$ (Chou & Bentler, 1995; Curran et al., 1996; Finney & DiStefano, 2013), data was categorized in such a way that the coefficient of skewness was 2.5 for a right-skewed distribution and -2.5 for a left-skewed distribution. The data was first generated to show a continuous normal distribution and then it was categorized according to threshold values.

2.2. Data analysis

The *lavaan* (Rosseel, 2012) package included in the R software (R Core Team, 2020) was used to generate the data. *EGAnet* (Golino & Christensen, 2020) package was used for exploratory graphic analysis. There are two different methods when performing EGA. These are the graphical least absolute shrinkage and selection operator (GLASSO), and triangulated maximally filtered graph approach (TMFG). In this study, the TMFG method, which was found to give more accurate results (Golino et al., 2020) in many conditions, was used. While the codes written by the researchers were used to calculate the percent correct values from the EGA results, the Kappa values were obtained with the *caret* (Kuhn, 2020) package.

Analyzes for machine learning methods were performed in the Experimenter module of WEKA (Hall et al., 2009, Bouckaert et al., 2020) with 10-fold cross-validation (Lachenbruch & Mickey, 1968). Cross-validation which was performed by dividing data into a number of folds (usually 10 folds) is a method used when the data is not large enough to divide it into training and test data (Witten et al., 2017). Since the scale items were classified instead of subjects in this study, the data set was transposed, and hence the number of instances was limited to the number of items. Therefore, 10-cross-validation method was used in this study. Bootstrapping (Efron, 1983) method is used when the data sets are medium (approximately 1000) or larger

(more than 1000); otherwise, holdout methods are used for small (less than 1000 subjects) sample sizes in machine learning.

2.3. Model Evaluation Criteria

Percent correct values were used to compare the performance of EGA and machine learning methods in the study. The percentage of correctly classified items into the dimensions for 100 replications was calculated. For this purpose, first, it was checked whether the number of recommended factors was estimated correctly. If it was correct, then it was examined whether the items were correctly classified into the dimensions. The percent correct values were obtained by dividing the number of replications in which the items were in the correct factors by the number of replications (100). Since it was stated that percent correct values should be above 80% (Hartmann, 1977), it was used as cut off value for percent correct.

There are many criteria to evaluate the classification performance of machine learning methods. The most used criteria are accuracy (percent correct), error rate, precision, recall, sensitivity, specificity, receiver operating characteristic (ROC) curve, F criterion, and Kappa statistics. These values are calculated by creating a confusion matrix via the classification results. In classification, there are frequencies belonging to the instances classified into the cells of confusion (error) matrix. In an error matrix consisting of 2x2 classes a and b, there are frequencies belonging to instances classified correctly into classes a and b (True positive [TP] and True Negative [TN]). Also, there are instances classified into class b while it should be in class a (False positive [FP]), and in class b while it should be in class a (False Negative [FN]). Based on these frequencies, the accuracy rate is obtained by dividing the number of correctly classified instances to the total number of instances. Error rate is obtained by subtracting the accuracy rate from 1. Precision is calculated by dividing the value of TP by the sum of TP and FP. Sensitivity and Recall (True positive rate) measures are calculated by dividing the TP value to the sum of TP and FN. Specificity (True negative rate) value is calculated by dividing the TN value by the sum of TN and FP.

F value (Rijsbergen, 1979), another measure used in the evaluation of models, is an equally weighted function of precision and recall values, while $F\beta$ value is not a function of equal weights (Han et al., 2011). The Kappa statistic (Cohen, 1960; Fleiss, 1971), which evaluates the concordance in the confusion matrix, is evaluated as low if it is between 0-.20, acceptable if it is between .21-.40, medium if it is between .41-.60, very good if it is between .61-.80, and perfect if it is between .81-1.00 (Landis & Koch, 1977). The ROC curve (Egan, 1975) is a measure which is frequently used in binary classification. It is a graphical representation of TP value on the vertical axis and FP values on the horizontal axis regardless of class memberships or error cost (Witten et al., 2017). If the area value under this curve is around .50, it indicates that the model performance is low, and when it is approximately 1, the performance is high (Han et al., 2011).

There are many model evaluation criteria, but it is important to determine which one is suitable for the data set. In this sense, whether the dependent variable is binary or multinomial is an important case in choosing the evaluation criterion to be used. Since the ROC curve, precision, recall, and specificity measures are used when dependent variable is binary, these evaluation criteria were not used for the multinomial form of dependent variables in this study. Similarly, the F criterion is a measure that can be used if the number of observations in the class variable is unbalanced (Branco et al., 2015). For these reasons, percent correct values and Kappa concordance statistics were used as model evaluation criteria in the present study.

3. RESULT / FINDINGS

Findings of the study are presented in this section according to the order of the research problems given in the introduction section.

3.1. Comparison of Percent Correct Values

Percent correct (PC) values obtained from EGA and machine learning methods are presented in [Figure 1](#). In addition, PC values are given in [Appendix A](#) for the ones who want to examine these values in detail. The findings obtained in this section were examined for percent correct values of each method.

The increase in the average factor loading and the sample size increased the PC value in general. The factor loading was primarily effective on the classification performance of EGA. When the factor loading was .70, EGA had sufficient PC performance (>80%) even if sample size was small for normally distributed data. As the sample size increased, EGA had sufficient PC performance under the conditions where the average factor loading was .40. EGA had sufficient PC performance in 52.08% of all conditions.

BayesNet had sufficient PC performance in approximately 98% of the conditions where the average factor loading was .70. When the conditions in which average factor loading was .40 were examined, the increase in the sample size increased the PC performance of the methods. BayesNet had sufficient PC performance in 52.08% of all conditions.

J48Consolidated had sufficient PC performance in all conditions where the average factor loading was .70, the number of items was 10, the number of dimensions was 2, and sample sizes were 200 and 500 regardless of the distribution of the data. However, the PC performance was below 80% under the conditions where sample sizes were 500 and 1000 and average factor loading was .40 for the normally distributed data. TJ48Consolidated had sufficient PC performance in 10.41% of all conditions.

KStar had sufficient PC performance in all conditions where the sample sizes were 100 and 200 and the average factor loading was .70. With the increase of the sample size to 500, it did not have sufficient PC performance for normally distributed data sets. In all conditions where the sample size was 1000, the PC value was below 80%. When average factor loading was .40, the number of dimensions was 2, the distribution of data was normal, and KStar had sufficient PC performance. Generally, KStar had sufficient PC performance in 40.63% of all conditions.

NaiveBayes had sufficient PC performance in all conditions where the average factor loading was .70. The conditions where average factor loading was .40 and the number of items was 10 positively affected the PC performance of the method. The sample size being 500 and above made the PC performance of the method independent from the distribution. However, the method had a better performance in conditions where data were normally distributed and the sample sizes were 100 and 200. NaiveBayes had sufficient PC performance in 67.71% of all conditions.

RandomForest had sufficient PC performance in all conditions where the average factor loading was .70. However, it had not sufficient PC performance under any conditions where the average factor loading was .40 and the number of items was 5. Increasing the number of items and sample size increased the PC performance of the method. RandomForest had sufficient PC performance in 67.71% of conditions.

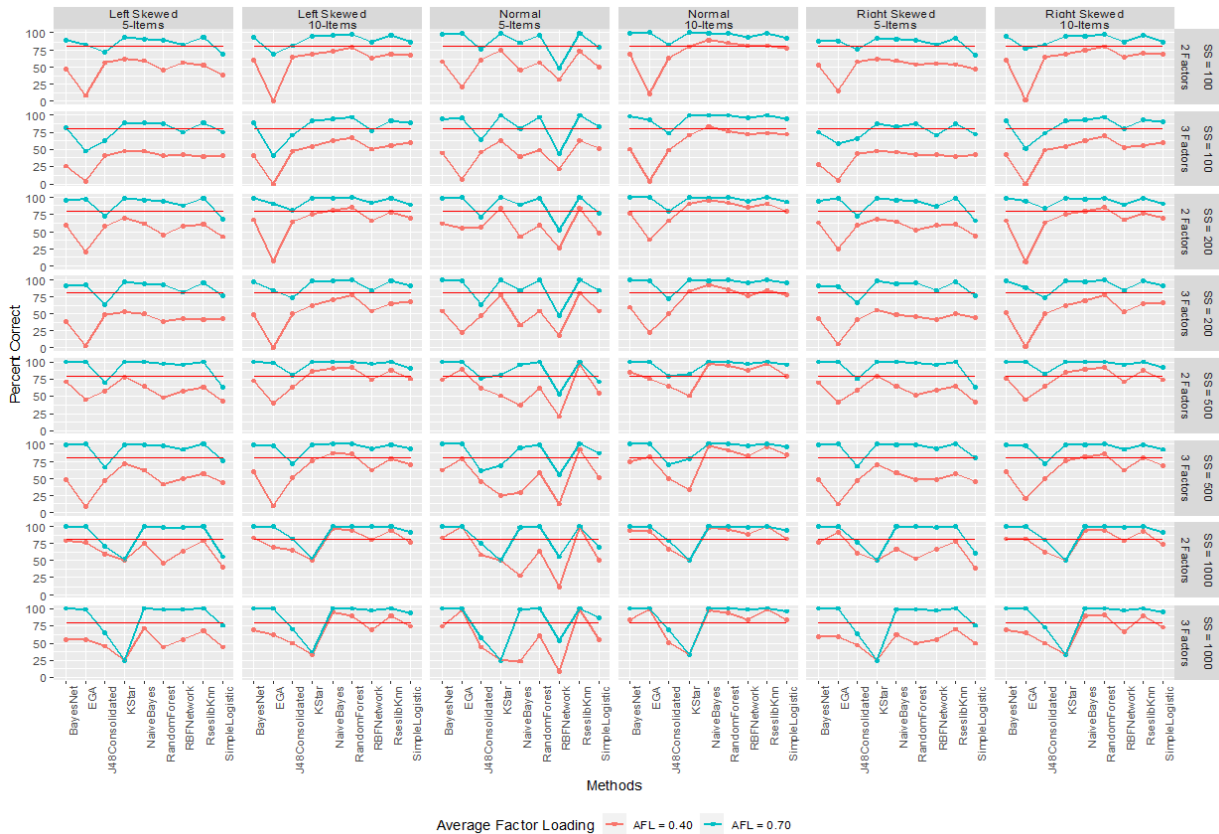
RBFNetwork had generally sufficient performance under conditions where the average factor loading was .70. However, it had not sufficient PC performance under any conditions where data were normally distributed and the number of items was 5. As the sample and the number of items increased, the PC values of the method also increased. RBFNetwork had sufficient PC performance in 45.83% of all conditions.

RseslibKnn had sufficient performance in all conditions where the average factor loading was .70. The PC value of the method was bigger than 80% in the conditions where sample sizes were 500 and 1000 and the number of items was 10. The number of dimensions did not have any effect on the PC performance of the method. It was especially noteworthy that it had 100%

PC value under all conditions where sample size was 1000 and average factor loading was 0.70. RseslibKnn had sufficient PC performance in 69.69% of all conditions.

SimpleLogistic had not sufficient performance under any conditions where the number of items was 5 in skewed data. The method had more than 80% PC values in 3 dimensional structures compared to 2 dimensional ones. However, the method had not sufficient PC performance under any conditions where the average factor loading was .40, and sample sizes were 100 and 200. Its PC values were higher than 80% in only one condition where the sample size was 500. SimpleLogistic had sufficient PC performance in 32.29% of all conditions.

Figure 1. Comparison of percent correct values of the methods.



3.2. Comparison of Kappa Concordance Values

Kappa values obtained from EGA and machine learning methods are presented in Figure 2. In addition, Kappa values are given in Appendix B for researchers who would like to examine the details.

EGA’s Kappa values varied between .69 and 1.00 for all simulation conditions. Accordingly, EGA had a very good matrix concordance in all conditions. However, it should be kept in mind that kappa values were calculated only with replications where the number of dimensions was estimated correctly. In other words, Kappa values should be evaluated together with percent correct values. According to these results, it can be said that EGA could classify items at a fairly good level in cases where the number of dimensions was estimated correctly.

BayesNet had good Kappa values above .60 in all conditions where the average factor loading was .70. When the conditions with an average factor loading of 0.40 were examined, it had more acceptable Kappa values that were obtained in large samples compared to small ones, in 2 dimensions compared to 3 dimensions, and in normal distribution compared to skewed distributions. In about 60% of the conditions where the average factor loading was .40,

moderate concordance was observed. BayesNet had good Kappa values in 66.66% of all conditions.

J48Consolidated had good Kappa values in 50% of the conditions where the average factor loading was .70, and the other 50% of the conditions had medium or acceptable Kappa values. In 25% of the conditions where the average factor loading was .40, medium and above concordance was observed, while acceptable concordance was obtained in other conditions. In this method, in general, higher Kappa values were obtained in 2 dimensions compared to 3 dimensions, for 5 items compared to 10 items per dimension. Changing the skewness and sample size did not cause a noteworthy change in the Kappa values. J48Consolidated had good Kappa values in 30.21% of all conditions.

KStar had perfect concordance in all conditions where sample sizes were 100, 200 and average factor loading was .70. Kappa values were slightly lower under conditions where sample size was 500 and there were normally distributed data than the skewed ones. However, Kstar had insufficient concordance under conditions where sample size was 1000 and the number of dimensions was 3. When the conditions with an average factor loading of .40 were examined, it tended to show a higher and better level of concordance in conditions where the number of dimensions was low, the number of items was 5, and the distribution of variables was normal. KStar had good Kappa values in 56.25% of all conditions.

NaiveBayes had good concordance in all conditions where the average factor loading was .70. Under conditions where the average factor loading was .40, it had much better Kappa values obtained in large samples compared to small ones and 3-dimensional structure compared to 2-dimensional ones. However, a better concordance was observed under conditions where data was skewed, and the number of items was 5 when compared to 10 items. The opposite of this case was true when there were 10 items per dimension. NaiveBayes had good Kappa values in 72.92% of all conditions.

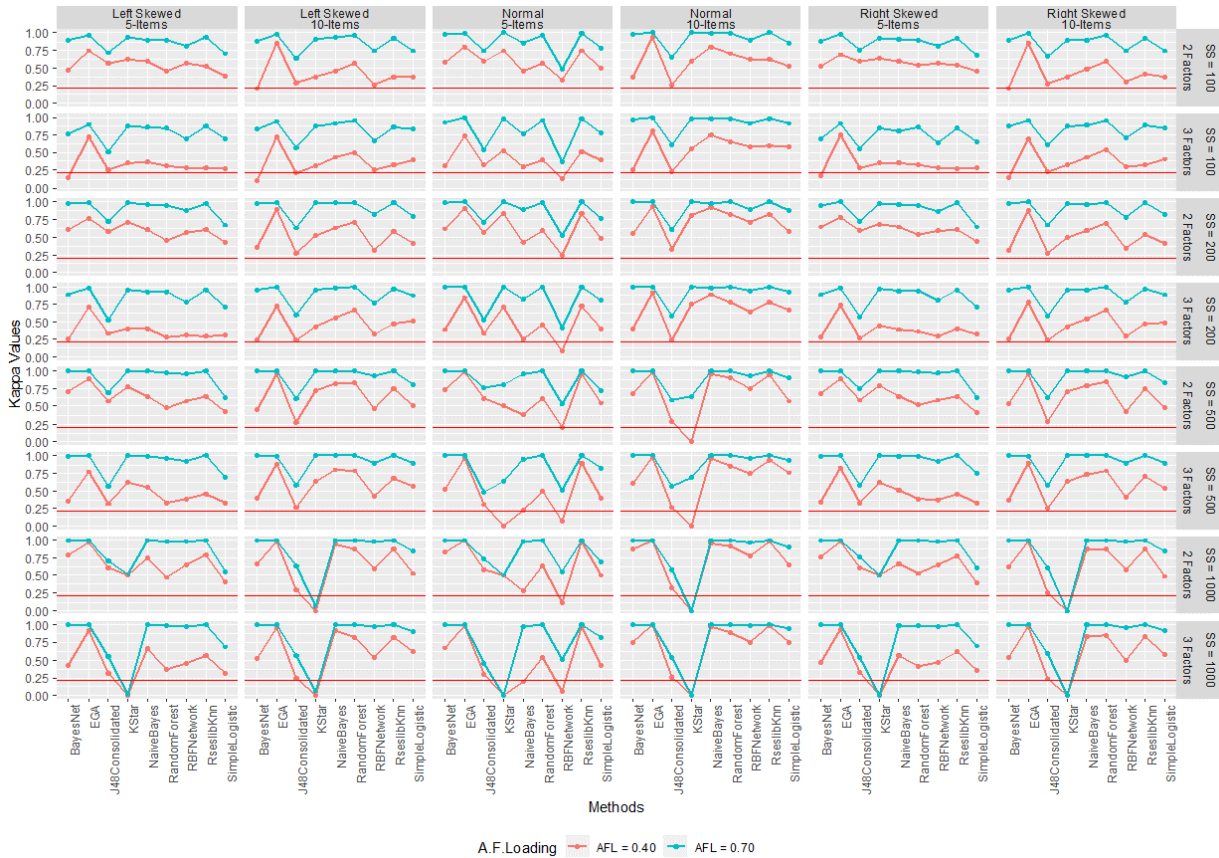
RandomForest had perfect Kappa values in all conditions where the average factor loading was .70. Under conditions where the average factor loading was .40, the increase in the number of items per dimension, sample size and the decrease in skewness increased the performance of the method. Generally, acceptable Kappa values were obtained. NaiveBayes had good Kappa values in 75% of all conditions.

RBFNetwork generally had perfect concordance in all conditions where the average factor loading was .70, except for conditions that the data were normally distributed, and the number of items was 5. In the conditions where the average factor loading was .40, the concordance was generally above the acceptable level, except for the conditions with a normal distribution and 5 items per factor. Overall, as the sample size got larger, the concordance increased. While the increase in the number of items per dimension decreased the performance in skewed data, it had the opposite in normal distributions. RBFNetwork had good Kappa values in 50% of all conditions.

RseslibKnn had perfect concordance in all conditions where the average factor loading was .70. There was an acceptable concordance when the average factor loading was .40 and the sample size was small. Under conditions where the sample size was over 200, a fairly good concordance was observed. In general, getting a larger sample size increased the concordance, especially in skewed data. The change in the number of dimensions did not have a noteworthy effect on the overall concordance for RseslibKnn. Especially in all conditions where the sample sizes were 500 and 1000 and the average factor loading was .70, it is noteworthy that the agreement was 1. The decrease in the number of items per dimensions and skewness of data increased the concordance. RseslibKnn had good Kappa values in 81.25% of all conditions.

SimpleLogistic had a very good or perfect concordance in all conditions where the average factor loading was .70. In all conditions where the average factor loading was .40, SimpleLogistic had an acceptable or above concordance level. In general, the increase in the number of items and the sample size increased concordance. Increasing the number of dimensions and skewness of data decreased the concordance. SimpleLogistic had good Kappa values in 54.17% of all conditions.

Figure 2. Comparison of Kappa concordance values of the methods.



4. DISCUSSION and CONCLUSION

In this study, the usability of exploratory graph analysis (EGA) and machine learning methods in deciding which item should be included in which dimension in the exploratory factor analysis was examined. The results obtained for different conditions were successively discussed for the performance of methods with regard to different sample sizes, average factor loadings, the number of items per dimensions, the number of dimensions, and the distribution of data.

When the findings obtained for different conditions were evaluated together, it was seen that machine learning methods gave comparable results to EGA. Machine learning methods showed high performance, especially in small and medium sample sizes. For example, in all conditions where the average factor loading was .70, BayesNet, Naive Bayes, RandomForest, and RseslibKnn methods had bigger values than 80% PC values similar to the values of EGA. BayesNet, Simple Logistic and RBFNetwork methods had also an acceptable or high PC performance under many conditions such as different sample sizes, factor loadings, and the number of items. These methods had better classification performance than that of EGA when factor loading was .40. Kappa concordance values also support these results. In general, higher percent correct and Kappa values were obtained in conditions where the average factor loading was .70 compared to the average factor loading .40.

Under conditions where the average factor loadings and the number of items per dimension were low, percent correct values below 80% were obtained regardless of the number of factors, skewness of data, and sample size. However, in conditions where the average factor loading was low and the number of items per dimension was high, sample size was small and data were normally distributed and PC values of Naive Bayes, RandomForest, RBFNetwork, RseslibKnn, and SimpleLogistic methods were close to 80% or above. These methods showed the same performance even if there was skewness in large sample sizes. Kappa values also greatly supported such a result. Especially when the number of items per dimension was more than 5, it was seen that these methods performed well even if the average factor loading was low. The fact that the methods were Bayesian, decision trees, artificial neural networks and instance-based showed that classification decisions can be made with different statistical and mathematical based methods. In addition, it was observed that the performance of some methods such as RBFNetwork and Kstar decreased in the conditions having 5 items and large sample sizes. This interesting result was considered to be obtained due to the mathematical structure of those methods.

Machine learning methods generally do not require any assumptions (except the conditional independence assumption for the Naïve Bayes). The results of this study showed that the number of categories, skewness of data, and sample size had an effect on the classification performance of these methods. Although they were not based on factor analysis, the results of other studies revealed that sample size (Beleites et al., 2013; Brain & Webb, 1999; Chu et al., 2012; Figueroa et al., 2012; Heydari & Mountrakis, 2018; Hamalainen & Vinni, 2006; Shao et al., 2013), feature selection (Chu et al., 2012), and the number of nominal classes (Minaei-Bidgoli et al., 2003; Nghe et al., 2007) had effects on the performance of machine learning methods. On the other hand, studies on factor analysis using machine learning generally focused on factor retention (e.g. Goretzko & Bühner, 2020; Iantovics et al., 2019). Therefore, the results of the present study provide researchers with a reference point in using and selecting the most suitable machine learning method for their data structure to decide on which items will be included in which factors. For example, assume that when a researcher cannot decide on which item belongs to which dimension after EFA analysis because an item can load more than one dimension at the same time (cross loading), the researcher in such a situation can try different methods given in the present study and place the item into the appropriate dimension by taking into account the conditions similar to her/his own study. In addition, in cases where it is necessary to perform item parceling, items can be grouped by using methods that give accurate results in the current study.

Due to many simulation conditions handled, the discussions were formed from generalized results for different conditions in the present study. Researchers who perform exploratory factor analysis can choose machine learning methods and classify scale items according to the characteristics of their data sets (sample size, average factor loading, and skewness of the data). In this case, they can compare the percent correct and Kappa values obtained from their study with the results of this study. For example, let us consider a method where PC value was obtained as 100% in current study. If the researcher obtains a very low value when he/she uses this method in his/her own data set, he/she may consider re-classifying the items. Thus, it will be possible to examine whether the items are in the right dimension or not. In addition, assume that researchers have been given a basis for decision-making. However, it should also be taken into account that this study does not cover all of the real situations that may actually occur. The level of similarity of the characteristics of the real data set with the conditions examined in the current study should also be taken into consideration.

In this study, eight machine learning methods based on different statistical and mathematical basis included in the WEKA software were examined. In future studies, the performance of

other methods such as Bayesian, artificial neural networks, instance based, rule based, decision tree, and support vector machine can also be examined. In addition, the number of conditions used in this research can be increased or the performance of EGA and machine learning methods used in the current study can be compared for different conditions such as inter-factor correlations. Since this study was carried out with simulated data sets, the performance of the EGA and machine learning methods can be examined over real data sets in similar conditions.

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

Ilhan Koyuncu: Investigation, Resources, Visualization, Software, Formal Analysis, Writing original draft, Methodology, Supervision, and Validation. **Abdullah Faruk Kilic:** Investigation, Resources, Software, Formal Analysis, Writing original draft, Methodology, Supervision, and Validation.

ORCID

Ilhan Koyuncu  <https://orcid.org/0000-0002-0009-5279>

Abdullah Faruk Kilic  <https://orcid.org/0000-0003-3129-1763>

5. REFERENCES

- Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning* 6, 37-66.
- Akpınar, H. (2014). *Veri madenciliği veri analizi [Data mining data analysis]*. Papatya Yayınları.
- Alpaydin, E. (2010). *Introduction to machine learning: Adaptive computation and machine learning series*. MIT Press.
- Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, 24(4), 468–491. <https://doi:10.1037/met0000200>
- Azqueta-Gavaldón, A. (2017). Developing news-based economic policy uncertainty index with unsupervised machine learning. *Economics Letters*, 158, 47-50.
- Baker, R. S. J. (2010). Machine learning for education. *International Encyclopedia of Education*, 7(3), 112-118.
- Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2, 53-58.
- Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.). Information Age.
- Barker, K., Trafalis, T., & Rhoads, T. R. (2004). Learning from student data. In *Proceedings of the 2004 Systems and Information Engineering Design Symposium* (pp. 79-86). IEEE.
- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186-203. https://doi.org/10.1207/s15328007sem1302_2
- Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta*, 760, 25-33.

- Belvederi Murri, M., Caruso, R., Ounalli, H., Zerbinati, L., Berretti, E., Costa, S., ... Grassi, L. (2020). The relationship between demoralization and depressive symptoms among patients from the general hospital: network and exploratory graph analysis: Demoralization and depression symptom network. *Journal of Affective Disorders*, 276(June), 137–146. <https://doi.org/10.1016/j.jad.2020.06.074>
- Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk - predicting student dropouts using administrative student data from German universities and machine learning methods. *Journal of Educational Machine learning*, 11(3), 1-41. <https://doi.org/10.5281/zenodo.3594771>
- Bouckaert, R. R. (2008). Bayesian network classifiers in Weka for Version 3-5-7. *Artificial Intelligence Tools*, 11(3), 369-387.
- Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2020). *WEKA manual for version 3-9-5*. University of Waikato.
- Brain, D., & Webb, G. (1999). On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Fourth Australian Knowledge Acquisition Workshop*, University of New South Wales (pp. 117-128), December 5-6, Sydney, Australia.
- Branco, P., Torgo, L., & Ribeiro, R. (2015). A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv:1505.01658*.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford.
- Bulut, O., & Yavuz, H. C. (2019). Educational machine learning: A tutorial for the " Rattle" package in R. *International Journal of Assessment Tools in Education*, 6(5), 20-36.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10
- Chattopadhyay, M., Dan, P. K., & Mazumdar, S. (2011). Principal component analysis and self-organizing map for visual clustering of machine-part cell formation in cellular manufacturing system. In *Systems Research Forum* (Vol. 5, No. 01, pp. 25-51). World Scientific Publishing Company.
- Chou, C. P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In Rich H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications*. Sage.
- Chu, C., Hsu, A. L., Chou, K. H., Bandettini, P., Lin, C., & Alzheimer's Disease Neuroimaging Initiative (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *Neuroimage*, 60(1), 59-70.
- Cleary, J. G., & Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. In *Machine Learning Proceedings 1995* (pp. 108-114). Morgan Kaufmann.
- Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20(1), 37-46.
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7), 27–29. <https://doi.org/10.1.1.110.9154>
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1(1), 16-29. <https://doi.org/10.1037/1082-989X.1.1.16>
- Efron, B. (1983). Estimating the error rate of a prediction rule: Improvements on crossvalidation. *J. Amer. Stat. Ass.*, 78, 316–331.
- Egan, J. P. (1975). *Signal detection theory and ROC analysis*. Academic Press.

- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(1), 8.
- Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte, NC: IAP.
- Fischer, R., & Alfons Karl, J. (2020). The network architecture of individual differences: Personality, reward-sensitivity, and values. *Personality and Individual Differences*, 160(February), 109922. <https://doi.org/10.1016/j.paid.2020.109922>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(3), 286–299. <https://doi.org/10.1037/1040-3590.7.3.286>
- Golino, H. F., & Christensen, A. P. (2020). *EGAnet: Exploratory Graph Analysis -- A framework for estimating the number of dimensions in multivariate data using network psychometrics*. Retrieved from <https://CRAN.R-project.org/package=EGAnet>
- 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), 1–26. <https://doi.org/10.1371/journal.pone.0174035>
- Golino, H. F., Moulder, R., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., ... Boker, S. M. (2020). Entropy fit indices: New fit measures for assessing the structure and dimensionality of multiple latent variables. *Multivariate Behavioral Research*, 1–29. <https://doi.org/10.1080/00273171.2020.1779642>
- Golino, H. F., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., ... 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>
- Goretzko, D., & Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. *Psychological Methods*, 25(6), 776–786. <https://doi.org/10.1037/met0000262>
- Gorsuch, R. L. (1974). *Factor analysis*. W. B. Saunders.
- Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265–275.
- Guess, A., Munger, K., Nagler, J., & Tucker, J. (2019). How accurate are survey responses on social media and politics?. *Political Communication*, 36(2), 241–258.
- Güre, Ö. B., Kayri, M., & Erdoğan, F. (2020). Analysis of factors effecting PISA 2015 mathematics literacy via educational machine learning. *Education and Science*, 45(202), 393–415.
- Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS, Political Science & Politics*, 48(1), 80.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Peter, R., & Witten, I. H. (2009). The WEKA machine learning software: An update. *SIGKDD Explorations*, 11(1), 10–18.
- Hamalainen, W., & Vinni, M. (2006). Comparison of machine learning methods for intelligent tutoring systems. In *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 525–534). Springer Berlin/Heidelberg.
- Han, J., J. Pei, & Kamber, M. (2011). *Machine learning: Concepts and techniques*. Elsevier.

- Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 10(1), 131-156. <https://doi.org/10.1901/jaba.1977.10-103>
- Hegde, J., & Rokseth, B. (2020). Applications of machine learning methods for engineering risk assessment—A review. *Safety Science*, 122, 104492.
- Heydari, S. S., & Mountrakis, G. (2018). Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sensing of Environment*, 204, 648-658.
- 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>
- Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, 32(1), 51-62. <https://doi.org/10.1080/10447318.2015.1087664>
- Iantovics, L. B., Rotar, C., & Morar, F. (2019). Survey on establishing the optimal number of factors in exploratory factor analysis applied to machine learning. *Wiley Interdisciplinary Reviews: Machine learning and Knowledge Discovery*, 9(2), 1-20. <https://doi.org/10.1002/widm.1294>
- Ibarguren, I., Pérez, J. M., Muguerza, J., Gurrutxaga, I., & Arbelaitz, O. (2015). Coverage-based resampling: Building robust consolidated decision trees. *Knowledge-Based Systems*, 79, 51-67. <https://doi.org/10.1016/j.knsys.2014.12.023>
- John, G. H., & Langley P. (1995). Estimating continuous distributions in Bayesian classifiers. In P. Besnard & S. Hanks (Eds.), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 338–345). San Francisco, Morgan Kaufmann.
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science*, 349(6245), 255-260, <https://doi.org/10.1126/science.aaa8415>
- 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>
- Kılıç, A. F., & Koyuncu, İ. (2017). Ölçek uyarlama çalışmalarının yapı geçerliği açısından incelenmesi [Examination of scale adaptation studies in terms of construct validity]. In Ö. Demirel & S. Dinçer (Eds.), *Küreselleşen dünyada eğitim [Education in a globalizing world]* (pp. 1202–1205). Pegem Akademi.
- Kjellström, S., & Golino, H. (2019). Mining concepts of health responsibility using text mining and exploratory graph analysis. *Scandinavian Journal of Occupational Therapy*, 26(6), 395–410. <https://doi.org/10.1080/11038128.2018.1455896>
- Kline, P. (1994). *An easy guide to factor analysis*. Routledge.
- Koyuncu, İ., & Gelbal, S. (2020). Comparison of machine learning classification algorithms on educational data under different conditions. *Journal of Measurement and Evaluation in Education and Psychology*, 11(4), 325-345.
- Koyuncu, İ., & Kılıç, A. F. (2019). The use of exploratory and confirmatory factor analyses: A document analysis. *Education and Science*, 44(198), 361-388. <https://doi.org/10.15390/EB.2019.7665>
- Kuhn, M. (2020). *caret: Classification and Regression Training*. Retrieved from <https://cran.r-project.org/package=caret>
- Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10(1), 1-11.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Landwehr, N., Hall, M., & Frank, E. (2006). *Logistic model trees*. Kluwer Academic Publishers.

- Larose, D. T., & Larose, C.D. (2014). *Discovering knowledge in data: An introduction to machine learning*. John Wiley and Sons.
- Li, C.-H. (2016a). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936–949. <https://doi.org/10.3758/s13428-015-0619-7>
- Li, C.-H. (2016b). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, 21(3), 369–387. <https://doi.org/10.1037/met0000093>
- Li, N., Qi, J., Wang, P., Zhang, X., Zhang, T., & Li, H. (2019). Quantitative structure–activity relationship (QSAR) study of carcinogenicity of polycyclic aromatic hydrocarbons (PAHs) in atmospheric particulate matter by random forest (RF). *Analytical Methods*, 11(13), 1816–1821.
- Mele, M., & Magazzino, C. (2020). A machine learning analysis of the relationship among iron and steel industries, air pollution, and economic growth in China. *Journal of Cleaner Production*, 277, 123293.
- Minaei-Bidgoli, B., D.A. Kashy, G. Kortemeyer, & W. Punch (2003). Predicting student performance: An application of machine learning methods with an educational web-based system. In *Proceedings of 33rd Frontiers in Education Conference*, (pp. 13–18). Westminster, CO.
- Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.
- Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, (pp. T2G–7). IEEE.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd. ed.). McGraw-Hill.
- Osborne, J. W. (2015). What is rotating in exploratory factor analysis? *Practical Assessment Research & Evaluation*, 20(2), 1–7.
- Panayiotou, M., Santos, J., Black, L., & Humphrey, N. (2020). Exploring the dimensionality of the social skills improvement system using exploratory graph analysis and bifactor-(S–1) modeling. *Assessment*, 1–15. <https://doi.org/10.1177/1073191120971351>
- Pérez, J. M., Muguerza, J., Arbelaitz, O., Gurrutxaga, I., & Martín, J. I. (2007). Combining multiple class distribution modified subsamples in a single tree. *Pattern Recognition Letters*, 28(4), 414–422. <https://doi.org/10.1016/j.patrec.2006.08.013>
- Pu, Y., Apel, D. B., & Hall, R. (2020). Using machine learning approach for microseismic events recognition in underground excavations: Comparison of ten frequently-used models. *Engineering Geology*, 268, 105519.
- Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers, Inc.
- R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from <https://www.r-project.org/>
- Reich, Y., & Barai, S. V. (1999). Evaluating machine learning models for engineering problems. *Artificial Intelligence in Engineering*, 13(3), 257–272.
- Rijsbergen CV. (1979). *Information retrieval* (2nd ed.). Butterworth.
- Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, 21(1), 135–146.
- Romero, C., & Ventura, S. (2013). Machine learning in education. *WIREs Machine Learning Knowledge Discovery* 3(1), 12–27.

- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36.
- Shao, L., Fan, X., Cheng, N., Wu, L., & Cheng, Y. (2013). Determination of minimum training sample size for microarray-based cancer outcome prediction—an empirical assessment. *PloS one*, 8(7), e68579. <https://doi.org/10.1371/journal.pone.0068579>
- Sumner, M., Frank, E., & Hall, M. (2005, October). Speeding up logistic model tree induction. In *European conference on principles of machine learning and knowledge discovery* (pp. 675-683). Springer, Berlin, Heidelberg.
- Sun, Y., Kamel, M. S., & Wang, Y. (2006). Boosting for learning multiple classes with imbalanced class distribution. In *Sixth international conference on data mining (ICDM'06)* (pp. 592-602). IEEE.
- Tabachnik, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Pearson.
- Tezbaşaran, E., & Gelbal, S. (2018). Temel bileşenler analizi ve yapay sinir ağı modellerinin ölçek geliştirme sürecinde kullanılabilirliğinin incelenmesi [An investigation on usability of principal component analysis and artificial neural network models in the process of scale development]. *Mersin University Journal of the Faculty of Education*, 14(1), 225-252.
- 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>
- West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56-75). Sage.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Machine learning: Practical machine learning tools and techniques* (4th Edition). Morgan Kaufmann.
- Wojna, A., Latkowski, R. (2018): Rseslib 3: Open source library of rough set and machine learning methods. In: *Proceedings of the International Joint Conference on Rough Set (LNCS*, vol. 11103, pp. 162-176). Springer.
- Wojna, A., Latkowski, R., Kowalski, (2019). *RSESLIB: User guide*. Retrieved from <http://rseplib.mimuw.edu.pl/rseplib.pdf>
- Zhang, F., & Yang, X. (2020). Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection. *Remote Sensing of Environment*, 251, 112105. <https://doi.org/10.1016/j.rse.2020.112105>

6. APPENDICES

APPENDIX A

Percent correct values of the methods.

Skewness	Items per Factor	Methods	Sample Size																
			100			200			500			1000							
			Average Factor Loadings						Number of Factors										
			0.40		0.70		0.40		0.70		0.40		0.70		0.40		0.70		
2	3	2	3	2	3	2	3	2	3	2	3	2	3	2	3				
Left Skewed	5 Items	BayesNet	46.70	25.25	89.30	81.30	59.60	39.05	96.50	91.05	71.10	49.00	99.60	98.90	78.60	54.85	100	99.90	
		EGA	9.00	3.00	83.00	47.00	21.00	4.00	98.00	93.00	45.00	8.00	100	100	77.00	55.00	100	99.00	
		J48Consolidated	56.80	40.50	71.80	62.05	58.10	48.65	72.90	63.25	57.80	46.60	70.20	66.20	60.60	46.05	70.50	64.25	
		KStar	61.90	47.50	94.10	89.30	70.80	52.55	98.70	97.00	78.50	71.55	100	99.70	50.00	25.00	51.60	25.45	
		NaiveBayes	59.20	47.60	90.40	88.55	61.40	50.10	96.00	94.40	65.00	62.00	99.80	99.10	74.70	70.90	100	100	
	10 Items	RandomForest	45.20	40.15	89.20	87.30	44.80	38.45	94.80	93.00	47.60	42.00	97.70	97.45	46.80	44.40	98.90	98.75	
		RBFNetwork	56.80	42.05	82.30	75.45	57.10	42.65	87.90	81.90	57.20	49.90	96.10	93.05	64.60	55.05	98.80	98.55	
		RseslibKnn	52.60	39.75	93.60	89.45	60.10	41.80	98.50	95.80	63.50	56.50	100	99.95	78.70	66.70	100	100	
		SimpleLogistic	38.90	40.40	69.50	75.35	42.80	43.75	68.20	76.75	42.40	44.65	62.90	75.45	40.50	43.45	56.40	75.50	
		BayesNet	60.35	40.93	93.75	88.43	68.00	49.27	99.00	97.37	72.65	60.07	100	99.90	83.60	68.30	100	100	
	5 Items	EGA	1.00	0.00	69.00	41.00	7.00	1.00	91.00	85.00	40.00	10.00	99.00	98.00	70.00	62.00	100	100	
		J48Consolidated	64.70	47.60	82.00	71.47	64.15	49.47	81.40	73.07	63.55	51.60	81.10	71.70	64.90	49.83	81.80	70.43	
		KStar	68.45	53.80	95.70	91.00	75.80	62.10	99.30	97.60	86.00	76.13	99.95	99.80	50.00	33.33	53.00	36.33	
		NaiveBayes	72.75	62.70	97.10	94.33	81.55	70.60	99.40	98.70	91.10	86.37	99.95	99.97	97.10	94.73	100	100	
		RandomForest	78.60	66.57	97.95	96.40	85.65	78.13	99.70	99.37	91.40	86.17	99.90	99.93	93.65	88.67	100	100	
	Normal	10 Items	RBFNetwork	63.05	50.93	87.25	78.33	65.95	54.33	91.65	84.27	73.40	61.77	96.70	93.13	80.15	68.83	99.45	97.67
			RseslibKnn	69.15	55.30	96.05	91.00	78.95	65.17	99.35	98.00	87.40	78.30	100	99.87	93.75	88.70	100	100
			SimpleLogistic	68.30	59.97	87.10	88.60	70.65	67.67	89.85	91.10	75.25	71.00	90.65	93.40	76.60	74.70	91.95	93.83
			BayesNet	57.90	44.30	97.50	94.75	61.80	53.70	99.50	99.00	73.60	62.20	100	100	83.40	74.85	100	100
			EGA	21.00	6.00	99.00	95.00	55.00	23.00	100	98.00	89.00	79.00	100	100	100	97.00	100	100
5 Items		J48Consolidated	60.20	46.20	75.50	64.55	56.90	47.95	71.90	63.15	61.60	45.15	76.70	60.50	58.40	44.65	74.50	57.60	
		KStar	74.50	63.25	99.50	99.60	83.60	77.90	100	100	50.00	25.00	81.40	69.20	50.00	25.00	50.00	25.00	
		NaiveBayes	45.60	39.00	85.20	80.75	43.10	33.80	89.30	83.85	37.50	28.75	96.20	95.20	28.90	23.95	98.20	98.40	
		RandomForest	56.80	48.70	96.50	97.15	59.40	54.15	99.30	99.00	61.40	57.50	99.80	99.55	64.10	60.30	99.80	100	
		RBFNetwork	32.60	22.15	48.20	43.80	25.50	18.20	52.10	47.35	21.00	12.20	53.00	54.75	12.30	8.10	56.30	53.50	
10 Items		RseslibKnn	73.80	62.45	99.00	99.45	84.50	79.65	100	100	96.10	93.00	100	100	98.70	98.60	100	100	
		SimpleLogistic	50.00	52.15	78.10	82.80	47.90	54.30	77.10	84.45	55.40	51.75	71.80	87.25	50.50	55.20	68.90	86.90	
		BayesNet	68.55	49.83	99.20	98.20	77.50	59.73	99.80	99.93	84.35	73.87	100	100	94.10	84.07	100	100	
		EGA	12.00	3.00	100	93.00	38.00	23.00	100	98.00	76.00	81.00	100	100	93.00	99.00	100	100	
		J48Consolidated	62.90	48.50	82.45	74.10	66.55	49.47	80.30	72.27	63.95	50.43	79.50	71.03	66.55	50.40	79.10	69.00	
Right Skewed		10 Items	KStar	80.00	70.57	100	99.50	90.30	83.13	100	99.97	50.00	33.33	82.60	79.23	50.00	33.33	50.00	33.33
			NaiveBayes	89.80	83.63	99.45	99.27	95.85	92.57	99.20	98.70	97.85	97.67	100	100	98.10	97.97	100	100
			RandomForest	85.60	77.10	99.70	99.40	91.70	85.33	99.85	100	95.00	90.53	100	99.97	96.15	92.90	100	100
			RBFNetwork	80.85	71.87	94.25	95.00	85.40	75.97	94.40	95.80	88.25	83.00	96.65	98.07	89.10	84.00	98.65	99.33
			RseslibKnn	80.85	73.57	99.85	99.53	91.00	84.57	99.95	99.97	97.65	96.33	100	100	99.75	99.57	100	100
	5 Items	SimpleLogistic	76.70	72.73	92.30	94.57	79.15	78.13	94.15	95.13	79.15	84.07	95.45	96.30	82.30	83.57	94.75	96.87	
		BayesNet	52.80	28.40	87.80	75.40	63.80	42.60	95.40	91.60	69.20	48.10	99.70	98.85	75.90	58.90	100	100	
		EGA	16.00	5.00	88.00	59.00	25.00	6.00	99.00	90.00	41.00	13.00	100	100	92.00	59.00	100	100	
		J48Consolidated	58.60	43.85	76.50	65.40	59.30	42.15	72.60	66.55	58.80	47.05	75.40	67.00	62.00	47.45	75.90	63.50	
		KStar	62.50	48.00	92.40	87.70	69.20	55.25	98.30	97.80	78.90	70.55	100	100	50.00	25.00	50.20	25.00	
	10 Items	NaiveBayes	58.80	46.20	91.40	83.80	64.30	49.25	96.40	94.65	63.70	58.25	99.80	99.40	67.20	62.35	99.60	99.60	
		RandomForest	53.60	42.15	90.00	87.95	52.80	46.30	94.90	94.95	52.00	47.75	98.70	99.30	52.60	49.55	99.60	99.40	
		RBFNetwork	55.90	41.70	82.40	70.55	59.30	42.20	86.10	83.85	58.90	48.75	96.50	93.50	66.20	55.35	98.90	98.00	
		RseslibKnn	54.10	39.60	92.90	87.15	60.70	50.50	98.70	97.00	63.90	56.75	100	100	77.60	70.35	100	100	
		SimpleLogistic	46.40	41.45	67.60	72.75	43.80	44.50	65.40	76.25	40.60	45.55	63.40	79.50	39.10	48.85	61.70	76.15	
	5 Items	BayesNet	60.55	42.57	95.25	91.07	65.70	51.00	98.30	97.63	77.10	58.80	99.90	99.80	81.35	69.30	100	100	
		EGA	2.00	0.00	77.00	52.00	5.00	2.00	95.00	88.00	45.00	21.00	100	98.00	82.00	64.00	100	100	
		J48Consolidated	64.25	48.43	82.80	73.70	63.65	49.43	83.80	72.90	64.50	50.00	81.70	72.20	63.15	48.87	80.55	72.60	
		KStar	68.40	54.70	95.10	91.50	75.20	62.00	99.10	97.63	85.55	75.53	99.95	99.87	50.00	33.33	50.05	33.40	
		NaiveBayes	74.20	62.93	95.20	92.73	79.40	69.30	98.05	96.93	89.70	82.00	100	99.77	93.80	89.43	100	99.93	
10 Items	RandomForest	79.80	70.00	98.40	97.07	84.90	77.47	99.65	99.27	92.50	85.57	100	99.93	94.05	90.10	100	100		
	RBFNetwork	65.25	53.63	87.50	80.93	67.45	53.40	89.25	84.77	71.10	61.50	95.90	92.40	79.60	66.43	98.95	97.57		
	RseslibKnn	70.70	55.30	96.00	92.97	76.65	65.50	99.40	98.30	87.35	80.33	100	99.87	93.40	89.53	100	100		
	SimpleLogistic	69.15	60.40	87.55	89.67	70.30	65.90	90.90	91.73	74.15	69.33	91.40	92.80	74.30	72.20	91.90	94.80		

APPENDIX B

Kappa concordance values of the methods

Skewness	Items per Factor	Methods	Sample Size																
			100			200			500			1000							
			Average Factor Loadings																
			0.40		0.70		0.40		0.70		0.40		0.70		0.40		0.70		
Number of Factors																			
		2		3		2		3		2		3		2		3			
Left Skewed	5 Items	BayesNet	0.47	0.15	0.89	0.76	0.60	0.25	0.97	0.88	0.71	0.36	1.00	0.99	0.79	0.43	1.00	1.00	
		EGA	0.74	0.72	0.97	0.91	0.77	0.70	1.00	0.99	0.89	0.78	1.00	1.00	0.97	0.92	1.00	1.00	
		J48Consolidated	0.57	0.25	0.72	0.52	0.58	0.33	0.73	0.53	0.58	0.32	0.70	0.56	0.61	0.31	0.71	0.55	
		KStar	0.62	0.35	0.94	0.87	0.71	0.40	0.99	0.96	0.78	0.62	1.00	1.00	0.50	0.00	0.52	0.01	
		NaiveBayes	0.59	0.36	0.90	0.86	0.61	0.40	0.96	0.93	0.65	0.55	1.00	0.99	0.75	0.66	1.00	1.00	
	10 Items	RandomForest	0.45	0.31	0.89	0.85	0.45	0.28	0.95	0.92	0.48	0.33	0.98	0.97	0.47	0.37	0.99	0.99	
		RBFNetwork	0.57	0.29	0.82	0.70	0.57	0.31	0.88	0.77	0.57	0.39	0.96	0.92	0.65	0.46	0.99	0.98	
		RseslibKnn	0.53	0.28	0.94	0.87	0.60	0.29	0.98	0.95	0.64	0.46	1.00	1.00	0.79	0.57	1.00	1.00	
		SimpleLogistic	0.39	0.27	0.70	0.69	0.43	0.31	0.68	0.71	0.42	0.33	0.63	0.69	0.41	0.32	0.56	0.69	
		BayesNet	0.21	0.11	0.88	0.83	0.36	0.24	0.98	0.96	0.45	0.40	1.00	1.00	0.67	0.52	1.00	1.00	
	Normal	5 Items	EGA	0.85	0.72	0.98	0.94	0.89	0.73	0.99	0.99	0.96	0.88	1.00	1.00	0.98	0.97	1.00	1.00
			J48Consolidated	0.29	0.21	0.64	0.57	0.28	0.24	0.63	0.60	0.27	0.27	0.62	0.58	0.30	0.25	0.64	0.56
			KStar	0.37	0.31	0.91	0.87	0.52	0.43	0.99	0.96	0.72	0.64	1.00	1.00	0.00	0.00	0.06	0.05
			NaiveBayes	0.46	0.44	0.94	0.91	0.63	0.56	0.99	0.98	0.82	0.80	1.00	1.00	0.94	0.92	1.00	1.00
			RandomForest	0.57	0.50	0.96	0.95	0.71	0.67	0.99	0.99	0.83	0.79	1.00	1.00	0.87	0.83	1.00	1.00
10 Items		RBFNetwork	0.26	0.26	0.74	0.67	0.32	0.32	0.83	0.76	0.47	0.43	0.93	0.90	0.60	0.53	0.99	0.97	
		RseslibKnn	0.38	0.33	0.92	0.86	0.58	0.48	0.99	0.97	0.75	0.67	1.00	1.00	0.88	0.83	1.00	1.00	
		SimpleLogistic	0.37	0.40	0.74	0.83	0.41	0.52	0.80	0.87	0.50	0.56	0.81	0.90	0.53	0.62	0.84	0.91	
		BayesNet	0.58	0.31	0.98	0.93	0.62	0.39	0.99	0.99	0.74	0.52	1.00	1.00	0.83	0.68	1.00	1.00	
		EGA	0.80	0.74	1.00	0.99	0.91	0.84	1.00	1.00	0.98	0.97	1.00	1.00	1.00	1.00	1.00	1.00	
Right Skewed		5 Items	J48Consolidated	0.60	0.32	0.75	0.54	0.57	0.34	0.72	0.53	0.62	0.30	0.77	0.49	0.58	0.30	0.74	0.46
			KStar	0.74	0.53	1.00	0.99	0.84	0.70	1.00	1.00	0.50	0.00	0.81	0.63	0.50	0.00	0.50	0.00
			NaiveBayes	0.46	0.30	0.85	0.77	0.43	0.25	0.89	0.82	0.38	0.22	0.96	0.95	0.29	0.19	0.98	0.98
			RandomForest	0.57	0.39	0.97	0.96	0.59	0.46	0.99	0.99	0.61	0.50	1.00	1.00	0.64	0.53	1.00	1.00
			RBFNetwork	0.33	0.13	0.48	0.37	0.25	0.09	0.52	0.42	0.21	0.07	0.53	0.51	0.12	0.05	0.56	0.51
	10 Items	RseslibKnn	0.74	0.52	0.99	0.99	0.84	0.72	1.00	1.00	0.96	0.90	1.00	1.00	0.99	0.98	1.00	1.00	
		SimpleLogistic	0.50	0.39	0.78	0.78	0.48	0.41	0.77	0.80	0.55	0.40	0.72	0.83	0.51	0.43	0.69	0.83	
		BayesNet	0.37	0.25	0.98	0.97	0.55	0.40	1.00	1.00	0.69	0.61	1.00	1.00	0.88	0.76	1.00	1.00	
		EGA	0.93	0.80	1.00	0.99	0.94	0.91	1.00	1.00	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00	
		J48Consolidated	0.26	0.23	0.65	0.61	0.33	0.24	0.61	0.58	0.28	0.26	0.59	0.57	0.33	0.26	0.58	0.53	
	10 Items	KStar	0.60	0.56	1.00	0.99	0.81	0.75	1.00	1.00	0.00	0.00	0.65	0.69	0.00	0.00	0.00	0.00	
		NaiveBayes	0.80	0.75	0.99	0.99	0.92	0.89	0.98	0.98	0.96	0.97	1.00	1.00	0.96	0.97	1.00	1.00	
		RandomForest	0.71	0.66	0.99	0.99	0.83	0.78	1.00	1.00	0.90	0.86	1.00	1.00	0.92	0.89	1.00	1.00	
		RBFNetwork	0.62	0.58	0.89	0.92	0.71	0.64	0.89	0.94	0.76	0.75	0.93	0.97	0.78	0.76	0.97	0.99	
		RseslibKnn	0.62	0.60	1.00	0.99	0.82	0.77	1.00	1.00	0.95	0.94	1.00	1.00	0.99	0.99	1.00	1.00	
SimpleLogistic		0.53	0.59	0.85	0.92	0.58	0.67	0.88	0.93	0.58	0.76	0.91	0.94	0.65	0.75	0.90	0.95		
BayesNet		0.53	0.17	0.88	0.69	0.64	0.28	0.95	0.89	0.69	0.35	1.00	0.99	0.76	0.47	1.00	1.00		
EGA		0.69	0.76	0.98	0.92	0.79	0.73	1.00	0.99	0.89	0.83	1.00	1.00	0.99	0.93	1.00	1.00		
J48Consolidated		0.59	0.29	0.76	0.56	0.59	0.27	0.73	0.57	0.59	0.34	0.75	0.58	0.62	0.33	0.76	0.53		
KStar		0.63	0.35	0.92	0.84	0.69	0.44	0.98	0.97	0.79	0.62	1.00	1.00	0.50	0.00	0.50	0.00		
10 Items	NaiveBayes	0.59	0.35	0.91	0.80	0.64	0.39	0.96	0.94	0.64	0.51	1.00	0.99	0.67	0.56	1.00	0.99		
	RandomForest	0.54	0.32	0.90	0.86	0.53	0.36	0.95	0.94	0.52	0.39	0.99	0.99	0.53	0.41	1.00	0.99		
	RBFNetwork	0.56	0.29	0.82	0.64	0.59	0.30	0.86	0.80	0.59	0.38	0.97	0.92	0.66	0.47	0.99	0.97		
	RseslibKnn	0.54	0.27	0.93	0.84	0.61	0.40	0.99	0.96	0.64	0.46	1.00	1.00	0.78	0.62	1.00	1.00		
	SimpleLogistic	0.46	0.29	0.68	0.66	0.44	0.32	0.65	0.70	0.41	0.33	0.63	0.74	0.39	0.36	0.62	0.70		
	BayesNet	0.21	0.14	0.90	0.87	0.31	0.26	0.97	0.96	0.54	0.38	1.00	1.00	0.63	0.54	1.00	1.00		
	EGA	0.85	0.70	0.99	0.96	0.88	0.78	1.00	0.99	0.96	0.90	1.00	1.00	0.99	0.97	1.00	1.00		
	J48Consolidated	0.28	0.23	0.66	0.61	0.27	0.24	0.68	0.59	0.29	0.25	0.63	0.58	0.26	0.23	0.61	0.59		
	KStar	0.37	0.32	0.90	0.87	0.50	0.43	0.98	0.96	0.71	0.63	1.00	1.00	0.00	0.00	0.00	0.00		
	NaiveBayes	0.48	0.44	0.90	0.89	0.59	0.54	0.96	0.95	0.79	0.73	1.00	1.00	0.88	0.84	1.00	1.00		
10 Items	RandomForest	0.60	0.55	0.97	0.96	0.70	0.66	0.99	0.99	0.85	0.78	1.00	1.00	0.88	0.85	1.00	1.00		
	RBFNetwork	0.30	0.30	0.75	0.71	0.35	0.30	0.79	0.77	0.42	0.42	0.92	0.89	0.59	0.50	0.98	0.96		
	RseslibKnn	0.41	0.33	0.92	0.89	0.53	0.48	0.99	0.97	0.75	0.70	1.00	1.00	0.87	0.84	1.00	1.00		
	SimpleLogistic	0.38	0.41	0.75	0.84	0.41	0.49	0.82	0.88	0.48	0.54	0.83	0.89	0.49	0.58	0.84	0.92		

Examining the Discrimination of Binary Scored Test Items with ROC Analysis

Sait Cum ^{1,*}

¹Ministry of National Education, Izmir Provincial Directorate, Turkey

ARTICLE HISTORY

Received: Mar. 11, 2021

Revised: Oct. 06, 2021

Accepted: Nov. 08, 2021

Keywords:

Item analysis,
Discrimination,
Roc analysis,
Test development,
Psychometrics.

Abstract: In this study, it was claimed that ROC analysis, which is used to determine to what extent medical diagnosis tests can be differentiated between patients and non-patients, can also be used to examine the discrimination of binary scored items in cognitive tests. In order to obtain various evidence for this claim, the 2x2 contingency table used in the ROC analysis was adapted in accordance with the logic of item discrimination. It was suggested in the article that the areas under the ROC curves (AUC) obtained by using the sensitivity and specificity values calculated with the adapted contingency table can be considered as a measure of item discrimination. The results of the statistical analyses made on the simulation data showed that the AUC values were positively and highly correlated with the D , r_{bis} and a parameter values of the items, and the AUC values from different sized samples were consistent. Additionally, ROC analysis was more stable against range narrowing than other methods. In this respect, it was concluded that very large groups were not needed to examine item discrimination with the proposed method.

1. INTRODUCTION

Osterlind (1990) defined the test item as a unit of measurement that includes a stimulus and a prescriptive response form created to examine mental attributes. Test items provide inferences about a number of psychological and cognitive structures related to the knowledge, ability, or personal characteristics of respondents based on their performance. In binary scored items, the value of “1” indicates that the item is answered correctly, and “0” indicates that the item is answered incorrectly or is left blank. These kinds of items are frequently encountered in achievement and ability tests, in which it is aimed to measure maximum performance. When writing items for purposes such as test development or item pooling, it is necessary to determine the psychometric properties of the items, which is important to obtain valid and reliable measurements. Psychometric properties that provide information about the aspects of items such as difficulty, discrimination, and probability of the item to be answered correctly with chance can be predicted by various statistical or mathematical techniques. Decisions on the using of the item in the test can be made based on these properties. The individuals who possess the knowledge/skill measured by an item are expected to answer that item correctly, and the

*CONTACT: Sait Cum  saitcum@hotmail.com  Ministry of National Education, Izmir Provincial Directorate, Turkey

individuals who don't have that knowledge are likely to respond to that item incorrectly. The power of the item to separate these two groups is defined as the item discrimination. It can be stated that the measurement can achieve its goal as long as the groups are distinguished from each other. It can be emphasized that discrimination is an important item characteristic since if the measurement reaches its purpose, it is valid.

Prediction to determine the discrimination of the items can be made by various methods. It can be said that the commonly used classical approaches are to determine the biserial correlation coefficients between the item-total test scores (r_{bis}) and to determine the difference between the correct response rate in the upper group and the correct response rate in the lower group (D). These values, which are determined through these approaches and range from -1 to 1, are called item discrimination index.

The item discrimination index, which is calculated over small groups or homogeneous groups according to the ability levels of individuals, can provide misleading information due to the range narrowing (Ebel & Frisbie, 1991; Fulcher & Davidson, 2007). In this respect, Çüm, Gelbal and Tsai (2016) found in their study that item discrimination indexes calculated with the biserial correlation coefficient method showed considerable differences between different small samples.

Discrimination of items can also be examined based on the Item Response Theory (IRT). In IRT, the slope of the item characteristic curve is accepted as the item discrimination parameter (a parameter). Although it is stated that the parameter value theoretically changes in the $-\infty$ and $+\infty$ ranges, it usually takes the values between 0 and 2 (or 0 and 3) (DeMars, 2016; Hambleton & Swaminathan, 1985).

In this study, it was claimed that item discrimination can also be examined with the ROC analysis (Receiver Operating Characteristic Analysis). In this respect, a discrimination prediction method, which is not included in the literature, was discussed for the first time in the study.

ROC analysis (Receiver Operating Characteristic Analysis) provides the opportunity to evaluate the performance of medical tests, statistical classifiers, prediction models and algorithms. With the ROC curves created within the scope of the analysis, a graph showing the discrimination performance of a medical diagnostic test (0 = *no disease*, 1 = *disease*) is obtained (Zou et al., 2012). ROC curves are images created to summarize the accuracy of diagnostic predictions (1-0) and they can be used regardless of the source of these predictions. In addition, by comparing the generated ROC curves, the accuracy of different methods used for predictions might be compared (Gönen, 2007). ROC analysis is based on a 2x2 contingency table (Table 1).

Table 1. The basis of the ROC analysis is a 2x2 contingency table.

		True Status (Gold Standard)	
		Positive (+)	Negative (-)
Test Result (Result of Diagnosis)	Positive (+)	True Positive (TP)	False Positive (FP)
	Negative (-)	False Negative (FN)	True Negative (TN)
Total		TP+FN	FP+TN

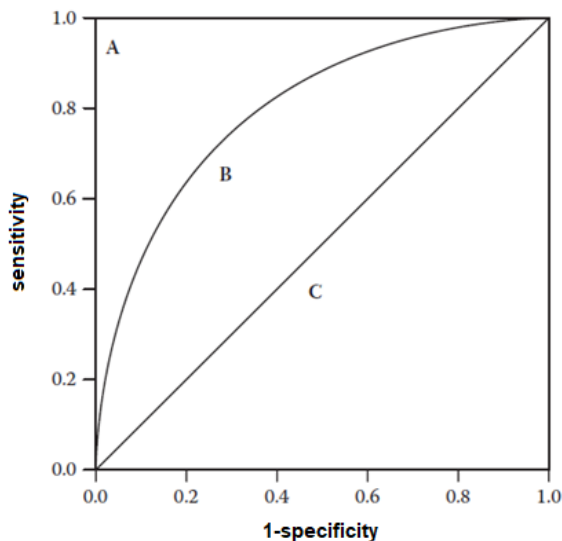
The variable given as the “test result” in the table, for example, refers to the decisions (the result of diagnosis) based on the scores (values) obtained as a result of a medical test whose effectiveness is examined. The values obtained from the result of the test are included in the

analysis as a continuous variable. This continuous variable is then transformed into a two-category (positive, negative) variable, which separates values above and below a given cut-off score. The values in the table will change when the cut-off score changes. In ROC analysis, all possible cut-off points can be tested and optimal cut-off score can be determined by various statistical techniques.

The variable given as a true status in the table is a two-category variable obtained in usually from a more reliable reference test as a result of clinical follow-up or decisions made by a gold standard council, and it separates people into who are really positive or negative. Based on this 2x2 table, two important parameters are explained. The first one is the probability of the diagnostic test to classify a healthy person (negative) as healthy, namely specificity; secondly, the probability of the test correctly classifying a patient person (positive) as a patient is sensitivity (Alonzo & Pepe, 2002; Krzanowski & Hand, 2009; Ruopp et al., 2008; Zou et al., 2012).

Considering the change of TP, FN, TN, FP values for each possible cut-off point, the sensitivity is calculated as $TP / (TP + FN)$, and the specificity is calculated as $TN / (FP + TN)$. The ROC curve is the graph obtained from the pairs of sensitivity and 1-specificity calculated from each of the possible cut-off points (Zou et al., 2012).

Figure 1. ROC curve.



In Figure 1, the ROC curve (B) is shown at a location in the area between point A and reference axis C. It can be stated that the diagnostic test distinguishes patients and non-patients as well as the ROC curve converges to point A. The C axis is obtained by connecting the points representing the randomness of this distinction. As the curve gets closer to this axis, the discriminative effectiveness of the test decreases. It can be stated that the area under the curve (AUC) is the measure that is generally used to summarize the analysis and provides the opportunity to evaluate the effectiveness of the test. Since the area under the curve will be equal to the area of the square when the curve reaches point A, the AUC value is maximum 1; When the curve coincides with the C axis, the area under curve will be equal to the area of the triangle, so, the AUC value will be 0.5 and this value expresses the randomness in identifying individuals (Krzanowski & Hand, 2009; van Erkel & Pattynama, 1998).

The method proposed in this study was started with the adaptation of the 2x2 contingency table that was taken as basis in the analysis in order to determine the item discrimination with ROC analysis and to use AUC values as the item discrimination measure. For this purpose, the modifications made on contingency table were shown in Table 2.

Table 2. Contingency table adapted to predict item discriminations.

		Item Score	
		True (1)	False (0)
Test Total Score	High-scoring Group	High-scoring Group True (HGT)	High-scoring Group False (HGF)
	Low-scoring Group	Low-scoring Group True (LGT)	Low-scoring Group False (LGF)
Total		HGT+LGT	HGF+LGF

It can be said that a test item is discriminative to the extent that it can distinguish between individuals who have the attribute measured by item and those who do not. Based on the assumption that the item and the test measure the same attribute, in other words, the test is one-dimensional, individuals with high test total scores are expected to answer the item correctly, and individuals with low test total scores are expected to answer the item incorrectly. This underlies the logic of discrimination prediction based on internal criteria. The proposed method also provides an opportunity to examine the discrimination based on optimal internal criteria. The high and low scoring groups mentioned in the table are not the groups consisting of a definite and fixed number of individuals. Some individuals in these groups move to the other group at each cut-off score tested. The combination of these two groups forms the whole group in each case. In the adapted contingency table, the number of individuals in the high-scoring group who answered the item correctly to the HGT section, the number of individuals who answered the item incorrectly to the HGF section; the number of those who are in the low-scoring and who answered the item correctly is written in the LGT section, and the number of those who answered the item incorrectly is written in the LGF section by trying all possible cut-off points.

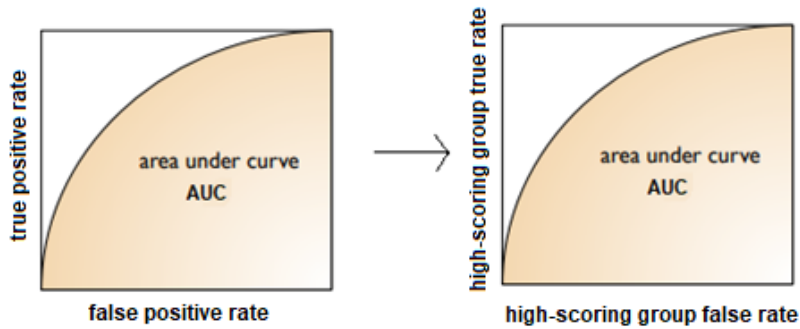
In this case, the sensitivity value gives the probability of an individual in the high-scoring group to answer the item correctly and is calculated as $HGT / (HGT + LGT)$. The specificity value gives the possibility of an individual in the low-scoring group to answer the item incorrectly and is calculated as $LGF / (HGF + LGF)$. The ROC curve, which will describe the item discrimination, is formed from the sensitivity and 1-specificity pairs obtained in the context of all possible cut-off points in accordance with the original analysis. The area under the curve (AUC) determines a measure of the item's discrimination. It is expected that the number of individuals who answered the item correctly from the high-scoring group will increase and the number of those who answered the item from the low-scoring group incorrectly will increase as close to the optimal cut-off score. The approach chosen to determine possible cut-off scores does not affect the basic logic of the analysis. For example, for a 10-item test, the starting point of the cut-off points is 1 point less than the score of the respondent who received the lowest score from the test; the endpoint can be determined to be 1 point more than the score of the respondent who got the highest score from the test. The cut-off points between them can be calculated as the average of each consecutive score pair. In the case where the lowest score obtained from the 10-item test exemplified is 1 and the highest score is 8, all possible cut-off points can be determined as follows:

$$0, \frac{1+2}{2}, \frac{2+3}{2}, \frac{3+4}{2}, \frac{4+5}{2}, \frac{5+6}{2}, \frac{6+7}{2}, \frac{7+8}{2}, 9$$

$$= 0, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 9$$

The estimation of the area under the ROC curve (AUC) for medical diagnostic tests has been formulated by some researchers (Bamber, 1975; Krzanowski & Hand, 2009; Pepe, 2003). In the method proposed in this study, it was envisaged that AUC values could be used as a measure of the discrimination of items and the formulas were arranged as follows based on the studies of the mentioned researchers (AUC was visualized in Figure 2).

Figure 2. Area under curve.



Medical diagnostic tests AUC formula includes test results randomly selected from patient and non-patient populations as variables. In the rearranged method, variables were changed as follows: the test result randomly selected from the high-scoring group was defined as the "H" variable and the test result randomly selected from the low-scoring group was defined as the "L" variable. Four classification possibilities arise from these variables.

S variable denotes test score and *t* variable denotes cut-off score:

- 1- The probability that an individual from population H will be correctly classified, high-scoring group true:

$$p(S > t | H).$$

- 2- Probability of an individual from population L being misclassified, low-scoring group false:

$$p(S > t | L).$$

- 3- The probability that an individual from population L will be correctly classified, low-scoring group true:

$$p(S \leq t | L).$$

- 4- Likelihood that an individual from population H will be misclassified, high-scoring group false:

$$p(S \leq t | H).$$

$$\forall x \in (0,1)$$

$$AUC = \int_0^1 ROC(x) dx$$

$$x \rightarrow 0 \text{ as } t \rightarrow +\infty$$

$$x \rightarrow 1 \text{ as } t \rightarrow -\infty$$

$$AUC = \int_{-\infty}^{+\infty} p(S > t | \bar{U}, S = t | A) dt$$

As being unique to this study, the ROC table and the AUC formula used in this study were rearranged in order to examine the discriminations of binary scored test items for the first time.

However, it should be noted that these adaptation attempts were made with regard to item discrimination logic, they do not change mathematical basis of the analysis.

In the examinations made with the method suggested in this article, it was expected to determine the measures reflecting the item discrimination feature with less errors. Because it was thought that taking into consideration many possible cut-off scores in the calculations made with the proposed method would increase the precision of the measurements. The value for the area under the ROC curve, proposed as a measure of item discrimination, is a combination of observations for the functioning of the item under a number of different conditions. On the other hand, it was thought that the 2x2 contingency table, which was taken as the basis in ROC analysis and adapted to examine the item discrimination with this method, coincides logically and psychometrically with the concept of item discrimination. Comparison of the proposed method in this study with the currently used methods was important in terms of revealing the advantages and disadvantages of the approach. In addition, the sample size is also a matter of debate when it comes to choosing a method for determining the psychometric properties of the items. In this context, it was considered important to test the consistency of the proposed method between different samples in terms of size and score distribution. In the literature review, no study was found in which ROC analysis was used to examine item discrimination. In this sense, it can be stated that this study is important for the psychometrics literature. This study will pave the way for other advanced studies. The usage of the proposed method on the determination of the psychometric properties of the items can be advanced and extended by other psychometrists. It is also thought that ROC analysis can be easily carried out with many statistical software, especially SPSS, and it will provide convenience for test developers and test practitioners in terms of ease of calculation.

1.1. Purpose of the Study

The aim of this study was to compare the item discrimination measures obtained from different methods with simulation data and to examine the consistency of the measurements obtained from ROC analysis between samples of different sizes and different distribution characteristics. For this purpose, the answers to the following questions were sought in the study.

- 1- What are the correlations between the upper-lower groups item discrimination indexes (D), biserial correlation coefficients of the item-total test scores (r_{bis}), a parameters from IRT-2PLM, and the AUC values obtained from 20 items and 1000 respondents?
- 2- What are the correlations between the AUC values obtained from 20 items and 100, 200, 400, 1000 respondent groups, and is there a statistically significant difference between these AUC values?
- 3- To what extent are the values determined by different item discrimination prediction methods invariant in case of range narrowing?

2. METHOD

This study is a basic (pure) simulation research aimed at producing new information.

2.1. Data Set

Within the scope of the study, 20 binary scored test items were simulated. Values of the discrimination parameters (a) of these items vary between 0 and 2 and values of their difficulty parameters (b) ranged between -2 and 2. In addition, a group of 1000 respondents whose ability values (θ) ranged between -2 and 2 were simulated.

2.2. Data Analysis

The data handled within the scope of the study were produced in WinGen software and made ready for the analysis. To find the answer to the first research question, D indexes were

calculated based on the correct response rates of the items in the upper-lower groups in terms of test scores, biserial correlation coefficients were calculated between the scores of each item and the total test scores, and finally, the areas under the ROC curve (AUC values) for each item were calculated (proposed method). Since the data were simulated based on the Item Response Theory, the generated a parameters were directly used. Correlations among the item discrimination measures obtained based on four different methods were determined by Spearman's rank-order correlation coefficient method.

In order to answer the second research question, AUC predictions were made by using randomly determined samples of 100, 200 and 400 respondents from the sample of 1000 respondents produced. Correlations between values obtained from samples of different sizes were examined by Spearman's rank difference correlation coefficient method. In addition, the significance of the differences between the predictions was examined with Kruskal Wallis-H. Same analyzes were made and reported with other methods in order to make comparisons.

In order to answer the third research question, the dataset was sorted in ascending order in terms of total test scores. The score range is narrowed by dividing the lowest-scoring 33% and the highest-scoring 33% of the group. The correlation coefficients between the item discrimination measures obtained from these narrowed-range groups and the full dataset were calculated by Spearman's rank difference correlation coefficient method. These analyzes were performed for each of the four different methods.

When using smaller datasets selected from the full dataset, analyzes based on IRT were performed with R (ShinyItemAnalysis) to obtain the a parameters. TAP and SPSS V23 statistical softwares were also used to analyze the research data. For the ROC analysis, the positive value of the real state variable was determined as "1" (items scored as 1-0). Sensitivity and specificity values were determined based on the assumption that larger test scores indicate more positive test results. Nonparametric approach was preferred for the predictions of the areas under the ROC curve.

3. RESULTS / FINDINGS

In order to find an answer to the first research question of the study, the values regarding the discrimination of the items were predicted based on the proposed method and the other three methods, and the findings were given in [Table 3](#).

Table 3. Values predicted by different methods regarding item discrimination.

Item	1	2	3	4	5	6	7	8	9	10
D	0.470	0.250	0.770	0.560	0.710	0.350	0.460	0.500	0.320	0.720
r_{bis}	0.672	0.262	0.808	0.595	0.800	0.380	0.572	0.531	0.381	0.752
a	0.999	0.118	1.717	0.705	1.580	0.270	0.870	0.679	0.405	1.247
AUC	0.837	0.620	0.880	0.775	0.873	0.674	0.773	0.747	0.681	0.846
Item	11	12	13	14	15	16	17	18	19	20
D	0.350	0.740	0.420	0.670	0.620	0.660	0.200	0.400	0.610	0.510
r_{bis}	0.425	0.821	0.450	0.745	0.643	0.707	0.192	0.705	0.775	0.679
a	0.538	1.539	0.476	1.430	0.992	1.271	0.630	1.473	1.537	1.321
AUC	0.701	0.888	0.709	0.851	0.797	0.829	0.586	0.861	0.874	0.831

Hosmer and Lemeshow (2000) stated that if the AUC value is equal to 0.5, no discrimination can be mentioned, it is acceptable if the value is between 0.7 and 0.8, perfect if it is between 0.8 and 0.9, and an extraordinary distinction if it is greater than 0.9. Considering this view, it

can be suggested that the AUC value should be above 0.7 for a good item discrimination. When Table 3 was analyzed, it was seen that items with AUC values below 0.7 (item no 2, 6, 9, 17) have low D and r_{bis} values. It was also determined that these items had very low or low discrimination in terms of a parameters. This determination was made according to Baker's (2001) criteria that the value of a parameters can be interpreted as very low discrimination in the range of 0.01 - 0.34, low discrimination in the range of 0.35 - 0.64, medium discrimination in the range of 0.65 - 1.34, high discrimination in the range of 1.35 - 1.69, and very high discrimination is greater than 1.70.

The correlation coefficients between the values in Table 3 were determined and the related findings were given in Table 4.

Table 4. Correlations between predictions made by different methods.

Method	D	r_{bis}	a	AUC
D	1			
r_{bis}	0.912*	1		
a	0.830*	0.970*	1	
AUC	0.838*	0.979*	0.977*	1

*Correlations are statistically significant at the 0.01 level.

Based on the findings, it can be interpreted that D, r_{bis} , a , and AUC values provide similar information on determining the item discrimination. All correlation coefficients showed a positive and high correlation between all pairs of prediction methods. In addition, the a parameters obtained based on the Item Response Theory showed the highest correlation with the AUC values obtained based on the ROC analysis among all other methods. This was noted because the Item Response Theory currently prevails among the test theories.

In order to find the answer to the second research question, the AUC values of the same items from 100, 200 and 400 groups determined randomly from the full data set of 1000 respondents were predicted and the correlations of the obtained values between the different groups were given in Table 5.

Table 5. Correlations between AUC values obtained from different sized samples.

Sample size	100	200	400	1000
100	1			
200	0.916*	1		
400	0.836*	0.930*	1	
1000	0.791*	0.912*	0.986*	1

*Correlations are statistically significant at the 0.01 level.

When Table 5 was examined, it was determined that there were positive high correlations between AUC values obtained from different sized samples. The lowest correlation coefficient (0.791) was among the samples consisting of 1000 and 100 respondents while the highest correlation coefficient (0.986) was among the samples consisting of 1000 and 400 respondents. In addition, the statistical significance of the differences between the AUC values obtained from different samples was examined with the Kruskal Wallis-H test and it was found that the p value of the test was 0.876. Findings showed that the predictions for the areas under the ROC curve were similar between different sized samples and the differences between the values are not statistically significant. In this regard, it can be inferred that large samples are not required to examine item discriminations with the proposed method.

Similar comparisons were also made between predictions made by other methods from different sized samples. The correlation coefficients between the D values obtained from four different sized samples ranged from 0.867 to 0.997. The correlation coefficients between r_{bis} values ranged from 0.783 to 0.977. Finally, correlation coefficients between the a parameters were in the range of 0.823 and 0.979. In addition, Kruskal Wallis-H test results showed that there was no statistically significant differences between the compared predictions. In this sense, it cannot be claimed that the AUC method provides an advantage over the other methods regarding this comparison.

The third research question was about examining the effect of range narrowing on the predictions. Accordingly, predictions were obtained from the lower 33% and upper 33% parts of the dataset in terms of the total test scores with different methods. Correlations of these predictions with each other and with the full dataset were determined for each method. Findings were given in Table 6.

Table 6. Correlations between narrowed-range datasets and full dataset.

	Lower 33%- Full Data	Upper 33%- Full Data	Lower 33%-Upper 33%
D	0.102	0.126	-0.624**
r_{bis}	0.640**	0.421	0.405
a	0.496*	0.257	0.691**
AUC	0.744**	0.586**	0.526*

** Correlations are statistically significant at the 0.01 level.

* Correlations are statistically significant at the 0.05 level.

In the analyzes performed in terms of the invariance of item characteristics in case of range narrowing, it was determined that the most unstable indexes were the D indexes, which were calculated with the correct response rates of the upper and lower groups. This finding indicated that it would not be appropriate to use this method with homogeneous groups in terms of test scores. On the other hand, AUC values were the measures that showed the best performance compared to others, especially in terms of higher correlation between narrowed-range datas and full data values. Accordingly, it can be argued that the use of ROC analysis would be more appropriate than other methods when determining item discriminations with homogeneous groups.

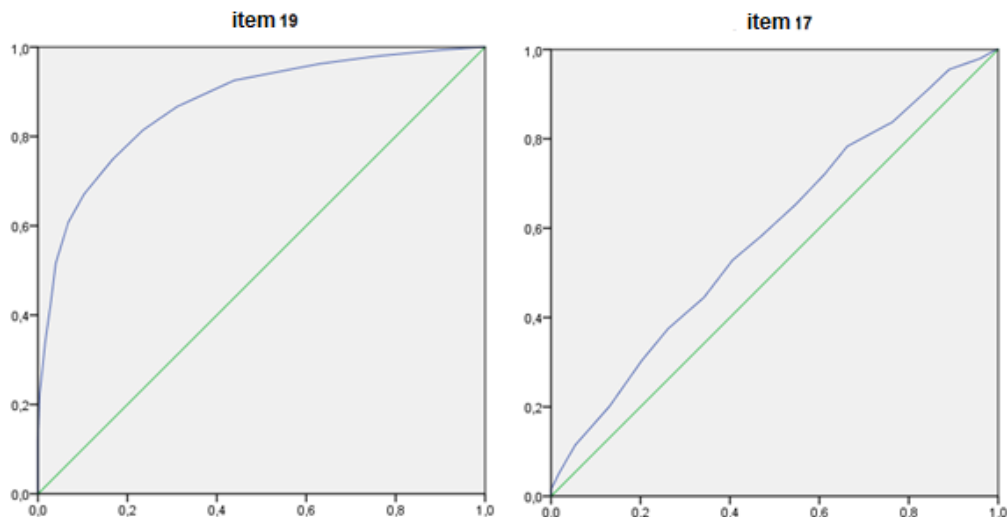
4. DISCUSSION and CONCLUSION

As a result of the findings obtained from the comparisons made in this study, various evidence has been obtained for the claim that the ROC analysis, which is used to determine the degree of discrimination between patients and non-patients, especially with medical diagnostic tests, can also be used to examine item discrimination. In the proposed method, it was determined that the AUC values, which were accepted as a measure of item discrimination, positively and highly correlated with the D, r_{bis} and a parameter values of the items. In addition, it was interpreted that the items with AUC values below 0.7 were low or very low discriminative items based on the values of D, r_{bis} and a parameters.

Based on the aforementioned findings, it was concluded that the criteria proposed by Hosmer and Lemeshow (2000) for interpreting the area under the ROC curve can also be accepted if the analysis is used for the study of item discriminations. It can be stated that the AUC values obtained by the proposed method should take at least 0.7 in order for the discrimination of the items to be acceptable, and the discrimination of the items increases as this value gets closer to 1. In this study, it was concluded that AUC values were consistent between different sized samples and that large samples were not required to examine item discrimination with the

proposed method. In addition, it was determined that the AUC values were affected less negatively compared to other methods if the score distributions in the group were homogeneous. This was noted as a very important advantage of determining item discriminations with ROC analysis. It should also be mentioned that, with the proposed method, item discriminations can be examined not only with AUC values, but also with ROC curve graphs.

Figure 3. ROC curves showing the discrimination of two different items.



It can be stated that as the ROC curve gets closer to the upper left corner of the graph, the discrimination of the examined item increases. As seen in [Figure 3](#), the discrimination of item 19 is high, and the item 17 is low. It is thought that the method proposed in this study may also be advantageous in terms of ease of interpretation by providing visual expression of the discrimination of the items.

The ROC curves method adapted to item analysis can be recommended to test developers, test practitioners, and other researchers since it provides consistent predictions, and it does not require very large groups for these predictions. In this respect, proposed method can be added to the literature as an alternative method. Other researchers working in the field of psychometrics may develop or criticize the method from various aspects as well.

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



Sait Cum  <https://orcid.org/0000-0002-0428-5088>

5. REFERENCES

- Alonzo, A.T., & Pepe, S. M. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, 3(3), 421-432. <https://doi.org/10.1093/biostatistics/3.3.421>
- Baker, F.B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.
- Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4), 387–415. [https://doi.org/10.1016/0022-2496\(75\)90001-2](https://doi.org/10.1016/0022-2496(75)90001-2)

-
- Çüm, S., Gelbal, S., & Tsai, C-P. (2016). Examination of the consistency of the sato test theory item parameters obtained from different samples. *Journal of Measurement and Evaluation in Education and Psychology*, 7(1), 170-181. <https://doi.org/10.21031/epod.69276>
- DeMars, C. (2016). *Madde tepki kuramı [Item response theory]*. Nobel.
- Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement*. Prentice-Hall Inc.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.
- Gönen, M. (2007). *Analyzing Receiver Operating Characteristic Curves with SAS®*. SAS Institute Inc.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer.
- Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression*. John-Wiley & Sons, INC.
- Krzanowski, W.J., & Hand, D.J. (2009). *ROC curves for continuous data*. Chapman and Hall/CRC Press.
- Osterlind, S. J. (1990). Toward a uniform definition of a test item. *Educational Research Quarterly*, 14(4), 2-5.
- Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction*. University Press, Oxford.
- Ruopp, D. M., Perkins, J. N., Whitcomb, W. B., & Schisterman, F. E. (2008). Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal*, 3, 419-430. <https://doi.org/10.1002/bimj.200710415>
- Van Erkel, A.R., & Pattynama, P.M. (1998). Receiver operating characteristic analysis: Basic principles and applications in radiology. *European Journal of Radiology* 27, 88-94. [https://doi.org/10.1016/S0720-048X\(97\)00157-5](https://doi.org/10.1016/S0720-048X(97)00157-5)
- Zou, H. K., Liu, A., Bandos, I.A., Onho-Machado, L., & Rockette, E. H. (2012). *Statistical evaluation of diagnostic performance topics in roc analysis*. CRC Press.

A Comparison of Latent Class Analysis and the Mixture Rasch Model Using 8th Grade Mathematics Data in the Fourth International Mathematics and Science Study (TIMSS-2011)

Turker Toker ^{1,*}, Kathy Green ²

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

²University of Denver, Morgridge College of Education, Department of Research Methods and Information Science, USA

ARTICLE HISTORY

Received: Feb. 28, 2021

Accepted: Nov. 10, 2021

Keywords:

Latent class analysis,
The Mixture Rasch model,
Psychometrics,
Mathematics,
Validation.

Abstract: This study provides a comparison of the results of latent class analysis (LCA) and mixture Rasch model (MRM) analysis using data from the Trends in International Mathematics and Science Study – 2011 (TIMSS-2011) with a focus on the 8th-grade mathematics section. The research study focuses on the comparison of LCA and MRM to determine if results obtained differ when the assumed psychometric model differs. Also, a log-linear analysis was conducted to understand the interactions between latent classes identified by LCA and MRM. Response data to the three booklets were used to run latent class analysis using Mplus 7.31 (Muthén & Muthén, 2012a) for LCA and WINMIRA (von Davier, 2001a). The findings of this paper do not reveal unequivocally whether a model based on primarily qualitative differences (LCA), that is, different strategies, instructional differences, curriculum etc. or a model including additional factors of quantitative differences within strategies (MRM) should be used with this particular dataset. Both of the tests provided similar results with more or less similar interpretations. Both techniques fit the data similarly, a result found in prior research. Nonetheless, for tests similar to TIMSS exams, item difficulty parameters can be useful for educational researchers giving potential priority to use of MRM.

1. INTRODUCTION

Latent class analysis (LCA) is a subgroup of structural equation modeling which is used to find categorical groups or subtypes of cases, in the present case based on responses to test items (McCutcheon, 1987). Mixture Rasch models, which combine Rasch models with latent class analysis, have been used to identify latent classes who might use different problem-solving techniques or who use different skills in response to test items. The purpose of this study was to compare of the results of latent class analysis and mixture Rasch model analysis for a major international assessment in mathematics. Latent class analysis and mixture Rasch model analysis are two approaches to identification of latent classes in data. The purpose of the two approaches and likely the outcomes overlap but assumptions about the nature of the data and the information derived from each approach differ. The existence of multiple latent classes in

*CONTACT: Turker Toker ✉ turker.toker@usak.edu.tr 📍 Usak University, Faculty of Educational Sciences, Department of Educational Measurement and Evaluation, Turkey

e-ISSN: 2148-7456 / © IJATE 2021

test data speaks to the validity of test scores, particularly with the mixture Rasch model. If multiple latent classes are found in test data, distinct groups of participants exist for whom the construct varies, making cross-country comparisons suspect.

In this study, results of two statistical techniques for latent class estimation based on students' responses were compared.

1.1. Latent Class Analysis and the Mixture Rasch Model

Since both techniques are used in educational sciences, it is important to summarize their similarities and differences. Rasch models assume that participants who have the same ability have similar item solution techniques, skills, and psychological procedures used for solution (Fischer & Molenaar, 2012). However, studies in cognitive psychology and standardized testing have suggested that participants at the same ability level might use totally different techniques and strategies and take different paths to arrive at a solution (Sigott, 2004; Sternberg, 1985). If so, the test construct may change for different participants depending on the paths they take for solving the items, which is a threat to construct validity. LCA and the MRM are statistical models used to examine this threat.

Analysis of examinee responses to test items typically rests on the assumption that item parameters are homogeneous across examinees; that is, the items are assumed to behave in the same way for all examinees. In a conventional Rasch analysis, a single difficulty parameter is estimated for each item, and all item difficulty estimates are located on a single dimension along with a single ability parameter for each examinee. However, when examinees systematically differ in the ways they understand or solve items, this assumption may no longer hold. Differences in item solution processes, for example, can give rise to differences in item position parameters and hence to different latent classes.

The fundamental concept underlying LCA is straightforward: some of the parameters of a statistical model differ across unobserved subgroups. These subgroups, which are posited to be nations in this case, are the categories of a categorical latent variable (Vermunt & Magidson, 2004). The mixture Rasch model, on the other hand, is based on the Rasch model (Rasch, 1960), and was introduced by Rost (1990). It is a mixture of a latent trait approach and a latent class approach to model qualitative and quantitative ability differences. The model assesses a set of items as a whole. Therefore, it is the set of item parameters for all items that is tested for differences between latent classes rather than each item parameter being tested individually (Frick, Strobl, & Zeileis, 2015).

LCA estimates relationships between indicator variables due to class membership only. Also, it calculates class membership probabilities instead of fixed class memberships. For example, if there are four suspected classes in a data set the probability of a participant being in each class might be as follows: 0.76, 0.14, 0.08, and 0.02. Since LCA does not provide fixed class memberships for each case, another step takes place within the model selection process called "quality of the classification of latent class membership" (Wang & Wang, 2012). A criterion value from Nagin's (2005) study is used to determine the quality (.70 and higher). Finally, LCA requires each latent class to be defined in a meaningful manner so variance within the population can be described. As a result of this, latent class interpretation is a very important step of LCA.

However, in the MRM, because each class of participants shows a different pattern of response, there are different parameter estimates for each class. The class-related differences in item parameter estimates (the relative difficulty of items) provides differences in how the construct being examined is understood by that class's respondents. Unlike LCA, the class assignment method the MRM uses is a fixed assignment procedure called modal class. One important point is that LCA's path for class membership divides the sample into different groups. Final class

membership probabilities provide percentages rather than fixed class membership. At first, one might emphasize that LCA’s procedure can provide statistical optimization. However, while gaining statistical optimization, classification interpretability and usability can be lost. Also, in the case of a follow up study with same participants, 72% of one case cannot be invited to a focus group while 28% of the same case stays in another group (Dallas & Wilse, 2013).

The solution the mixture Rasch model provides on this matter is using item difficulty parameters. Since the main product of each class is item difficulty parameters, interpretation of classes is derived from differences in item difficulties. Therefore, there is no need to evaluate the quality of the classification of latent class membership, and to define the latent classes for modeling purposes in the MRM.

2. METHOD

2.1. Research Goal

This study evaluated and compared the performance of LCA and MRM methods. Both techniques were used in terms of questionnaire validation to see if TIMSS-2011 data yielded different sub-groups within the selected nations.

2.2. Sample and Data Collection

Data used in this study were taken from the TIMSS-2011 8th grade mathematics section administered in 2011. Students’ responses to the items were used for analyses. There were 26.596 8th grade students from four different nations. The reason to select these nations was mainly their performance shown on the exam and their cultural differences. For country specific descriptive information, see [Table 1](#).

Table 1. *Gender and Age of TIMSS-2011 Subjects (based on booklet selection).*

Nation	Count		Gender (%)				Mean Age	
			Girl		Boy			
	Selected	Population	Selected	Population	Selected	Population	Selected	Population
Turkey	1.225	6.928	48.70	49	51.30	51	14.08	14.00
USA	1.990	10.477	49.70	51	50.30	49	14.22	14.20
Singapore	1.229	5.927	49.40	49	50.60	51	14.39	14.80
Finland	768	4.266	50.30	48	49.70	52	14.75	14.40

Note: Gender is shown in percentages.

The TIMSS-2011 8th grade mathematics test consisted of 217 items which included 118 multiple-choice items in 14 different booklets. Each booklet contained 10-18 items. Six of the mathematics blocks were released. Only Booklets One, Four, and Six were used due to having a larger number of released items in those booklets but only results from booklet six will be discussed. The total number of released items included in these booklets is 40.

2.3. Analyzing of Data

Response data to the three booklets were used to run latent class analysis using Mplus 7.31 (Muthén & Muthén, 2012a) for LCA and WINMIRA (von Davier, 2001a). Competing models were selected by using information criterion values which the Bayesian information criterion (BIC) and Akaike’s information criterion (AIC) for LCA and Pearson Chi-square value and Cressie-Read statistic (Cressie & Read, 1984) for the MRM. Although the original study contained 3 different booklets results from booklet four will be used due to limitations on word count. Also, it is important to emphasize that both techniques provided similar results for booklets one and six.

The latent class structure of the TIMSS-2011 8th-grade mathematics data was assessed by both analyses. Following that a log-linear analysis was conducted to see if defined latent classes were similar.

3. RESULTS / FINDINGS

3.1. Latent Class Analysis

The fit statistics and information criterion indices for the models, which ranged from 1 to 4 latent classes, are shown in Table 2. Based on the p -values of the LMR LR test ($p = 0.29$) and the BLRT test ($p = 0.14$), both were statistically nonsignificant at the 4-class model; hence, the test failed to reject the 3-class model in favor of a four or more class model. Also, non-decreasing BIC (21392) of the 4- class model supported evidence for the 3-class model, the non-decreasing AIC (21207) of the 4-class model supported evidence for the 3-class model. Hence, the fit of the 3-class model was decided to be adequate and the selected model for further analysis for Booklet Four.

Table 2. LCA Model Fit Indices for Booklet Four.

Model	BIC	AIC	LMR LRT p -value	BLRT p -value
1-class	N/A	N/A	N/A	N/A
2-class	21371	21256	<0.001	<0.001
3-class	21332	21157	<0.001	<0.001
4-class	21392	21207	0.29	0.14

Note. BIC = the Bayesian information criterion; AIC = Akaike’s information criterion; LMR LRT = Lo-Mendell-Rubin Likelihood Ratio Test; BLRT = Bootstrap Likelihood Ratio Test.

3.1.1. Classification quality

The final class sizes and percentages for the latent classes are given in Table 3. Table 3 shows that 473 students (27.1%) were assigned to Class 1, 694 students (39.0%) were assigned to Class 2, and 579 students (33.9%) were assigned to Class 3. The average latent class posterior probabilities for the most likely latent class membership are reported in Table 4. The probability for most likely latent class membership for students assigned to the first class was 0.87, while the probability of misclassification was 0.13. Similarly, for students assigned to the second class, the probability of correct class membership was 0.76, while the probability of misclassification was 0.24; for students assigned to the third class, the probability of correct class membership was 0.80, while the probability of misclassification was 0.20. All average latent class probabilities for most likely latent class membership exceeded 0.70. Furthermore, entropy was .69 which show that latent class membership classification quality was adequate enough for the 3-class model.

Table 3. Final Latent Class Size and Percentage for Booklet Four.

Classes	Size	Percentage
1	473	27.1 %
2	694	39.0 %
3	579	33.9 %

Table 4. Average Latent Class Probabilities for Most Likely Latent Class Membership for Booklet Four.

Classes	Probability of Class 1 Membership	Probability of Class 2 Membership	Probability of Class 3 Membership
1	0.87	0.13	0.00
2	0.09	0.76	0.15
3	0.00	0.20	0.80

3.1.2. Definition of latent classes

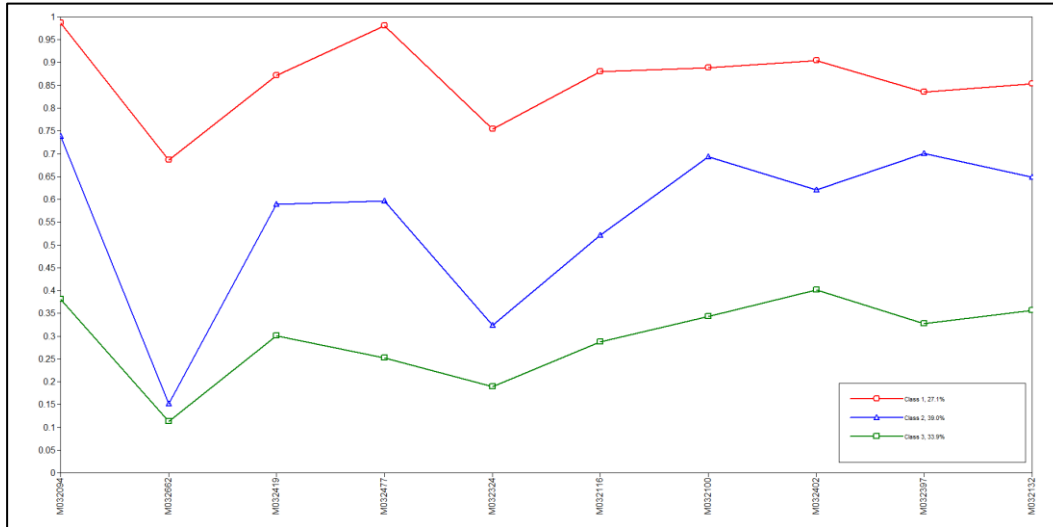
The differences in the sample population were explored by analysis of the estimated item-response probability of endorsing “Correct Response” for each of the 10 items. The three latent classes—highly skilled students, moderately skilled students, and somewhat skilled students—were labeled by the researcher based on the observed pattern of item response probabilities. The highly skilled students’ class, denoted as Class 1 consisting of 473 students, had the highest item-response probabilities for each of the 10 items. Class 2, which contained 694 students with the second highest item-response probabilities for each of the 10 items, as moderately skilled students; Class 3 was defined as somewhat skilled students, which contained 579 students and had the lowest item-response probabilities for each of the 10 items. The unconditional latent class probabilities and the conditional probabilities for endorsing “Correct Answer” are reported by latent class in [Table 5](#).

Table 5. Three-Class Latent Class Membership for Booklet Four.

Item	Probability of Class 1	Probability of Class 2	Probability of Class 3
	Unconditional		
	0.27	0.40	0.33
	Conditional “Correct Answer”		
M032094	0.99	0.74	0.38
M032662	0.69	0.15	0.11
M032419	0.87	0.59	0.30
M032477	0.98	0.60	0.25
M032324	0.76	0.32	0.19
M032116	0.88	0.52	0.29
M032100	0.89	0.69	0.34
M032402	0.90	0.62	0.40
M032397	0.84	0.70	0.33
M032132	0.85	0.65	0.36

Conditional probability profiles for endorsing the “Correct answer” for the 3-Class model are shown in Figure 1.

Figure 1. Conditional Probability Profiles of Endorsing “Correct Answer” for 3-Class LCA Model for Booklet Four (Mplus Version 7.31).

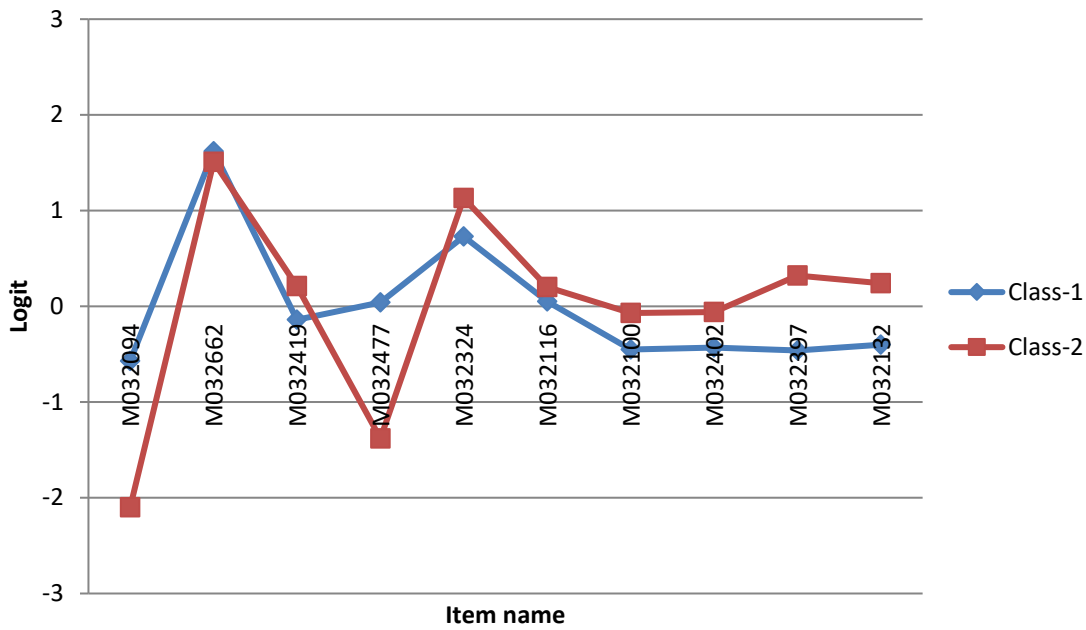


3.2. Mixture Rasch Model

The dataset consisted of 10 items with 1746 participants. To determine the appropriate number of classes, one, two, three, and four latent class solutions were fit to the data (see Table 2). *P*-values for Booklet Four of Cressie-Read and Pearson Chi-square were .13 and .15. Since the two-class model had the highest *p*-value, a two-class solution was selected for Booklet Four. Class size values for each class presents that class 1 was expected to include about 66% of the sample. Class 2 was expected to include about 34% of the sample. The class sizes indicate that about 66 percent and 34 percent of the sample can be fitted by a mixed Rasch model which was assumed to hold in these classes. According to the Q-index, there was one item (M032662) with a *Z_q* value of 2.37 and *p*-value of .01 which shows lower discrimination in class one. In such cases, item removal is suggested from the scale only after examining the items content and additional information from the estimated model (von Davier, 2001b). Item category values for this item were acceptable. Out of 1,746 responses 1,251 students answered the item false and 495 students answered correct. Additionally, the item parameter value for class one was also acceptable with a value of .13. After examining the item category values and item fit, it is decided not to remove the item from analysis. All of the other items fit each class well ($.05 < p < .95$) (see Table 5).

Figure 2 shows that the two classes had similar item difficulty parameters for the first six items and different item difficulty parameters for the last four items. These four items were slightly easier for first class then for the second class. The lines display items on which the two classes seem to diverge and later to converge.

Figure 2. Class specific item parameter profiles for Booklet Four.



The majority of items were not markedly different in difficulty across classes. In general, all classes found the items to be relatively easy as logit position was generally negative (see Table 6 for specific values including standard error).

Table 6. Item parameters of Booklet Four by classes.

Item	Class-1		Class-2	
	Estimate	Error	Estimate	Error
M032094	-0.57	0.06	-2.10	0.29
M032662	1.62	0.09	1.51	0.09
M032419	-0.14	0.07	0.21	0.12
M032477	0.04	0.07	-1.38	0.21
M032324	0.73	0.07	1.13	0.10
M032116	0.05	0.07	0.20	0.12
M032100	-0.45	0.06	-0.07	0.13
M032402	-0.43	0.06	-0.06	0.13
M032397	-0.46	0.06	0.32	0.11
M032132	-0.40	0.06	0.24	0.12

A four-way log-linear analysis was conducted with variables nation, gender, LCA class membership, and the MRM class membership. The likelihood ratio chi-square with no parameters and only the mean was 2326.18. The value for the first order effect was 1897.99. The difference $2326.18 - 1897.99 = 428.19$ is displayed on the first line of Table 7.

Table 7. *K-Way and Higher-Order Effects for Booklet Four.*

	K	df	Likelihood Ratio	
			Chi-Square	<i>p</i>
K-way Effects	1	7	428.19	<.001
	2	17	1894.55	<.001
	3	17	3.43	1.00
	4	6	0.02	1.00

The significant *p* value (< .001) shows that there was a first order effect. The addition of a second order effect improved the likelihood ratio chi-square by 1894.55. This was also significant. But the addition of a third and a fourth order term did not significantly improve fit (*p*> .05).

Table 8. *Partial Associations for Booklet Four.*

Effect	df	Partial Chi-Square	<i>p</i>
LCA*NATION*MRM	6	.00	1.00
LCA*NATION*ITSEX	6	3.25	.78
LCA*MRM*ITSEX	2	.00	1.00
NATION*MRM*ITSEX	3	1.33	.72
LCA*NATION	6	65.39	<.001
LCA*MRM	2	1362.86	<.001
NATION*MRM	3	10.46	.02
LCA*ITSEX	2	4.41	.11
NATION*ITSEX	3	10.05	.02
MRM*ITSEX	1	.13	.72
LCA	2	42.13	<.001
NATION	3	216.13	<.001
MRM	1	169.78	<.001
ITSEX	1	.15	.70

Note. NATION= Countries, ITSEX=Gender, LCA= Latent Class Analysis Group Membership, MRM= Mixed Rasch Model Group Membership

Table 8 shows that there were statistically significant associations between nation and LCA class membership (*p*< .05), nation and the MRM class membership (*p*< .05), LCA class membership and MRM class membership (*p*< .05), and nation and gender (*p*< .05) for Booklet Four. All other interactions between other variables were not statistically significant (*p*> .05). Due to the purpose of this paper only the association between LCA and MRM results will be explained.

To further analyze the interactions of LCA class membership, and the MRM class membership variables a custom model was created with the significant two-way associations.

Table 9. *Goodness-of-Fit Tests for 2-way Interaction Model for Booklet Four.*

	Chi-Square	df	<i>p</i>	Adjusted	
				df ^a	<i>p</i>
Likelihood Ratio	7.99	26	1.00	10	.63

a. One degree of freedom is subtracted for each cell with an expected value of zero. The unadjusted df is an upper bound on the true df, while the adjusted df may be an underestimate.

In [Table 9](#), the goodness of fit test showed that the model fit the data adequately ($p > .05$). Also, a crosstab analysis for Booklet Four was done to see LCA class memberships and the MRM class membership agreement level. Although LCA and MRM analysis provided a different number of classes for Booklet Four, LCA’s class one (highly skilled students) overlapped 100 % with MRM class two. LCA class two (moderate skill students) overlapped with both MRM class one (81.3%) and class two (18.7%). LCA class three (somewhat moderate skilled students) overlapped with only MRM class one (see [Table 10](#)).

Table 10. Crosstabulation of LCA Class Membership vs. MRM Class Membership for Booklet Four.

		MRM GROUP MEMBERSHIP			
		Class 1	Class 2	Total	
LCA GROUP MEMBERSHIP	Class 1	Count	0	473	473
		% within LCA GROUP MEMBERSHIP	0.0%	100.0%	100.0%
	Class 2	Count	564	130	694
		% within LCA GROUP MEMBERSHIP	81.3%	18.7%	100.0%
	Class 3	Count	579	0	579
		% within LCA GROUP MEMBERSHIP	100.0%	0.0%	100.0%
Total	Count	1143	603	1746	
	% within LCA GROUP MEMBERSHIP	65.5%	34.5%	100.0%	

Please present the findings/results in this section. This section should give significant results obtained from the study clearly and concisely. Please present the findings/results in this section. This section should give significant results obtained from the study clearly and concisely.

4. DISCUSSION and CONCLUSION

For item parameters, both of the techniques calculate item logit values and standard errors. For LCA, item parameter estimates are on the logit scale, and therefore, can be somewhat difficult to interpret. The same information is given in a more interpretable scale under the MRM where item parameters are products of item difficulty measure for each class. However standard errors of the parameters have very close results for Booklet 4 (see [Table 6](#))

The decision on number of classes differs in the two techniques. BIC and AIC were used to evaluate fit for LCA. On the other hand, since Winmira2001 considered data as being sparse, Cressie-Read and Chi-square values were used for model fit purposes. However, based on BIC values, both techniques provided similar results (see [Tables 2](#)). So, it can be concluded that selecting one model over another model did not depend on fit values. Since a qualitative conclusion is important for LCA, model fit is not enough by itself. There are also other combinations of different values such as average estimated posterior probabilities for quality (Nagin, 2005) and entropy value (Clark, 2010). Moreover, latent classes should be defined in an interpretable way as well. For the MRM, the solution is simpler. If there is model fit based on fit indices the next step is simply interpretation of the model.

The two analyses had somewhat different solutions for the class weights for all booklets. It can be interpreted that latent class analysis puts the most cases into the middle class for three class solutions and to the second class for two class solution. LCA uses response probabilities in which students have the same probability of giving the correct answer within the same class. As a result of this, students in the same class have no quantitative differences. The only difference created and shown by LCA is between groups which is a product of qualitative

differences. In our case, this would be interpreted as item correct response values based on students' background. However, the mixture Rasch model, regardless of number of classes within the solution, sorts classes based on similarity in their response patterns which results in the placement of cases with an order where most student fall in to the first class, then second, then third etc. Since there are differences between item parameters within the same class for the MRM, interpretation changes and relies on two things: one being latent class membership and two being the class specific quantitative person parameter (Büsch, Hagemann, & Bender, 2010).

This study provides useful information about two commonly used techniques in educational research. Since the data used in this study are from a real data set, none of the techniques were tested under controlled circumstances such as different levels of amount and type of missing data, presence of outliers, sample size (bigger, smaller), item distributions, score distributions, etc. Monte Carlo simulation studies are recommended to see if the results differ under these different conditions.

Further, TIMSS multiple-choice items were dichotomous; use of items with varied responses scales is also recommended, as are studies with item content very different from a mathematics achievement test. For example, studies are recommended that compare LCA and MRM when the construct assessed is a personality variable or attitudinal as well as achievement. The comparison of both techniques is limited to dataset used in this study. Therefore, it is suggested that same study can be done using other type of questionnaires.

As with any statistical approach that uses binary variables, recoding categorical responses into dichotomous responses was one of the limitations of the study since student responses might result in different classification based on the multiple-choice responses. In any latent class model, the issue of reification is of great importance. Also using a real-world dataset limited the radius of effect area of the study since conclusions are limited to the current data.

Sampling techniques of TIMSS organizers is also another limitation. One simple example shows that number of students in Turkish and American educational systems are more than the whole population of Singapore and Finland. TIMSS requires each participant country to join with at least 4.500 students. Although this number covers most of the Singaporean and Finnish 8th-grade population, it is still small for systems like the US or Turkey (Rutkowski & Rutkowski, 2016). In this case, generalizability of the results is questionable.

Acknowledgments

This study was a part of an unpublished doctoral dissertation.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** University of Denver/Institutional Review Board, mary.travis@du.edu, February 4, 2016.

Authorship Contribution Statement

Authors are expected to present author contributions statement to their manuscript such as; **Turker Toker:** Investigation, Resources, Visualization, Software, Formal Analysis, and Writing - original draft. **Kathy Green:** Methodology and Supervision. Authors may edit this part based on their case.

ORCID

Turker Toker  <https://orcid.org/0000-0002-3038-7096>

Kathy Green  <https://orcid.org/0000-0002-0681-2937>

5. REFERENCES

- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. Springer Series in Statistics, 199-213. https://doi.org/10.1007/978-1-4612-1694-0_15
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345-370. <https://doi.org/10.1007/bf02294361>
- Büsch, D., Hagemann, N., & Bender, N. (2010). The dimensionality of the Edinburgh handedness inventory: An analysis with models of the item response theory. *Laterality: Asymmetries of Body, Brain and Cognition*, 15(6), 610-628. <https://doi.org/10.1080/13576500903081806>
- Clark, S. L. (2010). *Mixture modeling with behavioral data* (3405665) [Doctoral dissertation]. ProQuest Dissertations and Theses Global.
- Cressie, N., & Read, T. R. C. (1984a). Multinomial Goodness-Of-Fit Tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3), 440-464. <https://doi.org/10.1111/j.2517-6161.1984.tb01318.x>
- Dallas, A. D., & Willse, J. T. (2013). Survey analysis with mixture Rasch models. *Journal of Applied Measurement*, 15(4), 394-404. <https://europepmc.org/article/med/25232672>
- Fischer, G. H., & Molenaar, I. W. (Eds.). (2012). *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media.
- Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*, 75(2), 208-234. <https://doi.org/10.1177/0013164414536183>
- McCutcheon, A. L. (1987). *Latent class analysis*. SAGE.
- Muthén, L. K., & Muthén, B. O. (2012a). *Mplus* (Version 7.31) [Computer Software]. Los Angeles, Muthén&Muthén.
- Muthén, L. K., & Muthén, B. O. (1998). 2014. *Mplus User's Guide, 7th edition*. Muthén & Muthén.
- Nagin, D. (2005). *Group-based modeling of development*. Harvard University Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Danish Institute for Educational Research. <https://doi.org/10.4135/9781412961288.n335>
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271-282. <https://doi.org/10.1177/014662169001400305>
- Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, 45(4), 252-257. <https://doi.org/10.3102/0013189X16649961>
- Sigott, G. (2004). *Towards identifying the C-Test construct*. Peter Lang.
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. CUP Archive.
- Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. *The Sage Encyclopedia of Social Sciences Research Methods*, 2, 549-553. *Methods*. <https://doi.org/10.4135/9781412950589.n472>
- von Davier, M. (2001). *WINMIRA* [Computer software]. Institut für die Pädagogik der Naturwissenschaften
- von Davier, M. (2001b). *WINMIRA user manual* [Computer software manual]. Institut für die Pädagogik der Naturwissenschaften
- Wang, J., & Wang, X. (2019). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons.

6. APPENDIX

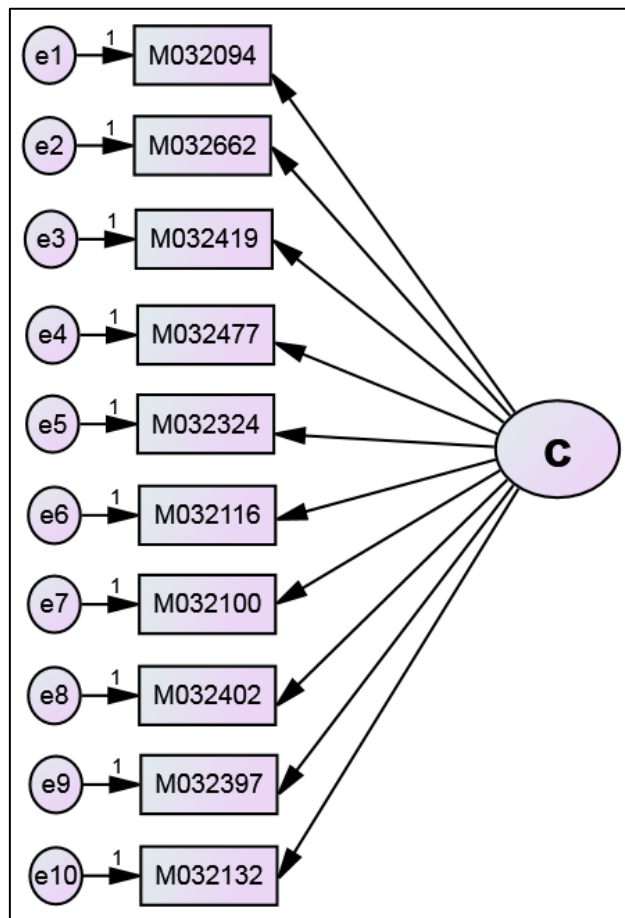
Appendix A: LCA 2 Class Model Specification for Booklet One (Other Classes Similar) (Mplus Version 7.11).

```
Mplus VERSION 7.11
MUTHEN & MUTHEN
05/17/2016 11:26 AM

INPUT INSTRUCTIONS

Title:
Booklet 1 2 Class Solution Latent Class Analysis.
Data:
  File is Booklet1 2 class.dat;
Variable:
  names          = IDSTUD M066 M021 M026 M095 M173 M016 M028 M014 M073 M002 M084 M029;
  usevariables   = M066 M021 M026 M095 M173 M016 M028 M014 M073 M002 M084 M029;
  categorical    = M066 M021 M026 M095 M173 M016 M028 M014 M073 M002 M084 M029;
  classes       = c(2);
Analysis:
  Type=mixture;
Plot:
  type is plot3;
  series is M066 (1) M021 (2) M026 (3) M095 (4) M173 (5) M016 (6)
           M028 (7) M014 (8) M073 (9) M002 (10) M084 (11) M029 (12);
Savedata:
  file is booklet1_2class_save.txt ;
  save is cprob;
  format is free;
output:
  tech11 tech14;
```

Appendix B: LCA model for Booklet Four (Amos Version 22).



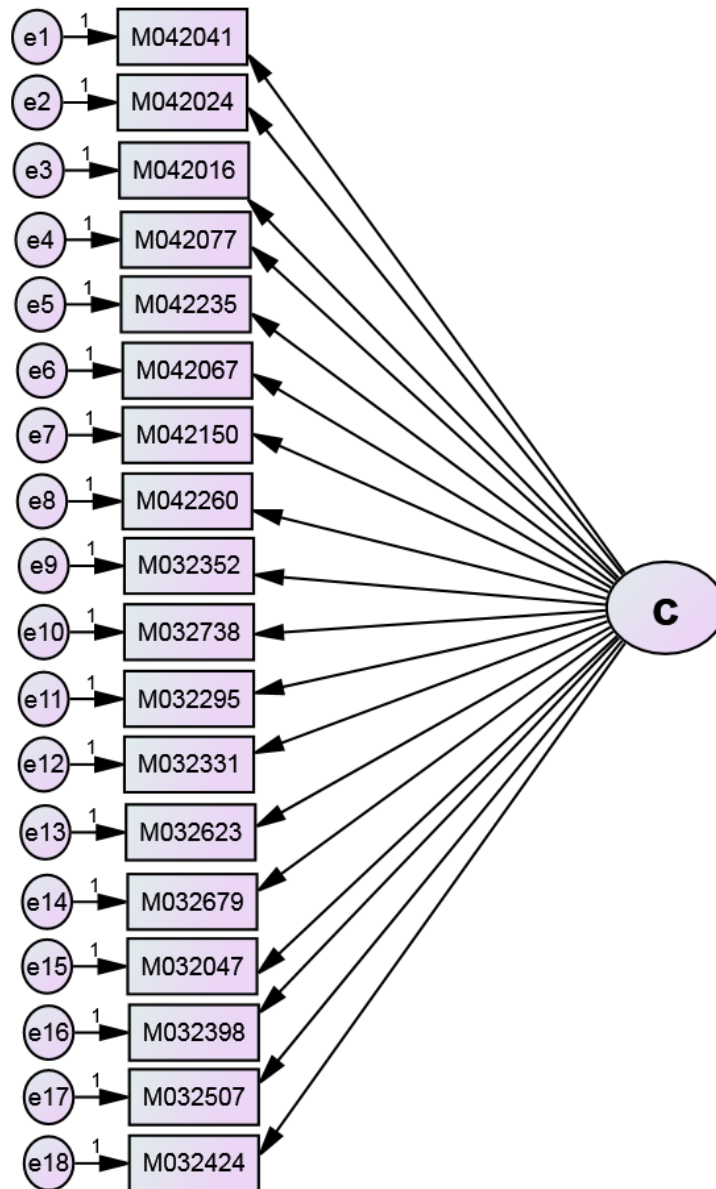
Appendix C: LCA 3 Class Model Specification for Booklet Four (Other Classes Similar)
(*Mplus* Version 7.11).

```
Mplus VERSION 7.11
MUTHEN & MUTHEN
05/17/2016 11:29 AM

INPUT INSTRUCTIONS

Title:
Booklet 4 3 Class Solution Latent Class Analysis.
Data:
  File is Booklet 4 3 Class.dat;
Variable:
  names           = IDSTUD M094 M062 M019 M077 M024 M016 M000 M002 M097 M032;
  usevariables    = M094 M062 M019 M077 M024 M016 M000 M002 M097 M032;
  categorical     = M094 M062 M019 M077 M024 M016 M000 M002 M097 M032;
  classes        = c(3);
Analysis:
  Type=mixture;
Plot:
  type is plot3;
  series is M094 (1) M062 (2) M019 (3) M077 (4) M024 (5) M016 (6)
           M000 (7) M002 (8) M097 (9) M032 (10);
Savedata:
  file is booklet4_3class_save.txt ;
  save is cprob;
  format is free;
output:
  tech11 tech14;
```

Appendix D: LCA model for Booklet Six (Amos Version 22).



Appendix E: LCA 4 Class Model Specification for Booklet Six (Other Classes Similar) (Mplus Version 7.11).

```
Mplus VERSION 7.11
MUTHEN & MUTHEN
05/17/2016 11:32 AM

INPUT INSTRUCTIONS

Title:
Booklet 6 4 Class Solution Latent Class Analysis.
Data:
  File is Booklet 6 4 Class.dat;
Variable:
  names          = IDSTUD M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17 M18;
  usevariables = M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17 M18;
  categorical = M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 M11 M12 M13 M14 M15 M16 M17 M18;
  classes = c(4);
Analysis:
  Type=mixture;
Plot:
  type is plot3;
  series is M1 (1) M2 (2) M3 (3) M4 (4) M5 (5) M6 (6)
           M7 (7) M8 (8) M9 (9) M10 (10) M11 (11) M12 (12) M13 (13)
           M14 (14) M15 (15) M16 (16) M17 (17) M18 (18);
Savedata:
  file is booklet6_4class_save.txt ;
  save is cprob;
  format is free;
output:
  tech11 tech14;
```

The Opinions of Field Experts on Online Test Applications and Test Security During the COVID-19 Pandemic

Hakan Kilinc ^{1,*}, Muhammet Recep Okur ¹, Ilker Usta ¹

¹Anadolu University, Open Education Faculty, Department of Distance Education

ARTICLE HISTORY

Received: Feb. 05, 2021

Revised: Oct. 20, 2021

Accepted: Nov. 16, 2021

Keywords:

Covid-19,
Online test applications,
Online test security,
Online learning
environments,
Focus group interview,
Case study.

Abstract: Within the scope of this study, it was aimed to determine the factors that should be considered regarding the usability and security of online test applications used as an assessment and evaluation tool during the COVID-19 pandemic. In this context, the case study method was used to obtain the opinions of field experts. Furthermore, in this study, using the focus group interview technique as a data collection technique, the criterion sampling method, one of the purposeful sampling methods, was used to determine participants. At this point, it was taken into consideration that the participants were experts in the field of open and distance learning. In this regard, a total of 15 field experts who have experienced online test during the emergency distance education period at Anadolu University, Turkey contributed to the focus group interviews consisting of three groups. The results obtained at the end of the study offer solutions for the usability of online test applications, the use of which has increased with the pandemic process and ensuring the security of these applications.

1. INTRODUCTION

With the Covid-19 pandemic spreading so rapidly and affecting the whole world, the flow and rhythm of life have changed worldwide (Zhao, 2020), and a flexible working model has begun to be implemented in working environments to reduce the effect of the pandemic and slow down its spread due to its feature of being highly contagious. The working-from-home model has become one of the most common methods applied in this context. Another field, the operating principle of which has changed, has been the field of education and training (Doghonadze et al., 2020). In this regard, education and training institutions were closed, and face-to-face education was suspended to prevent the increase in COVID-19 cases. In this context, the education of millions of students from all education levels in many countries around the world was interrupted (UNESCO, 2020a). In such a situation, education, one of the basic human rights, was interrupted (UN, 1984). To compensate for this situation, many educational institutions reacted quickly, and emergency remote education applications were put into practice worldwide. At this point, the courses taught in online environments came to the forefront. In this context, although there were many successful and unsuccessful applications in the emergency remote education process, one of the biggest discussions and deficiencies was

*CONTACT: Hakan Kilinc ✉ hakankilinc@anadolu.edu.tr 📍 Anadolu University, Open Education Faculty, Department of Distance Education

e-ISSN: 2148-7456 /© IJATE 2021

experienced in the assessment and evaluation processes (Bozkurt, 2020). One of these applications is online tests.

The assessment and evaluation process through online tests which are used for the determination of the learning levels of learners and make predictions for the future is carried out by using the information and communication technologies (Gülbahar, 2013; Şimşek, Balaban & Ergin, 2016; Yağcı et al., 2015). Online tests usually carried out in online learning environments through learning management systems such as Blackboard, Canvas, and Moodle and question types such as multiple-choice, true-false, short-long answer, gap-filling, and matching have positive effects for learners, instructor, and institutions. These positive effects can be listed as cost and time savings (Şimşek et al., 2016), storage of answers, providing appropriate and quick feedback, ensuring flexibility, high security by reducing human errors (Angus & Watson, 2009; Jordan & Mitchell, 2009), less effect of instructor (Anderson et al., 2005), and obtaining quick results (Kuhtman, 2004). Moreover, the fact that online tests are technology-based allows the use of multimedia elements instead of face-to-face exams (Liu, Papathanasiou & Yung-Wei, 2001). Additionally, online tests also have limitations such as requiring computer and internet access, the possibility of students cheating or the difficult control of whether the student himself/herself takes the exam, and difficult communication (Anderson et al., 2005; Özen, 2016; Sindre & Vegendla, 2015, Solak et al., 2020). In the studies conducted on online tests within the scope of their positive aspects and limitations, it has been revealed that online tests increase the academic achievement of students (Schmidt et al., 2019), contribute more to academic achievement than the traditional method (Yağcı et al., 2011) and that students feel more comfortable, fast, efficient, and safe compared to classical exams (Saban et al., 2010). Studies conducted on online tests that are frequently used in online learning environments focus on subjects such as learners' opinions on online tests (Koçak et al., 2006, Saban et al., 2010), the effect of online tests on academic achievement (Yağcı et al., 2011), the comparison of online tests with traditional exams (Saban et al., 2010; Yağcı et al., 2011), the effect of online tests on learning and motivation (Marriot, 2009), and the relationship between exam preferences (traditional-online) and performance (Hewson, 2012). However, in this study, online test applications applied in emergency distance education processes that have been implemented with the Covid-19 pandemic process, rather than this wide area on online tests until now, are mentioned. The unpreparedness of many institutions in the emergency distance education process (Bozkurt, 2020; Senel & Senel, 2021) has caused some disruptions in the implementation of online test applications, as in other distance education applications (Bozkurt, 2020; Can, 2020; d'Orville, 2020) From this point, it can be stated that there is a need for studies on the usability, development, and reliability of online test applications in the emergency of distance education process.

1.1. Importance of the Study

Some education and training institutions have suspended the assessment and evaluation practices carried out in face-to-face environments during the COVID-19 pandemic and have started using online test applications (d'Orville, 2020). In a study conducted by UNESCO (2020b), which explains this situation, it is stated that 58 out of 84 countries postponed or rescheduled exams, 23 countries introduced alternative methods such as online or home-based testing, while exams were continued in 22 countries and completely canceled in 11 countries. Therefore, the examination system has been changed in many countries during the pandemic period (Bozkurt, 2020; d'Orville, 2020). In addition to this situation, UNESCO (2020c) emphasized that educational institutions should prepare in the field of assessment and evaluation due to the COVID-19 pandemic. The unpreparedness of educational institutions for the COVID-19 pandemic have required online exam online tests to be carried out without sufficient validity and security reliability studies instead of traditional exams and tests

(d'Orville, 2020). However, with the emergency distance education applications implemented during the Covid-19 period, the unattended online tests in many institutions have threatened the security of the exam (Bozkurt, 2020; Can, 2020; Senel & Senel, 2021). The point that should be emphasized here is the usability of online test applications and the security of these exams (Can, 2020; Solak et al., 2020). In this context during the emergency distance education period, what points should be considered while developing online test applications and how a reliable assessment and evaluation process should be carried out emerge as an important situation that needs to be resolved. Moreover, in this period, increasing the security of online test applications is important to eliminate the negativities (possibility of cheating, etc.) caused by these exams. Based on this, it is thought that this study is important in guiding education and training institutions on online test applications, which are becoming increasingly important with the COVID-19 pandemic process, and that these applications can be carried out in line with their purpose. To reach these targeted outputs within the scope of the study, the opinions of field experts who have experienced online tests during the emergency distance education period are needed.

1.2. Aim of the Study

This study aims to determine the situations that should be considered regarding the usability and test security on online test applications in the COVID-19 pandemic process according to the opinions of field experts. In this context, answers to the following research questions are sought to achieve the purpose of the study:

- 1- What should be considered during the development of online test applications during the COVID-19 pandemic?
- 2- What should be taken into account to ensure the security of online test applications during the COVID-19 pandemic?

At this point, the reason for seeking the opinions of field experts can be listed as follows:

With the Covid-19 pandemic, which emerged unexpectedly, many institutions have had to implement online testing management. For the tests to be applied at this point to be useful and reliable, it is thought that the opinions of field experts who have experienced both the distance education process and the online test application during the pandemic period are guiding. Experts who have had this experience can provide fast and reliable online tests that need to be implemented urgently. To obtain the suggestions of field experts on this subject, their opinions are needed. Therefore, it is thought that one of the most effective ways to reach fast and reliable solutions during the Covid-19 pandemic, in which we have entered the emergency distance education process, is to consult the opinions of field experts who have experienced this process.

The research questions determined within the scope of the study are not related to the general online tests. Instead, these research questions were needed to determine how a reliable online test system should be implemented to minimize the problems that may arise after the use of unattended online test applications in the emergency distance education process.

2. METHOD

This study was designed with a case study, one of the qualitative research approaches. The case study that forms the pattern of this research is based on in-depth analysis of any event, individual or process based on data (Creswell, 2007). According to Yin (2003), case study is a method that investigates the phenomenon in the existing natural environment and is used in cases where the existing situation and the environment in which it is located are not separated by precise lines. Case studies are defined as a research method that works in a real case within its real-life framework, where the boundaries between the case and the environment in which it is located are not evident and there is more than one source of evidence or data (Yin, 1984).

From this point of view, in this research, a document analysis including the applications related to the study was carried out. Then the opinions of field experts were consulted on the development and reliability of online tests applied by many educational institutions during the Covid-19 pandemic period when the emergency distance education process was started.

2.1. Study Group

The criterion sampling method, one of the purposeful sampling methods, was preferred in this study. The basic understanding of the criterion sampling method is to study all situations that meet a predetermined set of criteria. A set of previously prepared criteria can be used here as developed by the researchers (Yıldırım & Şimşek, 2011). As a criterion, it was considered that the participants were experts in the field of open and distance learning. A total of 15 participants who had experienced online tests during the distance education period, contributed to the study performed within this scope. The demographic information of the participants is presented in Table 1.

Table 1. Demographic Information of Participants.

Participant (Pseudonym)	Gender	Title
Masal	Female	Prof. Dr.
Cengiz	Male	Prof. Dr.
Hasan	Male	Prof. Dr.
Ahmet	Male	Prof. Dr.
Gülay	Female	Prof. Dr.
Okan	Male	Prof. Dr.
Mustafa	Male	Prof. Dr.
Ali	Male	Prof. Dr.
Resul	Male	Assoc. Dr.
Anıl	Male	Assoc. Dr.
Tuna	Male	Assoc. Dr.
Hakan	Male	Assoc. Dr.
Cihan	Male	Assoc. Dr.
Fırat	Male	Assoc. Dr.
Esra	Female	Assoc. Dr.

2.2. Data Collection Tool

The interview is the basic data collection technique in the case study design (Ersoy, 2013). In revealing experiences and meanings related to the phenomena, the interview technique provides researchers with interaction, flexibility, and opportunities to examine them through probes (Yıldırım & Şimşek, 2008). In this context, the focus group interview was used to obtain data within the scope of this study. The focus group interview (Bloor et al., 2001), which is commonly used in academic studies, is a technique of using the effect of group dynamics, obtaining in-depth information, and generating ideas in the interview process between a small group and the moderator (Bowling, 2002; Yıldırım & Şimşek, 2008). Krueger (1994) defines the focus group interview as a carefully planned discussion in an environment where individuals can freely express their thoughts. The purpose of focus group interviews is to obtain in-depth, detailed, and multidimensional qualitative information about the participants' perspectives, lives, interests, experiences, tendencies, thoughts, perceptions, feelings, attitudes, and habits on a specified subject (Bowling, 2002). The important thing in focus group interviews is creating

an environment where participants can freely express their opinions. In this sense, the most significant advantage of focus group interviews is that new and different ideas emerge from within-group interaction and group dynamics (Kitzinger, 1995). Focus group interviews, led by a moderator, using the techniques of asking questions, discussing, and summarizing to reveal the thoughts and experiences of participants, consist of a maximum of 10 or 12 participants (Karabekir et al., 2015). Within the scope of this study, a total of three focus group interviews, consisting of five participants each, were held. A moderator guided each group in the focus group interviews. During the interview, it was ensured that the participants sat in a U shape, and a pen and paper were placed on the tables. In the introduction part, the purpose of the study was explained, and the interviews were started after the participants introduced themselves briefly. Care was taken to ensure that the group participants consisted of experts in the open and distance learning field. In the groups where semi-structured interviews lasted approximately 90 minutes were conducted, the moderators recorded the data.

Semi-structured interview questions consisting of two questions were used to obtain the participants' opinions about the online test applications carried out at Anadolu University and exam security during the COVID-19 pandemic period. Before the interviews, the content validity of the semi-structured interview form was ensured by obtaining the opinions of three experts working on qualitative research methods, and it was finalized.

2.3. Data Analysis

The content analysis method was used in the data analysis. The main purpose of content analysis is to reach the concepts and relationships that can explain the collected data. To this end, the data collected must first be conceptualized, then organized logically according to the resulting concepts, and the themes explaining the data must be determined accordingly (Yıldırım & Şimşek, 2011). In this context, at the end of the focus group interviews, the reports prepared by the moderators working in each group were transferred to the computer environment, and the content analysis stage was initiated. The content analysis performed within the scope of the study was conducted using NVIVO 12 qualitative data analysis program. Codes were determined as a result of the content analysis of the raw data obtained from the interviews. At this point, the collected data were analyzed separately by two independent researchers, and the numbers of consensus and disagreement were determined by comparing the coding made by independent researchers. Reliability was calculated using these numbers using the formula ($\text{Reliability} = \frac{\text{Agreement}}{\text{Agreement} + \text{Disagreement}}$) suggested by Miles and Huberman (1994). The reliability among researchers was found to be 90%. The codes obtained are presented in the findings section of the study.

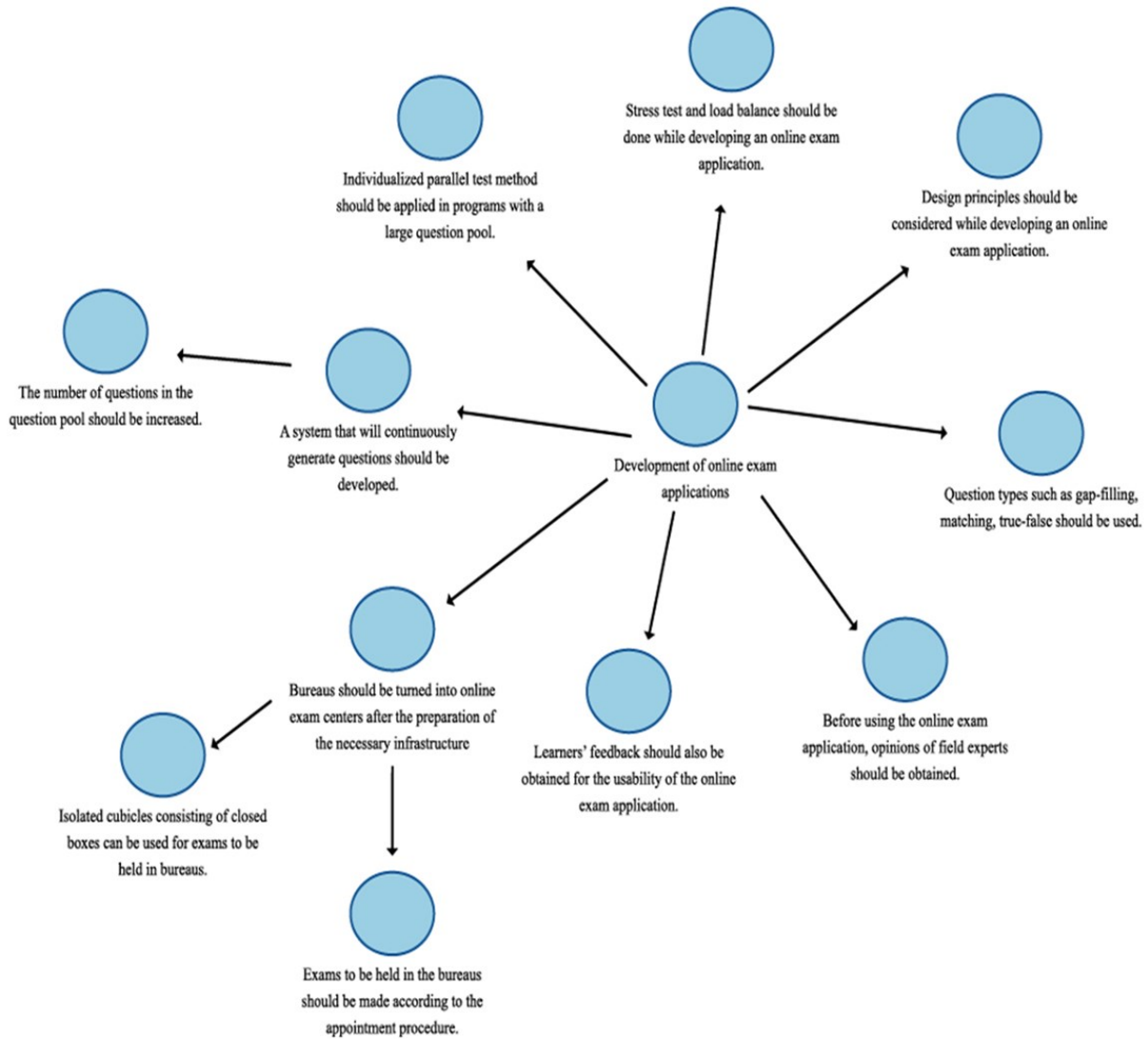
3. FINDINGS

The findings obtained in light of the questions asked to the participants within the scope of this study are presented below.

3.1. The Findings Obtained Regarding the Development of Online Test Applications During the COVID-19 Pandemic Period

The first question asked to the participants within the scope of the study was which points should be taken into consideration in the development of online test applications. The findings obtained in this context are presented in [Figure 1](#).

Figure 1. Recommendations for the development process of online test applications.



A total of 9 sub-themes were reached under the main theme of developing online test applications. Among these themes, the most repeatedly sub-theme was "Bureaus should be turned into online test centers after the preparation of the necessary infrastructure." The emergence of this sub-theme originates from the fact that the participants work at Anadolu University. The Open Education Faculty was established within Anadolu University in 1982. From 1982 to the present day, open education bureaus have been working throughout the country to provide services to learners within the scope of the Open Education Faculty with more than three million students, both active and passive. As of 2020, a total of 102 open education bureaus have been serving actively across Turkey. In this context, it can be stated that this theme, obtained for the use of bureaus as examination centers within the scope of the study, originates from the experience of the participants of Anadolu University Open Education Faculty. The prominent views obtained under this theme are as follows:

"Online tests supervised in bureaus can be conducted by appointment. Thus, bureaus can be turned into online test centers. At this point, necessary infrastructure should be prepared in bureaus." Cengiz

"Isolated cubicles consisting of closed boxes can be used in the online test system to be realized in bureaus. Students can take an exam by making an appointment in a supervised way at these points." Hasan

The subjects emphasized under the sub-theme "Bureaus should be turned into online test centers after the preparation of the necessary infrastructure" were the subjects of using isolated cubicles consisting of closed boxes in online tests to be carried out through bureaus and conducting exams to be held in bureaus according to the appointment method. In this context, it was also stated that the examination process should be spread over a wide period. Another sub-theme emphasized under the main theme of the development of online test applications was the theme "A system must be developed to generate questions continuously." Some of the opinions obtained in this context are as follows:

"The number of questions in the question pool should be increased so that online tests can be applied efficiently." Ahmet

"The more questions there are in the question pool, the more applicable online tests will be." Esra

"A system that will continuously generate questions by spreading the question preparation process over time should be developed." Cihan

Another sub-theme obtained under this main theme was the types of questions to be asked in exams. In this context, the prominent opinions among those obtained under the sub-theme "Question types such as gap-filling, matching, true-false should be used" are as follows:

"The opportunities of technology can be used in online tests. For example, question types such as gap-filling, matching, true/false questions can be diversified." Firat

"In addition to multiple-choice questions, other question types should also be included in technology-based online tests." Ali

In addition to these sub-themes obtained under the main theme of "Developing online test applications," the sub-themes of "Stress test and load balance should be performed while developing an online test application" and "The opinion of field experts should be obtained before using the online test application" were also included. In this context, field experts emphasized the necessity of performing all necessary tests before using online test applications and consulting experts in the field in this process. Moreover, under the sub-theme "For the usability of the online test application, students' opinions should also be obtained," it was stated that learners' opinions should also be consulted to see the equivalent of the online test application used in learners. Finally, under the sub-theme of "Design principles should be taken into consideration while developing an online test application," it was mentioned that design principles should be considered for the developed exam system to be user-friendly.

3.2. The findings obtained to ensure the security of online test applications carried out during the COVID-19 pandemic period

The second question asked to the participants within the scope of the study was about the factors that should be considered to ensure the security of online test applications. The findings obtained in this context are presented in Figure 2. 15 sub-themes were reached under the main theme of the security of online test applications. The themes were about how to hold the exams to be applied in online environments securely. The sub-themes obtained in this context focused on preventing the students who would take an exam from receiving help from anyone other than themselves. Accordingly, the most remarkable themes were the following themes: "Behavioral biometric techniques such as writing style on the keyboard, speaking style, and signing style should be used," "Information security processes such as identification, authentication and authorization should be used," "Physiological biometric techniques such as fingerprint recognition, eye-iris recognition should be used," "The full-screen lock application should be used," "Browser lock should be functionalized," and "360-degree cameras should be used." The opinions obtained within the scope of these themes were expressed to prevent

learners who would take an exam in online environments from receiving help from anyone other than themselves. The most repeated among these opinions are as follows:

"To prevent learners who will take an exam in online environments from receiving help from others, biometric techniques such as writing style on the keyboard, speaking style, and signing style should be used." Resul

"By using full-screen lock and browser lock applications, we can eliminate the possibility of learners accessing screens other than the exam screen." Cengiz

"If a 360-degree camera system is installed, the exam environment can be completely controlled." Anil

"Physiological biometric techniques such as fingerprint recognition and eye-iris recognition can be used to be sure of the student's identity in online tests to be carried out through bureaus." Tuna

Figure 2. Suggestions for the security of online test applications.



Furthermore, for learners not to receive help via other pages over the Internet or through any application during the exam, opinions were obtained under the following themes "A certain time limit should be set for each question," "Maximum 3 rights should be granted to enter the exam," "New questions should be directed in the new session when the exam is exited before its completion," "The session should be automatically terminated when the exam time is over,"

"Exam records (logs) should be followed through the learning management system," "Returning to the previous question-questions should be prevented." At this point, it can be stated that the experiences gained at Anadolu University played a role in the themes obtained. In the online test application applied by Anadolu University Open Education Faculty during the COVID-19 period, when a student enters the online test system, he/she can see all the courses in the relevant semester. He/she can choose for which course he/she wants to take an exam. Before the online test starts, the necessary exam rules are presented to the student.

When the exam is started, the internet browser opens in full screen, and the time starts to run. The system ends the exam when the student tries to exit the browser and open another web browser. The student continues the exam with the remaining question and the remaining time by logging into the system again. Only in this way, there is a right to a total of three re-entries. The right to re-enter the system has been defined in a limited number for students not to suffer due to browser problems, internet, electric cut-outs, etc. When the student starts the exam, each question comes up once. If the student leaves the question blank, he/she cannot return to the relevant question again. He/she saves the relevant answer of each question by marking it and moves on to the next question. No return can be made to the question left blank and answered. Therefore, it can be stated that the experiences gained by the participants during the online test application play a role in these themes. The prominent opinions among the opinions obtained are as follows:

"If a certain time limit is set for each question, learners will not have the opportunity to obtain information from other pages or other applications. At this point, field experts should determine how much time will be given to which question." Okan

"The application that allows cheating in the exam currently applied is that students can exit the exam and take the exam again from where they have left it. When the student exits the exam, new sessions and new questions should be directed, and a maximum of 3 rights should be granted." Masal

"By examining the exam logs kept by learning management systems, information about how long learners are browsing outside the exam screen can be obtained." Gülay

"Not allowing to return to the previous question will minimize the chance for learners to find answers to questions from elsewhere." Mustafa

Moreover, under the theme "The random distribution of questions and answers should be ensured in the application of the exam for the same course," it was emphasized that the possibility of learners taking the exam at the same time and in the same course to answer questions together should be eliminated. Furthermore, under the theme "Not all of the exam questions should be published, but a certain part," it was emphasized that the question pools prepared for use in online tests should consist of more and better-quality questions.

4. DISCUSSION and CONCLUSION

The first findings obtained within the scope of this study, in which the opinions of field experts were obtained on the development of online test applications that became more needed with the COVID-19 pandemic period and the security of these applications, were on the points that should be taken into consideration during the development of online test applications. In this context, it was concluded that visual design principles should come to the forefront in developing online test applications. It is an important element to use visual design principles to serve the purpose of the online tests to be carried out and to have a user-friendly interface. In a study performed by Albayrak (2014) on this issue, it was concluded that learner achievement was higher in online tests designed by considering visual design principles. In the studies conducted by Ortner and Caspers (2001) and Clough (2008), it was revealed that learners' achievement was affected by the quality of the exam. In another study carried out by Yağcı et

al., (2015), it was mentioned that an online test application developed in line with the visual design principles would reduce learners' anxiety about taking exams in online environments. Therefore, it can be said that visual design principles should be put into practice during the development of online test applications.

Another result obtained within the scope of the study was related to the necessity of receiving opinions from both field experts and learners in developing online test applications. In this context, after the online test application is developed, it should be opened to field experts before it is put into use, and it was concluded that opinions on the usability of the application should be obtained from them. In addition to this, after the developed application was put into use, it became necessary to obtain students' opinions about the experiences they gained during their use of the application. In this way, the deficiencies of the developed application will be determined, and studies will be carried out to eliminate the identified deficiencies by taking opinions from both field experts and learners. In the studies conducted by Koçak et al. (2006), Saban et al. (2010), and Yılmaz (2016) on this issue, it was emphasized that the views of students, which should be at the focus of learning and teaching activities, should be attached importance in the process of developing online test applications. A study performed by Sırakaya, Sırakaya and Çakmak (2015) investigated the attitude levels of distance learners toward online tests. At the end of the study, it was concluded that most of the learners had a positive opinion about online tests. According to Kınalıoğlu and Güven (2011), it is important to obtain opinions and suggestions from field experts in the process of developing online test applications for the application to be successful. Another important result reached at this point is that the stress test and load balance of the developed online test application should be performed. In line with this result obtained, Wang (2017) emphasized that stress testing and load balance should be considered to check the usability of online test applications.

Another result obtained within the scope of the opinions of field experts was on the question types to be used in online test applications. Accordingly, while developing online test applications, attention should be paid to the use of question types such as gap-filling, short-long answer, matching, true-false, in addition to multiple-choice questions. In line with this result, Borich (2013) indicated that question types such as true-false, matching, multiple-choice, completion, open-ended questions should be used in online tests.

Another issue emphasized during the development of online test applications was related to expanding the question pool. In line with this, in a study conducted on the design of online tests, Jiang et al. (2019) stated that more questions produced by field experts using various question types were important for the usability of online test applications, so by providing the diversity of questions, applications such as the individualized parallel testing method could be used. Therefore, it can be stated that the number of questions should be higher for a more efficient online testification process. In this context, under the theme of "a system that will continuously generate questions should be developed" obtained within the scope of the study, field experts emphasized that the number of questions should be high.

Another result obtained about the usability of online test applications is that after the necessary infrastructure is provided, online tests must be conducted in exam centers to be determined in certain regions. In this context, it was concluded that bureaus could be used within Anadolu University, the institution where the study was conducted. Likewise, it can be stated that other institutions can determine specific examination centers in certain regions and carry out online tests. At this point, the issues that should be focused on are providing the necessary infrastructure and conducting exams according to the appointment method. In a study conducted on online tests, Yağcı et al. (2015) stated that online tests could be held entirely in web-based environments or physical spaces such as exam centers, high schools, and conference halls. Therefore, it can be said that online test applications can be carried out in web

environments or physical spaces, considering the requests and infrastructure possibilities of institutions.

Another research question for which the opinions of field experts were obtained within the scope of the study was on the security of online tests. Based on the findings obtained in this context, the first result was on the use of biometric and physiological biometric techniques. It was concluded that measures such as writing style on the keyboard, speaking style, and signing style could be used within the scope of biometric techniques. It is stated that techniques such as fingerprint recognition, face recognition, and eye-iris recognition could be used within the scope of physiological biometric techniques. In this context, it can be said that biometric techniques can be used in online tests held in the web environment, and physiological biometric techniques can be used in online tests held through exam centers such as bureaus, high schools, and conference halls. In this way, the security of online test applications can be further increased, and behaviors such as cheating and plagiarism can be prevented. In line with these results obtained within the scope of the study, Kaya (2016) and Patil, Sharma and Patil (2019) mentioned that physiological biometric techniques such as face recognition could be used effectively in online test applications. Likewise, Bozkurt and Uçar (2018) and Senthil Kumar and Rathi (2021) emphasized that biometric and physiological biometric techniques should be used to eliminate limitations such as cheating and plagiarism in online tests.

It was concluded that another measure to ensure the security of online test applications was to employ information security processes such as identification, authentication, and authorization. The use of biometric and physiological biometric techniques and information security processes will eliminate the situation when someone other than the student who will take the exam in online environments will take the exam. Therefore, a more reliable and appropriate examination process will be carried out. Similar to this result, in the study conducted by Parker (2010), it was emphasized that possible threats to confidentiality, integrity, and compliance could be prevented by using information security processes in online tests. Considering that cheating incidents have increased in recent years (Jisc, 2020), it can be stated that it is a necessity to take security measures in this way.

Another dimension that would ensure the security of online tests within the scope of the study was the measures to be taken at the time of the exam. In this context, issues such as using the browser lock, applying the full-screen lock, granting up to 3 rights to enter the exam, directing new questions in the new session when exiting the system before the exam is completed, closing the system after the exam period are over, examining records through the learning management system, the inability to return to the previous question, and setting the time limit for each question were emphasized. Thanks to these measures to be taken, it is aimed to prevent learners from seeking answers to questions through another application or another Web page. Furthermore, giving the maximum of 3 rights to learners who exit the exam for any reason before completing the exam and updating the remaining questions as new questions when they take the exam again is another type of security that can be applied. Similar to these results obtained, Can (2020) and Karthika et al., (2019) stated that measures such as setting a certain time limit for each question, using screen and browser locks, and not displaying previous questions could be used in online test applications.

Another result obtained in ensuring the security of online tests is the use of a 360-degree camera system. In this context, the 360-degree camera application, including an audio system, it aims to fully control the environment in which students take online tests. In this way, learners will be prevented from attempting to cheat. In line with this result, Golden and Kohlbeck (2020) and Hylton et al. (2016) stated that the webcam-based exam system was useful for preventing misuse in online tests. Furthermore, in a study performed by Hoque, Ahmed, Uddin and Faisal

(2020), 360-degree cameras were used in the online test application process and were considered a successful application in terms of their results.

As a result, the increasing use of online learning environments, especially with the COVID-19 pandemic process, has become a situation that increases the importance of online test applications in evaluating the education and training process. Considering that instructor and learners exhibit positive attitudes toward online test applications (Can, 2020; Fatmasari, 2020, Gül, 2012; Harnkajornsuk et al., 2019; Sırakaya et al., 2015; Yılmaz, 2016), it emerges as an important point that these applications are usable and reliable (Kundu & Bej, 2021). However, the fact that institutions were generally unprepared for the emergency distance education applications implemented during the pandemic (Bozkurt, 2020; Senel & Senel, 2021) made it difficult to realize a healthy measurement-evaluation process. In this context, seeking the opinions of experts who have experience in online test applications during the emergency distance education period will guide the institutions on how these applications should be done and how to obtain reliable results. Therefore, as mentioned in the introduction section of the study, this can be shown as an example of the new perspectives brought to education and training practices with the COVID-19 pandemic process.

4.1. Recommendations

The recommendations that can be made based on the results obtained within the scope of the study are listed in the following way:

- Technological infrastructure systems and workforce requirements should be met by institutions to evaluate short and long-answer question types that can be used in online test applications.
- Assessment-evaluation, research and development (R&D) units should be established for technological infrastructures institutions should provide, and individuals with high technological knowledge should be employed in these units.
- While the question pool is increased during the development of online test applications, field domain experts should check the scope, validity, and reliability of the questions to be prepared. For this process, institutions should establish units within their organization and benefit from the knowledge and experience of field experts through these units.
- Individuals who will supervise the exam environment will be needed during the 360 -degree camera application process. In this regard, it is important to conduct exams according to the appointment procedure and use the workforce efficiently.
- Obtaining learners' biometric and physiological biometric properties to ensure the security of exams in online environments brings about ethical concerns. In this context, it is important to attach importance to data confidentiality and act in line with ethical rules.
- Learning processes should be addressed as a whole, and a process should be evaluated at the point of assessment and evaluation. In this context, in addition to a midterm and final exams to be carried out in online environments, elements such as data to be obtained from learning analytics, assignments, activities in the discussion forum should be evaluated in assessment and evaluation processes.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research and publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). Ethics Committee Approval and its number should be given by stating the institution name which gave the ethical approval. Ethics Committee Number: Anadolu University, 60779.

Authorship Contribution Statement

Hakan Kilinc: Investigation, Introduction, Importance of the Study, Purpose of the Study, Method, Data Analysis, Discussion, Conclusion, and Recommendations. **Muhammet Recep Okur:** Investigation, Method, Data Analysis, Visualization, Software, Findings, Discussion, Conclusion, and Recommendations. **Ilker Usta:** Investigation, Data collection, Discussion, Conclusion, and Recommendations

ORCID

Hakan KILINC  <https://orcid.org/0000-0002-4301-1370>

Muhammet Recep OKUR  <https://orcid.org/0000-0003-2639-4987>

Ilker USTA  <https://orcid.org/0000-0002-6403-6294>

5. REFERENCES

- Abduh, M. Y. M. (2021). Full-time online assessment during covid-19 lockdown: EFL teachers' perceptions. *Asian EFL Journal*, 28(1), 1-21.
- Albayrak, E. (2014). Elektronik Ortamlardaki Sınavlarda Tasarım Etmenlerinin Öğrencilerin Başarıları ve Elektronik Sınav Kaygılarına Etkisi [The Effects of Design Factors on Students' Success and Test Anxiety in Electronic Tests]. *International Online Journal of Educational Sciences*, 6(2), 460-474.
- Anderson, H. M., Cain, J., & Bird, E. (2005). Online course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education*, 69(1), 34-43.
- Angus, S.D., & Watson J. (2009) Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology*, 40, 255-272. <https://doi.org/10.1111/j.1467-8535.2008.00916.x>
- Bloor, M., Frankland, J., Thomas, M., & Robson, K. (2001). *Focus Groups in Social Research*. SAGE.
- Borich, G. D. (2013). *Effective teaching methods (8 ed.)*. England: Person Education.
- Bowling, A. (2002). *Research Methods in Health: Investigating Health and Health Services*. McGraw-Hill House.
- Bozkurt, A. (2020). Koronavirüs (Covid-19) pandemi süreci ve pandemi sonrası dünyada eğitime yönelik değerlendirmeler: Yeni normal ve yeni eğitim paradigması [Coronavirus (Covid-19) pandemic process and educational evaluations in the post-pandemic world: New normal and new education paradigm]. *AUAd*, 6(3), 112-142.
- Bozkurt, A., & Sharma, R. C. (2020). Emergency remote teaching in a time of global crisis due to CoronaVirus pandemic. *Asian Journal of Distance Education*, 15(1), i-vi. <https://doi.org/10.5281/zenodo.3778083>
- Bozkurt, A., & Uçar, H. (2018). E-Öğrenme ve e-sınavlar: Çevrimiçi ölçme değerlendirme süreçlerinde kimlik doğrulama yöntemlerine ilişkin öğrenen görüşlerinin incelenmesi [E-Learning and e-exams: Examination of learner views on identity verification methods in online assessment and evaluation processes]. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, 14(2), 745-755. <https://doi.org/10.17860/mersinefd.357339>
- Can, E. (2020). Koronavirüs (Covid-19) pandemisi ve pedagojik yansımaları: Türkiye'de açık ve uzaktan eğitim uygulamaları [Coronavirus (Covid-19) pandemic and pedagogical implications: open and distance education applications in Turkey]. *AUAd*, 6(2), 11-53.
- Clough, S. J. (2008). *Computerized versus paper and pencil assessment of socially desirable responding: score congruence, completion time, and respondent preferences* [Unpublished doctoral dissertation]. The University of Iowa, USA.
- Creswell, J. W. (2007). *Qualitative inquiry and research method: Choosing among five approaches*. Academic Press. Sage.

- d'Orville, H. (2020). COVID-19 causes unprecedented educational disruption: Is there a road towards a new normal? *Prospects*, 2020(49), 11-15. <https://doi.org/10.1007/s11125-020-09475-0>
- Doghonadze, N., Aliyev, A., Halawachy, H., Knodel, L., & Adedoyin, A. S. (2020). The Degree of Readiness to Total Distance Learning in the Face of COVID-19-Teachers' View (Case of Azerbaijan, Georgia, Iraq, Nigeria, UK and Ukraine). *Journal of Education in Black Sea Region*, 5(2), 2-41. <https://doi.org/10.31578/jeb.v5i2.197>
- Ersoy, A. (2013). Türk öğretmen adaylarının kültürlerarası deneyimlerinde karşılaştıkları sorunlar: Erasmus değişim programı örneği [Problems faced by Turkish teacher candidates in their intercultural experiences: Erasmus exchange program example]. *Eğitim ve Bilim*, 38(168),154-166.
- Fatmasari, R. (2020). Student satisfaction on distance education academic services. *In International Conference on Education, Science and Technology* (pp. 31-37). Redwhite Press.
- Golden, J., & Kohlbeck, M. (2020). Addressing cheating when using test bank questions in online Classes. *Journal of Accounting Education*, 52(2020), 1-14. <https://doi.org/10.1016/j.jaccedu.2020.100671>
- Gülbahar, Y. (2013). E-değerlendirme.[E-evaluation], K. Çağıltay, Y. Göktaş. (Eds.). *Öğretim Teknolojilerinin Temelleri: Teoriler, Araştırmalar, Eğilimler [Foundations of Instructional Technology: Theories, Research, Trends]*, pp. 651-663. Pegem Akademi Yayıncılık
- Harnkajornsuk, S., Chinda, B., Witayasakpan, S., Wongboonnak, S., & Bunto, P. A. S. (2019). Development of a Web-based Online Examination for Screening Gifted Students. *In Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 56-60).
- Hewson, C. (2012). Can Online Course-Based Assessment Methods Be Fair and Equitable? Relationships between Students' Preferences and Performance Within Online and Offline Assessments. *Journal of Computer Assisted Learning*, 28(5), 488-498. <https://doi.org/10.1111/j.1365-2729.2011.00473.x>
- Hoque, M. J., Ahmed, M. R., Uddin, M. J., & Faisal, M. M. A. (2020). Automation of Traditional Exam Invigilation using CCTV and Bio-Metric. *International Journal of Advanced Computer Science and Applications*, 11(6), 392-399.
- Hylton, K., Levy, Y., & Dringus, L. P. (2016). Utilizing webcam-based proctoring to deter misconduct in online exams. *Computers & Education*, 92(2016), 53-63. <https://doi.org/10.1016/j.compedu.2015.10.002>
- Jiang, J., Wu, B., Chang, L., Liu, K., & Hao, T. (2019). The Design and Application of an Web-Based Online Examination System. *International Symposium on Emerging Technologies for Education* (pp. 246-256). Springer, Cham.
- Jisc (2020). *The future of assessment: five principles, five targets for 2025*. <https://repository.jisc.ac.uk/7733/1/the-future-of-assessment-report.pdf>
- Jordan, S., & Mitchell T. (2009) e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. *British Journal of Educational Technology*, 40(2), 371-385. <https://doi.org/10.1111/j.1467-8535.2008.00928.x>
- Karabekir, M., Tozlu, E., & Şencan, M. N. M. (2015). Girişimci Adayı Üniversite Öğrencilerinin Girişimcilik Özelliklerinin Odak Grup Görüşmesi ile İncelenmesi [Investigation of Entrepreneurial Characteristics of University Students who are Entrepreneur Candidates with a Focus Group Meeting]. *SDÜ Fen Edebiyat Fakültesi, Sosyal Bilimler Dergisi*, 1(35), 203-216.

- Karthika, R., Vijayakumar, P., Rawal, B. S., & Wang, Y. (2019). Secure Online examination System for e-learning. *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)* (pp. 1-4). IEEE.
- Kaya, Z. (2016). *Biyometrik Güvenlik Sistemleri ve Yüz Tanımaya Dayalı Çevrimiçi Sınav Sistemi [Biometric Security Systems and Online Exam System Based on Face Recognition]* [Unpublished doctoral dissertation]. İstanbul Aydın University.
- Kınalıoğlu, İ. H., & Güven, Ş. (2011). *Uzaktan Eğitim Sisteminde Öğrenci Başarısını Ölçülmesinde Karşılaşılan Güçlükler ve Çözüm Önerileri [Difficulties Encountered in Measuring Student Achievement in Distance Education System and Solution Suggestions]*. XIII. Akademik Bilişim Konferansı, Malatya.
- Kitzinger, J. (1995). Qualitative research: introducing focus groups. *British Medical Journal*, *31*(7000), 299-302.
- Koçak, Ş., Yenilmez, E. D., & Yenilmez, E. (2006). Çevrimiçi Sınav Sistemlerinin Öğrenmeye Olan Etkileri Üzerine Bir Çalışma: Öğrenci Görüşleri [A Study on the Effects of Online Exam Systems on Learning: Student Views]. *Çukurova Üniversitesi İlahiyat Fakültesi Dergisi*, *6*(2), 171-189.
- Krueger, R.A. (1994). *Focus Groups: A Practical Guide For Applied Research*. SAGE.
- Kuhtman, M. (2004). Review of online student ratings of instruction. *College and University Journal*, *80*(1), 64-67.
- Senthil Kumar, A.V., & Rathi, M. (2021). Keystroke Dynamics: A Behavioral Biometric Model for User Authentication in Online Exams. In M. Khosrow-Pour (Ed.), *Research Anthology on Developing Effective Online Learning Courses* (pp. 1137-1162). IGI Global.
- Kundu, A., & Bej, T. (2021). Experiencing e-assessment during COVID-19: an analysis of Indian students' perception. *Higher Education Evaluation and Development*, *15*(2), 114-134. <https://doi.org/10.1108/HEED-03-2021-0032>
- Liu, M., Papathanasiou E., & Yung-Wei H. (2001) Exploring the use of multimedia examination formats in undergraduate teaching: results from the fielding testing. *Computers in Human Behavior*, *17*(3), 225-248. [https://doi.org/10.1016/S0747-5632\(01\)00008-5](https://doi.org/10.1016/S0747-5632(01)00008-5)
- Marriot, P. (2009). Students' Evaluation of The Use of Online Summative Assessment on an Undergraduate Financial Accounting Module. *British Journal of Educational Technology*, *40*(2), 237-254. <https://doi.org/10.1111/j.1467-8535.2008.00924.x>
- Ortner T. M. & Caspers, J. (2001). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, *27*(3), 157-163.
- Özen, Z. (2016). *Kimlik Doğrulaması için Tuş Vuruş Dinamiklerine Dayalı Bir Güvenlik Sisteminin Yapay Sinir Ağları ile Geliştirilmesi [Development of a Security System Based on Keystroke Dynamics for Authentication with Artificial Neural Networks]* [Unpublished doctoral dissertation]. İstanbul University.
- Parker, D. (2010). Our excessively simplistic information security model and how to fix it. *ISSA Journal*, *8*(7), 12-21.
- Patil, S., Sharma, Y. K., & Patil, R. (2019). Implications of Deep Learning-Based Methods for Face Recognition in Online examination System. *International Journal of Recent Technology and Engineering*, *8*(3), 14-27.
- Saban, A., Özer, H. İ., & Tümer, A. E. (2010). Çevrimiçi ders materyalleri ve çevrimiçi sınav sistemi ile ilgili öğrenci görüşleri [Student views on online course materials and online exam system]. *E-Journal of New World Sciences Academy*, *5*(4), 2238-2244.
- Schmidt, R. A., Lee, T. D., Winstein, C. J., Wulf, G., Zelaznik, H. N. (2019). *Motor control and learning: a behavioral emphasis*. Human Kinetics.

- Senel, S., & Senel, H. C. (2021). Remote Assessment in Higher Education during COVID-19 Pandemic. *International Journal of Assessment Tools in Education*, 8(2), 181-199. <https://doi.org/10.21449/ijate.820140>
- Şimşek, İ., Balaban, M., & Ergin, H. (2016). Eğitimde Ölçme ve Değerlendirme Çalışmalarında Web Tabanlı Uzman Sınav Sisteminin Kullanımı Üzerine Bir Araştırma [A Research on the Use of Web Based Examination System in Education Measurement and Evaluation Studies]. *Hasan Ali Yücel Eğitim Fakültesi Dergisi*, 13(26), 165-179.
- Sindre, G., & Vegendla, A. (2015). E-exams versus paper exams: A comparative analysis of cheating-related security threats and countermeasures. *NISK Journal*, 8(1), 34-45.
- Sırakaya, M., Sırakaya, D. A., & Çakmak, E. K. (2015). Uzaktan Eğitim Öğrencilerinin Çevrimiçi Sınava Yönelik Tutum Düzeylerinin İncelenmesi [Investigation of Distance Education Students' Attitudes towards Online Exam]. *Kastamonu Eğitim Dergisi*, 23(1), 87-104.
- Solak, H.İ., Ütebay, G., & Yalçın, B. (2020). Uzaktan eğitim öğrencilerinin basılı ve dijital ortamdaki sınav başarılarının karşılaştırılması [Comparison of distance education students' exam success in print and digital media]. *AUAd*, 6(1), 41-52.
- UN (1948). *Universal Declaration of Human Rights*. <https://www.un.org/en/universal-declaration-human-rights/index.html>
- UNESCO (2020a). *School closures caused by Coronavirus (Covid-19)*. <https://en.unesco.org/covid19/educationresponse>
- UNESCO (2020b). *Exams and assessments in COVID-19 crisis: fairness at the centre*. <https://en.unesco.org/news/exams-and-assessments-covid-19-crisis-fairness-centre>
- UNESCO (2020c). *Distance Learning Strategies, What do we know about effectiveness?* <http://www.unesco.org/covid19EDwebinar>
- Wang, F. (2017). *Research on the paperless examination in the university public computer laboratory*. Proceedings of the 2017 International Conference on E-Society, E-Education and E-Technology (pp. 52-55). ACM Digital Library.
- Yağcı, M., Ekiz, H., & Gelbal, S. (2011). *Çevrimiçi sınav ortamlarının öğrencilerin akademik başarılarına etkisi* [The effect of online exam environments on students' academic success]. 5th International Computer & Instructional Technologies Symposium, Fırat University, Elazığ, Turkey.
- Yağcı, M., Ekiz, H., & Gelbal, S. (2015). Yeni Bir Çevrimiçi Sınav Modeli Geliştirilmesi ve Uygulanması [Developing and Implementing a New Online Exam Model]. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, 16(1), 269-288.
- Yıldırım, A., & Şimşek, H. (2008). *Sosyal Bilimlerde Nitel Araştırma Yöntemleri (6. Baskı)* [Qualitative research methods in the social sciences. (6th Edition)]. Seçkin Yayınevi.
- Yıldırım, A., & Şimşek, H. (2011). *Sosyal Bilimlerde Nitel Araştırma Yöntemleri (8. Basım)* [Qualitative research methods in the social sciences. (8th Edition)]. Seçkin Yayınları
- Yılmaz, Ö. (2016). Çevrimiçi sınav görüş anketi [Online examination assessment survey]. *e-Kafkas Eğitim Araştırmaları Dergisi*, 3(3), 26-33.
- Yin, R. K. (1984). *Case Study Research: Design and Methods*. Sage Publications.
- Yin, R. K. (2003). *Case study research: Design and methods (3rd ed.)*. Sage.
- Zhao, Y. (2020). COVID-19 as a catalyst for educational change. *Prospects*, 49(1), 29-33. <https://doi.org/10.1007/s11125-020-09477-y>

Fuzzy logic expert system for evaluating the activity of university teachers

Vasile Florin Popescu^{1,*}, Marius Sorin Pistol²

¹National Defense University, Faculty of Security and Defense, Dep. Information Systems and Cyber Defense Romania

²European Commission, European Asylum Support Office, Malta

ARTICLE HISTORY

Received: Mar. 09, 2020

Revised: Oct. 31, 2021

Accepted: Nov. 21, 2021

Keywords:

Fuzzification,
Defuzzification,
Inference,
Mamdani,
Natural language.

Abstract: Assessing the performance of academics at different levels is increasingly difficult to achieve using traditional methods based mainly on numerical scores in evaluating teaching and research activity. The indexing of academic performance in various international databases with impact indices at different scales has led to the need for advanced computer models, such as expert systems based on fuzzy logic, proposed in this research, which address the evaluation of teachers even in the face of imprecise information and under conditions of uncertainty. In this research, as a contribution and novelty, a fuzzy logic model was developed in which an algorithm was simulated and implemented in Matlab using the Mandami toolkit, which allows inference of the rules of fuzzy logic and visualization. 3D. The system implementation was done by software in Matlab environment, using systems with fuzzy Mandami logic. The result of this pilot study was to test and validate the proposed model through a graphical interface, giving the results according to minimum criteria and with additional explanations.

1. INTRODUCTION

Fuzzy systems are an alternative to the traditional methods of dealing with affiliation and logic, which have their origins in ancient Greek philosophy and applications in the field of artificial intelligence. Despite its long-lived origins, it is a relatively new field and therefore there is still plenty of room for research. The application of fuzzy logic as a simple method for deciding on an unambiguous evaluation of university teachers based on ambiguity, vagueness, imprecision, or lack of input information requires several numerical parameters to work in terms of what is considered "significant error" and "error variation," but the exact values of these numbers are not critical unless good performance is required. Fuzzy logic does not require very precise numerical inputs, in terms of evaluating university teachers, is inherently robust, and can handle any reasonable number of inputs, but the complexity of the system increases significantly with the number of inputs and outputs. Rules based on simple language, such as IF X and Y THEN Z, are used to describe the desired response of the system in terms of linguistic variables rather than mathematical formulas. Their number depends on the number of inputs and outputs and the goals of the designer in controlling the response. Fuzzy systems, including fuzzy logic and

*CONTACT: Popescu Florin ✉ popescuveve@gmail.com 📍 National Defense University, Faculty of Security and Defense, Dep. Information Systems and Cyber Defense, Romania

fuzzy set theory, represent a rich and important extension of standard logic used in higher education assessment. The mathematics developed based on these theories is consistent, and fuzzy logic can be a generalisation of classical logic. Applications that can be generated from or adapted to fuzzy logic are widespread and provide the ability to model conditions that are vaguely defined despite the concerns of classical logicians. Many systems can be modelled, simulated and even physically implemented using fuzzy systems, such as the present study.

On many websites, in many scientific articles, or in talks at scientific symposia and conferences, many of us have read or heard terms such as artificial neural networks, fuzzy logic, genetic algorithms, genetic programming, evolutionary computation, expert systems, gravity algorithm, ant algorithm, particle group optimization, multiagent systems, and others. All of these terms are concepts that describe various methods and techniques for solving problems of moderate and high complexity based on the simulation of intelligent behavior, and they are grouped under the umbrella of two terms that sound similar but usually refer to different things: Artificial Intelligence and Intelligent Computing. Although there are supporters of the idea that the two terms actually refer to the same thing, most opinions hold that Artificial Intelligence and Intelligent Computing, while having similar goals, have fundamental elements that distinguish them. Artificial intelligence (AI) is the older of the two terms. The term was first used in its current meaning in 1956 at a scientific symposium at Dartmouth College in Hanover, USA. The Father IA, John McCarthy, defined it as "the science and technology of creating intelligent machines" (McCarthy, 1959) in the form of hardware or software. One of the most commonly cited definitions is that of the study and development of intelligent agents, where an intelligent agent is understood to be an autonomous entity that observes and interacts with the environment in an attempt to achieve goals.

The second term, Intelligent Computing (CI) or Computational Intelligence, was first used in 1990 by the IEEE Neural Networks Council, founded ten years earlier, which became a IEEE society in 2001 and later changed its name to IEEE Computational Intelligence Society to include new areas of interest such as fuzzy systems and evolutionary computing. The field of AI can be defined as a collection of natural inspiration techniques and computational methods that are distinct from the traditional techniques associated with AI and intended for the creation of intelligent systems. This collection includes subfields such as artificial neural networks, fuzzy systems, evolutionary computation, machine learning, Bayesian reasoning and so on. The core technologies of IC include artificial neural networks (RNA), fuzzy systems (SF), and evolutionary computation (EC), as well as hybrid intelligent systems that incorporate these technologies and other related paradigms. A taxonomy of intelligent systems, with the main research areas and components that define them, is shown in [Figure 1](#).

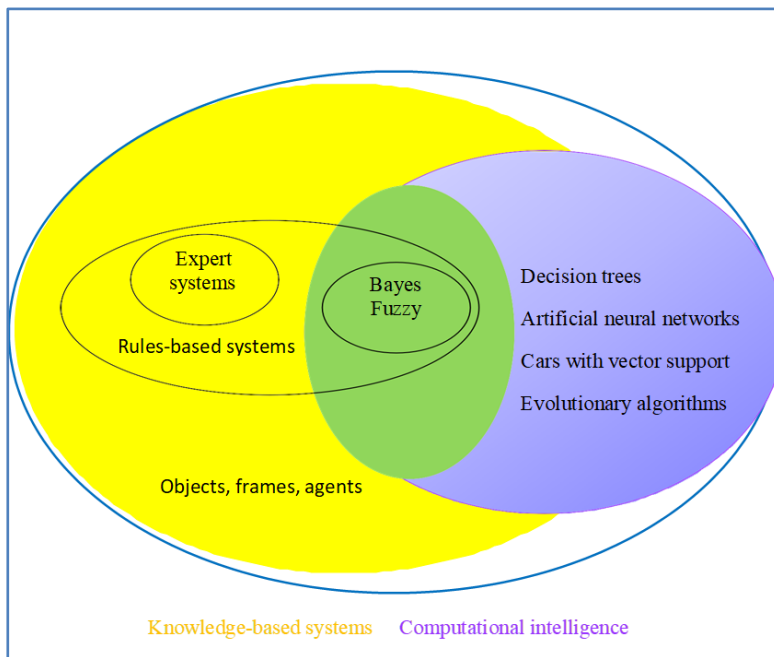
Sources of uncertainty:

- ✓ Imperfection of the rule
- ✓ Uncertainty of the evidence
- ✓ Confidence in the conclusion must be scaled
- ✓ Use of vague, imprecise language

Ways to express uncertainty:

- ✓ Probabilities
- ✓ Fuzzy logic
- ✓ Bayes' theorem
- ✓ Dumpster-Shafer theory

Figure 1. Intelligent systems.



Reasoning techniques in unsafe environments:

- ✓ Bayesian theory - probabilistic method
- ✓ Certainty theory
- ✓ Theory of possibility (fuzzy logic)
- ✓ Heuristic Methods

The possibility theory - a short history

- ✓ Parmenides (400 B.C.)
- ✓ Aristotle
 - "Law of the Excluded Middle" - every sentence must be True or False
- ✓ Plato
 - the third region between True and False
 - Lay the groundwork for fuzzy logic
- ✓ Lukasiewicz, & Tarski (1930)
 - Proposes a systematic alternative to Aristotle's bivalent logic - trivalent logic: True, False, Possible
- ✓ Zadeh, L. (1965).
 - Mathematically describes the theory of fuzzy sets and fuzzy logic: the membership function (True and False values) operates on the interval [0,1]
 - Proposed new computational operations for fuzzy logic
 - He considered fuzzy logic a generalization of classical logic
 - Published the first article about fuzzy crowds

When is it important to use fuzzy logic systems?

- ✓ Queries in natural language
- ✓ Representation of knowledge in expert systems
- ✓ Fuzzy control - when working with inaccurate phenomena (disturbed by noise)

2. METHOD

The model of fuzzy logic applied to the evaluation of university professors is an approximation method that can be used to formally model vague "knowledge" stored in a base of rules. The application of fuzzy logic in the evaluation of university employees at different levels is due to the advantages it offers in the following specific situations:

- Enables modelling of non-linear, complex, or imprecisely known processes for evaluation of university personnel, according to level;
- allows the implementation of the human experience of the evaluators, in this case in the construction of the inference rules, using the linguistic variables explained in the theoretical part.

2.1. Brief Description of the Concept of Fuzzy Logic, Mandami Model

The theoretical foundations of fuzzy logic were laid in 1965 by Lofti A. Zadeh, a professor at Berkely University in California. The term fuzzy logic was introduced by Zadeh at the same time as the proposal of fuzzy sets, but elements of fuzzy logic have been studied since 1920 (Garrido, 2012), reminiscent of the work of Łukasiewicz and Tarski, in which the so-called n -valued logic is proposed. However, the logic and fuzzy sets as they are known and used today are those proposed in the research of Zadeh.

In classical mathematics, an element is part of a set or not, whether it belongs to that set or not. In other words, the membership of elements in a given set is treated on a binary basis. Zadeh's theory of fuzzy sets, on the other hand, defines classical sets and their associated values in terms of crisp. Moreover, the new theory offers the possibility to evaluate step by step the membership of an element to a set by quantifying it using the so-called membership functions, which take values in the range $[0,1]$.

In fuzzy logic, the discrete values of Boolean logic (false and true) are replaced by a continuous membership function that takes values in the range $[0,1]$, where 0 stands for absolutely false and 1 for absolutely true. Consequently, an imprecise formulation has an associated truth value between 0 and 1.

For the new fuzzy sets, it was necessary to define the elementary operations for which it was proposed to use the complement against 1 for negation, the max operator for union and the min operator for intersection (Zadeh, 1996). Then fuzzy numbers, elementary algebraic operations with fuzzy numbers, fuzzy intervals and relations between fuzzy quantities were defined. Moreover, in order to maintain the connection with natural language and to allow a simple mathematical representation, the notion of modifier or qualifier, an equivalent to adjectives or adverbs in grammar, was introduced; thus, qualifiers are used such as: close, very, extremely, possible, with certainty, and so on.

The following are some important moments in the history of fuzzy systems. After the first moment in 1965, Professor Zadeh proposed the use of fuzzy algorithms in 1968 (Zadeh, 1968) and fuzzy decision systems in 1970 (Bellman & Zadeh, 1970). In 1971, he published the work *Quantitative Fuzzy Semantics* (Zadeh, 1971), in which he proposed the formal elements on the basis of which the methodology and the various types of applications of fuzzy logic were later developed. In 1973, he published a reference work (Zadeh, 1973) in which he defined linguistic variables and IF-THEN rules for the formation of knowledge bases. The first fuzzy controller for controlling an engine and a steam boiler was designed by Ibrahim Mandani in the mid-1970s (Mamdani & Assilian, 1975).

1987 is the year when the first commercial applications for different types of fuzzy controllers are developed and built, such as the fuzzy controller developed by Hitachi for the famous Japanese train Sendai, or those developed by Omron, another Japanese company that developed

the fastest fuzzy controller or the first fuzzy chip for microcomputers SUGE. Later, in 1993, the first application of fuzzy logic was registered for controlling a water treatment plant - also in Japan, of course.

In the mid-1980s, there were the first attempts at the theoretical foundation and practical development of fuzzy control systems based on fuzzy data sets combined with fuzzy learning. The foundations of fuzzy control systems are due to Professors Tomohiro Takagi of Meiji University (Tokyo) and Michio Sugeno of Doshisha University (Takagi & Sugeno, 1985).

Since the 1990s, applications of crowds and fuzzy logic in daily life have become more present and have developed rapidly. Numerous applications of this type can be found in the profile industry, the most famous being control systems for washing machines (Ahmed & Toki, 2016), ABS systems for braking (Subbulakshmi, 2014), autofocus systems for video cameras, elevator control systems (Patjoshi & Mohapatra, 2010) or philtres against spam messages (Vijayan et al., 2011).

The realisation of a flexible method for solving indeterminacy problems has been achieved through the development of fuzzy systems, which are based on fuzzy logic and are a special case of expert systems. Fuzzy logic works with the elements $A = \{x / x [0,1]\}$ and assigns a degree of belonging to the set to the object. The robustness of fuzzy logic is emphasised by the simultaneous control of numerical data and lexical knowledge (linguistic variables) by interpreting quantitative terms qualitatively.

The linguistic variable is a property and as a structure it includes (Chennakesava, 2008):

- a. The linguistic value u is an adverb, an adjective associated with the linguistic variable, indicating the name of the associated fuzzy set;
- b. The representation field U is a classical set on which fuzzy sets are defined. The set U is also called representation field, discourse universe or reference set;
- c. The membership function μ_F assigns to each element u the degree of membership in the fuzzy set F ;
- d. The degree of membership μ represents the extent to which an element belongs to a fuzzy set.

In order to understand the theory of fuzzy logic and the fuzzy set, it is necessary to present the elements on which it is based (Chuen, 1990). Let U be a set of objects with the general name $\{u\}$, which can be discrete or continuous. U is called the representation domain (universe of discourse) and u represents the generic elements of U . Denotation 1. Fuzzy set: a fuzzy set F contained in the representation domain U is characterized by the membership function μ_F which takes values in the range $[0, 1]$, i.e. $F: U [0,1]$.

The stages of the construction of a fuzzy logic system, the Mamdani model, are shown in [Figure 2](#). The theory of possibility, and implicitly Fuzzy Logic, is based on the following concepts:

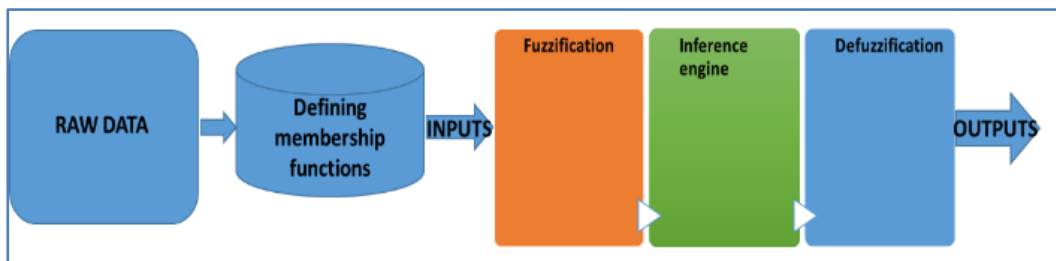
- Generalization of Boolean logic
- Manipulates the concept of partial truth
 - Classical logic - everything is expressed in binary terms
 - 0 or 1, white or black, yes or no
 - Fuzzy logic - the gradual expression of a truth.
 - Values between 0 and 1

The basic idea of this pilot study on the evaluation of academics using Fuzzy logic systems is:

- In accordance with the theory of certain information:
 - Florin Popescu is an associate professor

- In accordance with the theory of uncertain information:
 - ✓ In accordance with probability theory:
 - There is an 80% chance that Florin Popescu will be an associate professor
 - ✓ In accordance with fuzzy logic:
 - The degree of affiliation of Florin Popescu to the group of university lecturers is 0.8
- Definition of inputs and outputs by the expert
 - Gross input and output data
 - Fuzzification of input and output data
- Determining fuzzy variables and fuzzy sets based on membership functions
- Construction of a rule base by the expert
 - Decision matrix of knowledge base
- Evaluation of rules
 - Inference - transforming fuzzy inputs into fuzzy outputs by applying rules from knowledge base
- Defuzzification
- Interpretation of the results

Figure 2. Fuzzy logic system.



A. Fuzzification of input data

- Defining any set - 2 ways:
 - By enumerating the elements
 - ✓ Example: Crowd of students = {John, Stann, Brown}
 - By specifying a property of the set elements
 - ✓ Example: The set of numbers seems = $\{x \mid x = 2n, \text{ where } n - \text{natural number}\}$
- The characteristic function μ of a set
 - Let X be a universal set and x an element of the set ($x \in X$)
 - Classical logic
 - ✓ Let R be a subset of X: $R \in X$, R - regular set
 - ✓ Whether or not element x belongs to the set R
 - ✓ $\mu_R : X \rightarrow \{0, 1\}$, $e_R(x) = 1, x \in R / 0, x \notin R$
 - Fuzzy logic
 - ✓ Let F be a subset of X (discourse universe): $F \in X$, F – fuzzy sets
 - ✓ any element x belongs to the set F in a certain degree $\mu_F(x)$
 - ✓ $\mu_F : X \in [0, 1]$, $\mu_F(x) = g$, where $g \in [0, 1]$ degree of belonging of x to F
 - ✓ $g = 0 \in$ not belonging

A fuzzy set = a pair (F, μ_F) , where $\mu_F = \{1, \text{ if } x \text{ is total in } F; 0 \text{ if } x \text{ is not in } F / \text{ if } x \text{ is part of } F \text{ (x fuzzy number)}\}$.

2.2. Mamdani Model

- Basic idea:
 - a consequence of the rule is the form "output variable is part of a fuzzy set"
 - The result of the evaluation of the premises is applied for the function to which the consequence belongs.
 - Example: If x is in A and y is in B then z is in C.
- Typology (depending on how the result is applied to the function to which the consequence belongs)
 - Fuzzy sets result from clipped type
 - ✓ The function of belonging of the consequence is cut at the level indicated by the truth value of the result
 - Advantage: easy calculations
 - Disadvantage: possible information is lost
 - Fuzzy sets result of scaled type - Mamdani model
 - ✓ The function of belonging of the consequence is adjusted by scaling (multiplication) the truth value of the result
 - Advantage: less information is lost
 - Disadvantage: more complicated calculations.

Fuzzy Logic Toolbox™ software supports two types of fuzzy inference systems:

- Mamdani systems
- Sugeno systems

2.2.1. Mandami inference used

For simulation, modeling and validation of the results, 2 inference models were used in this research: inference (max-min) with a single rule and with multiple rules.

2.2.1.1. Mandami inference (max-min) with a single rule.

Modus Ponens generalized:

In fuzzy logic and rough reasoning, the most important rule of inference is generalized Modus Ponens.

In classical logic, this rule of inference is form $(p \wedge (p \rightarrow q)) \rightarrow q$, namely:

rule: if p, then q

premise: p

conclusion: q

In fuzzy logic, the corresponding inference rule is as follows:

rule: if x is A, then y is B

premise: x is A'

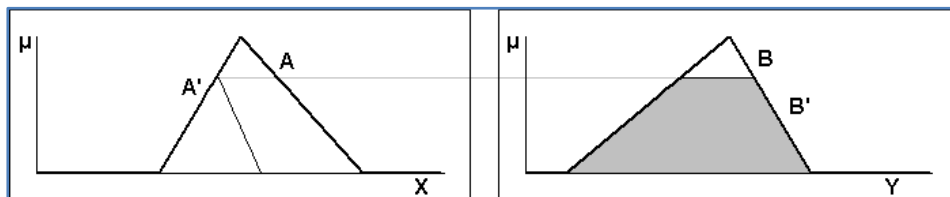
conclusion (consistent): and this B'

If $A' = A$ and $B' = B$, the rule is reduced to classical Modus Ponens. The matrix $A \rightarrow B$ is often denoted by R. The fuzzy inference process is seen as a transformation of one fuzzy set into another fuzzy set. The subset induced in B, B' is calculated as follows:

$$b_j' = \max(\min(a_i', r_{ij})).$$

There are several methods for defining the matrix R. In the following, we will use the Mandami inference type, whereby the set B' is a "cut" variant of B, at the height set by A' (see [Figure 3](#)). A' can be a normal subset of A, not just a single element.

Figure 3. *Mamdani type inference.*



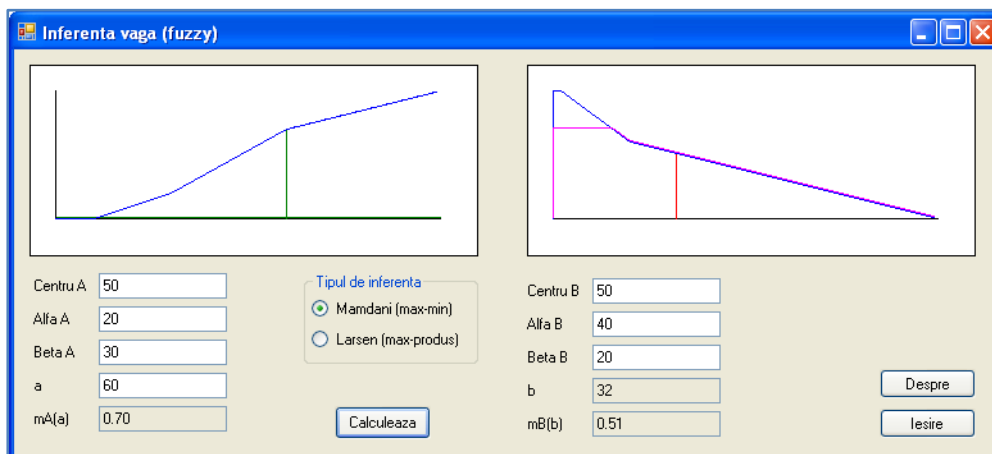
Defuzzification:

After determining the fuzzy set induced by an inference rule, in some applications a singular strict value must be determined based on this set. This process is called defuzzification. The most commonly used defuzzification technique is the centroid method (or centroid method):

$$x_{CG} = \frac{\sum_i x_i \cdot \mu_A(x_i)}{\sum_i \mu_A(x_i)}$$

In **Figure 4**, the program uses intervals (0,100) as universes of discourse. In the figure on the right, the blue bounded set is B, the pink bounded set is B' and the red marked value on the x-axis is the centroid.

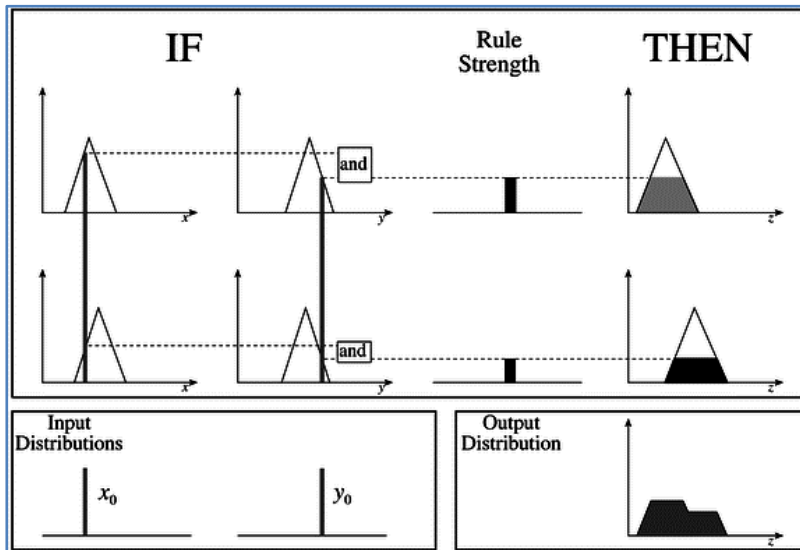
Figure 4. *Mamdani type inference and defuzzification.*



2.2.1.2. Mamdani inference with multiple rules. An example of a Mamdani-type fuzzy inference system is shown in **Figure 5**. To calculate the output of this system when the inputs are given, the following 6 steps must be followed:

1. It determines a lot of fuzzy rules;
2. The fuzzification of the entries is performed using the membership functions;
3. Combine the fuzzy inputs following the fuzzy rules to establish the activation powers of the rules;
4. Calculate the consistency of the rules by combining the activating powers of the rules with the membership functions of the outputs;
5. Combine consistencies to determine the output set;
6. Defuzzify the output set only if you want the output to be strict. The following is a detailed description of this process.

Figure 5. Mandami type fuzzy inference system with two rules and two strict inputs.



Creating fuzzy rules

Fuzzy rules are a set of statements that describe how the system can make a decision about estimating the output. Fuzzy rules have the following form: IF (input-1 is set-fuzzy-1) AND / OR (input-2 is set-fuzzy-2) AND / OR... THEN (output is set-fuzzy-output).

An example of a rule written in this way is the following: IF the project funding is sufficient AND the number of employees is low THEN the project risk is low. In the example in [Figure 6](#), there are two inputs x_0 and y_0 shown in the lower left corner. For these strict inputs, the degrees of affiliation are marked in the corresponding sets. Combining Multiple Antecedents When creating fuzzy rules, we use the operators AND, AND and sometimes NEGATION. The fuzzy operator AND is written as follows: $AB(x) T(A(x), B(x))$, where T is a function called the T-norm, $\mu_A(x)$ is the degree to which x belongs to set A , and $\mu_B(x)$ is the degree to which x belongs to set B . Although there are several ways to compute the function AND, the most commonly used is: $\min(\mu_A(x), \mu_B(x))$.

The fuzzy operator AND is a generalization of the Boolean logical operator AND in the sense that the truth value of a proposition is not just 0 or 1, but can be between 0 and 1. A T-norm function is monotone, commutative, associative and observes the conditions $T(0, 0) = 0$ and $T(x, 1) = x$.

The fuzzy operator OR is written as $AB S(A, B)$, where S is a function called T-conorm. Similar to the AND operator, this can be: $\max(\mu_A(x), \mu_B(x))$. The fuzzy operator OR is also a generalization of the Boolean logical operator OR to values between 0 and 1. A T-conorm function is monotonic, commutative, associative, and obeys the conditions $S(x, 0) = x$ and $S(1, 1) = 1$.

Calculation of consistencies:

First, the activation forces of the rules are calculated, as described earlier. In [Figure 6](#), it can be seen that the fuzzy operator AND is applied to the membership functions to calculate the activation forces of the rules.

Then, for a Mandami type fuzzy inference system, the output set is truncated at the level given by the activation force of the rule, as shown in the previous sections.

Aggregation of outputs:

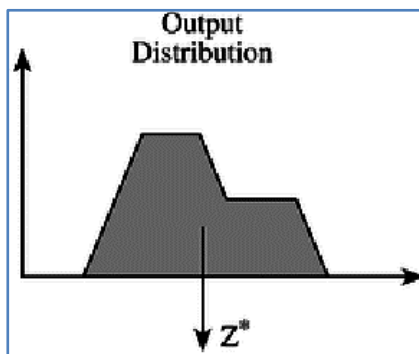
The outputs obtained after applying the fuzzy rules are combined to obtain the output set. This is usually done using the fuzzy operator OR. In [Figure 6](#), the membership functions on the right

are combined with the fuzzy operator OR to obtain the output set shown in the lower right corner.

Defuzzification:

It is often desirable to get a strict outcome. For example, if you are trying to classify handwritten letters on a blackboard, the fuzzy inference system must produce a strict number that can be interpreted. This number is obtained after the defuzzification process. The most commonly used defuzzification method is the Center of Gravity method, which is illustrated in [Figure 6](#).

Figure 6. Defuzzification using the center of gravity method.



2.3. Software Implementation of Fuzzy Algorithms

For the moment, we understand an algorithm to be a general method for solving a particular type of problem that can be implemented on a computer. In this context, an algorithm is the absolute essence of a routine. The algorithm underlying the fuzzy automaton contains an algebraic part in addition to the logical part. Thus, it is a mixed type algorithm organized as a finite sequence of steps and includes several specific operations. They meet the basic requirements to be implemented on the computer, that is, they are defined and effective. According to the famous equation of N. Wirth:

$$\text{algorithms} + \text{data structures} = \text{programs}$$

A program consists of a set of procedures, which are considered black boxes, and an associated set of data on which the procedures operate. The form that the algorithm takes in a computer implementation is subordinate to the programming style and depends in particular on the type of language. [Figure 2](#) shows the logical scheme of the fuzzy algorithm for a control system that accepts the structured programming standard. At the output, the program displays numerical values of the process at the current time or optionally graphical representations.

The software implementation of the fuzzy machine can also be based on a parallel algorithm. Parallel computing gives a new dimension to the construction of algorithms and programs. It is emphasized that parallel programming is not a simple extension of serial programming and that not all sequential algorithms can be parallelized.

Design of fuzzy systems with Matlab

In the Matlab software environment, there are specialized toolboxes for the analysis and design of intelligent control systems, which include the Fuzzy Logic Toolbox (FLT). This toolbox consists of a set of Matlab files (in two subdirectories: FUZZY and FUZZYDEMOS) that allow you to tackle the steps characteristic of synthesizing a fuzzy inference (FIS) based system. The subdirectory FUZZY contains function files divided into the following categories of specific functions and operations:

- Graphical user interface (GUI) functions; functions for editing the fuzzy inference system (FIS), membership functions (FA) and rules used, inference diagrams and associated control

surfaces; functions for the FIS generation command (by fusing - defining a FA for each variable involved in fuzzy rules, by mitigation based on the estimation of fuzzy inferences, and by transferring parameters between functions and variables, respectively. by generic evaluation of FA and visualization of the control surface); functions for implementation of other routines (FIS of Sugeno type, clusters of C-means type, etc.).

- Operations related to the difference of two FAs with different shapes (sigmoidal, Gaussian, pi, trapezoidal, triangular, Z, S, including their combinations), to concatenate matrices, to discretize the FIS, to evaluate multiple FAs, to edit lines of text, including active auxiliaries.

The FUZZYDEMOS Deputy Director (FUZZY LOGIC TOOLBOX DEMOS) contains several demonstration applications for basic fuzzy functions and operations, as well as complete fuzzy models of intelligent control systems.

Examples of functions and operations:

- Control functions for the graphical user interface with FUZZY for FIS editing
- MFEDIT for editing membership functions
- RULEEDIT for editing SURFVIEW rules viewing control surfaces
- RULEVIEW visualization of rules (RULEVIEW (FIS) and fuzzy inference diagrams for a FIS matrix - RULEVIEW ('FILENAME')).

In the construction and simulation of the Fuzzy model for evaluating university teachers, the following steps were:

Step 1: Set up the following diagrams in FIS Editor:

- ✓ block diagram of the fuzzy system for the evaluation of the didactic activity (see [Figure 7](#)), which has as variables the following:
 - Teaching Method;
 - Additional Resources;
 - Student Interactions;
 - Teaching Skills;
 - Implication.

The input variable “Explanations” and its sets are shown in [Figure 8](#).

Figure 7. Printscreen with block diagram of the fuzzy system for the evaluation of the didactic activity.

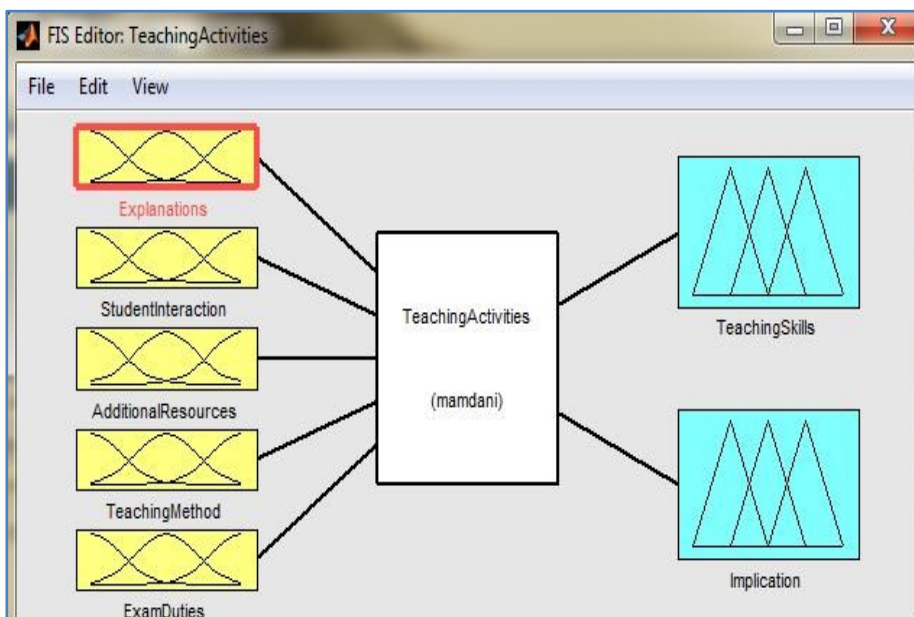
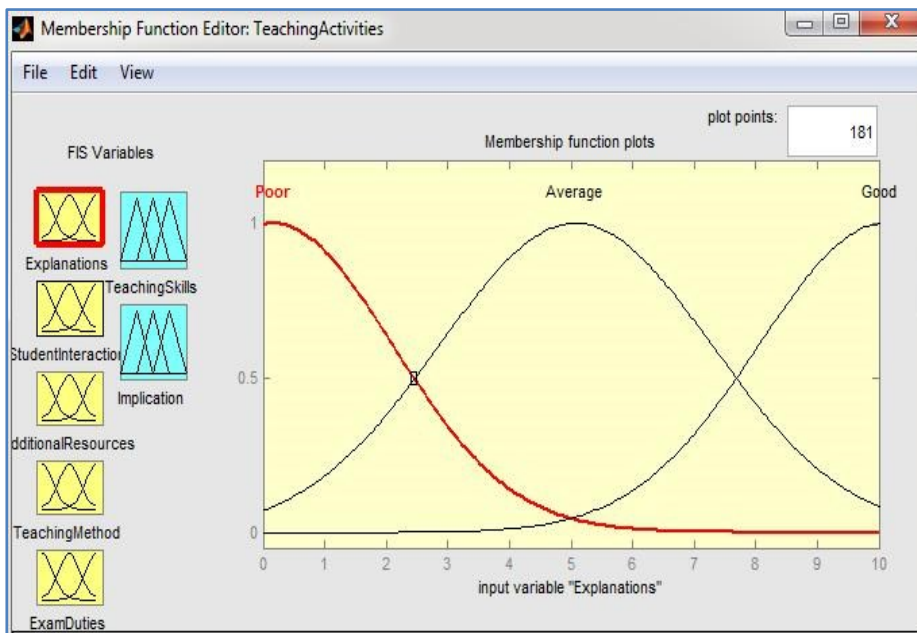
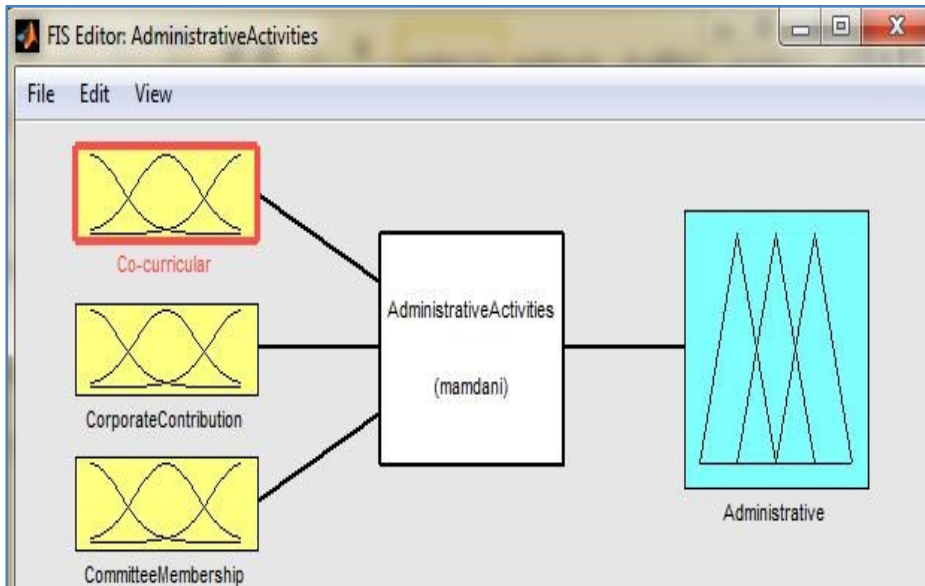


Figure 8. Printscreen with the input variable “Explanations” and its sets.



- ✓ The block diagram of the fuzzy system for the evaluation of the administrative activity (see Figure 9), which has as variables the following:
- Committee Membership;
 - Corporate Contribution;
 - Co-curricular.

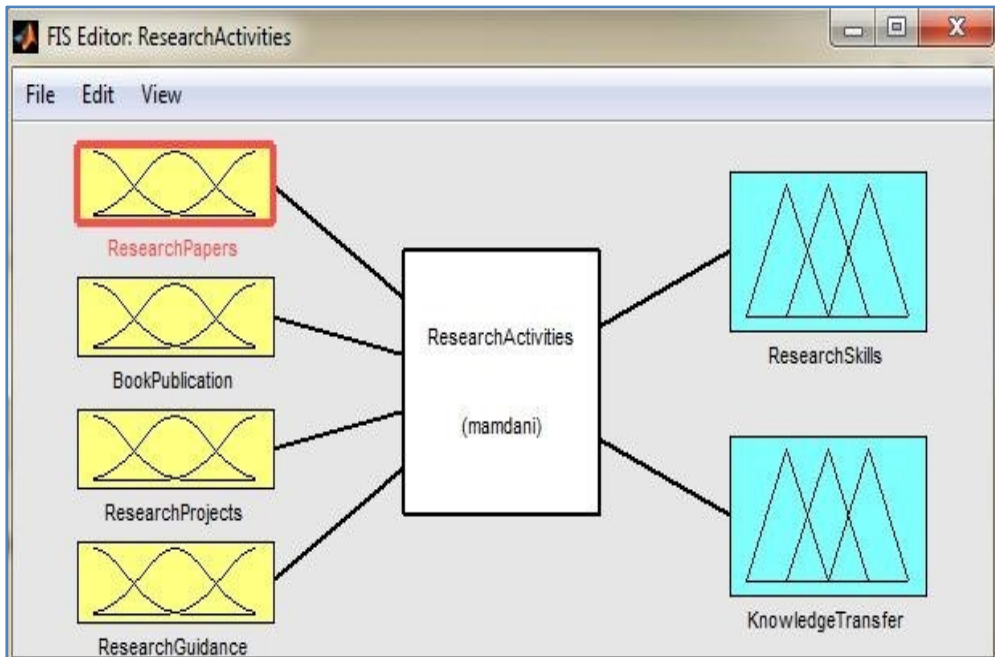
Figure 9. Block diagram of the fuzzy system for evaluating administrative activity.



- ✓ Block diagram of the fuzzy system for evaluating the research activity (Figure 10), which has the following variables:
- Research Projects;
 - Research Guidance;
 - Book Publications;
 - Reserch Paper.

All these variables underlie the evaluation related to ResearchSkills and Knowledge Transfer of each university professor (see Figure 10).

Figure 10. Block diagram of the fuzzy system for evaluating the research activity.



Within the main graphical interface screen (see Figure 11), there is a separation by color code, from top to bottom as follows: blue - data entry, green - results achieved, red - assessment of academics with minimum requirements. The minimum values for each university teaching position used in this simulation are presented within Table 1.

Figure 11. Graphical interface of the evaluation tool.

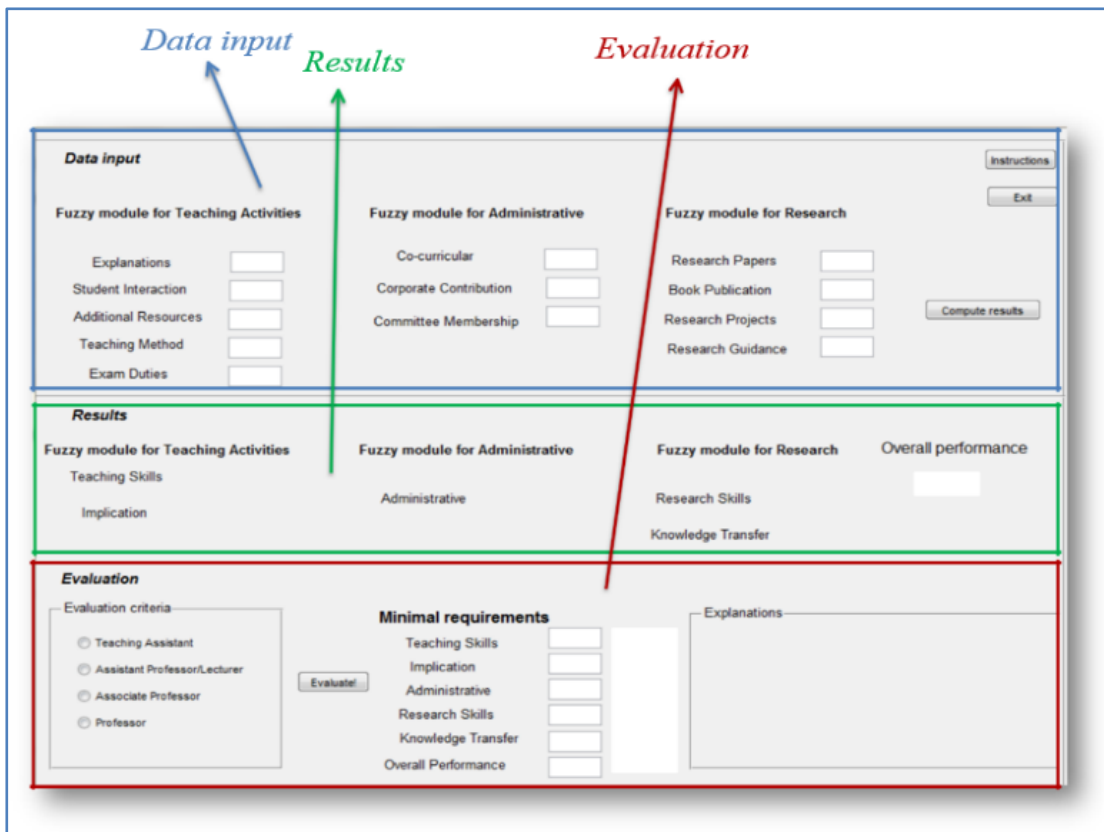


Table 1. Minimum values for each university teaching position.

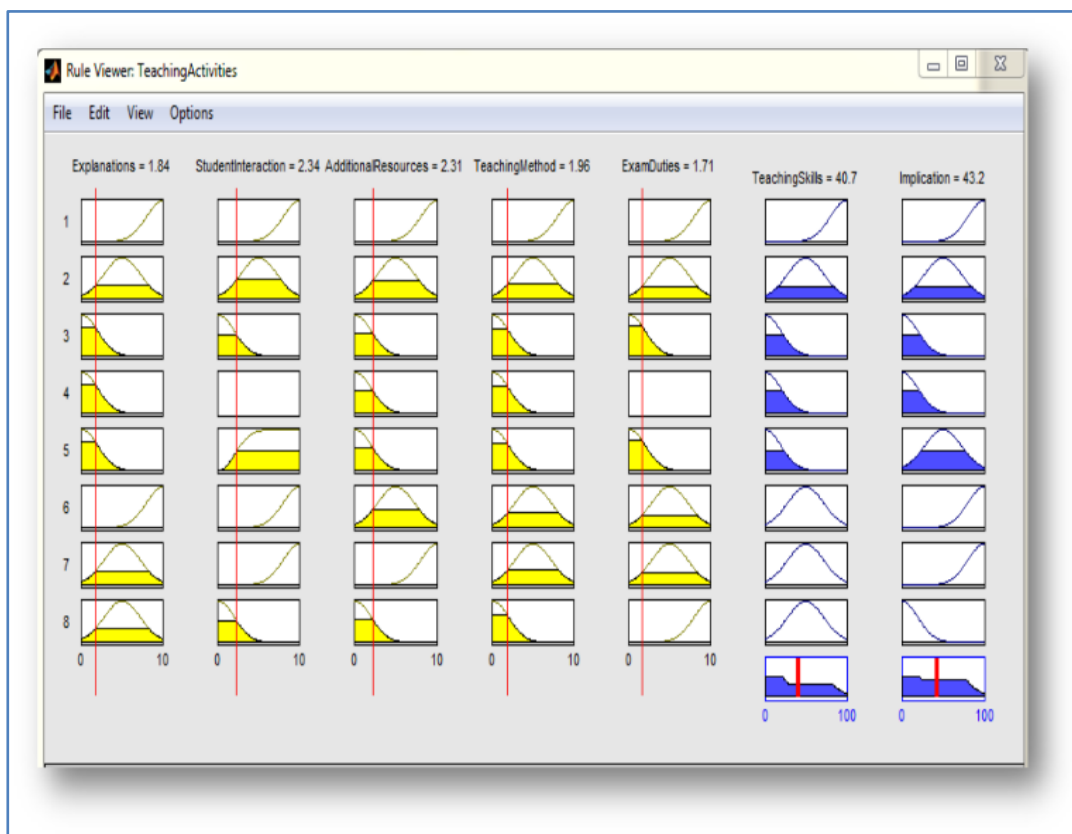
	Teaching assistant	Assistant professor	Associate professor	Professor
Teaching skills	30	40	60	70
Implication	30	40	60	60
Administrative	0	25	60	70
Research skills	10	30	60	70
Knowledge transfer	10	25	60	70
Overall performance	80	160	300	340

Step 2: Simulation of the minimum values for each teaching position university:

In this way, a minimum accepted score was established for each category of university employee: University Professor, Associate University Professor, University Lecturer and University Assistant.

The rule of fuzzy evaluation system for didactic activities (see Figure 12): if (Explanations is average) and (StudentInteraction is good) and (AdditionalResources is high) and (TeachingMethod is average) and (ExamDuties is average) then (TeachingSkills is average) (Implication is good).

Figure 12. The rule of the fuzzy activity evaluation system.



3. RESULT / FINDINGS

There are 2 situations with the afferent results of the simulation and validation phase: one in which the minimum conditions are met and the other in which they are not. The print screen in Figure 13 shows us a simulation with the results of the ideal model in which a teaching assistant meets the minimum requirements. Also in this pilot study, a situation was simulated in which a lecturer /assistant Professor does not meet the minimum requirements (see Figure 14).

Figure 13. Situation when are met the minimum requirements.

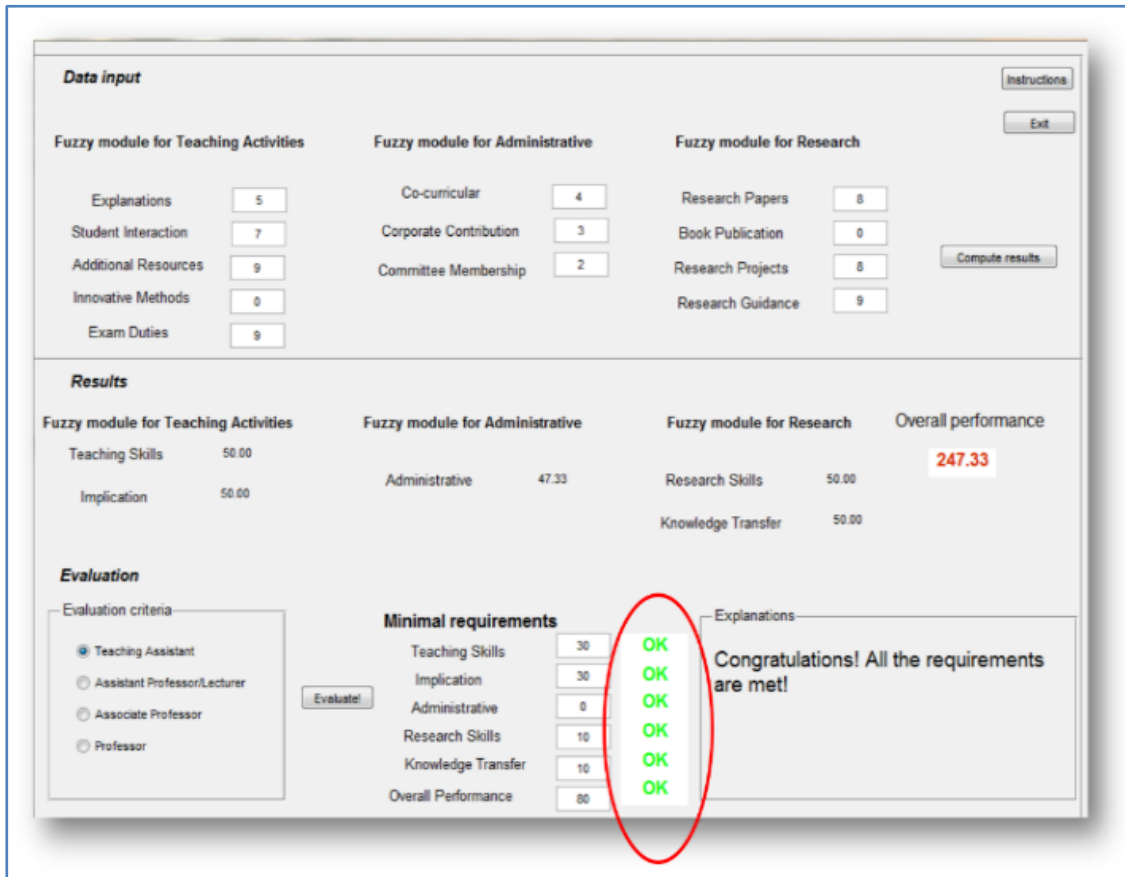
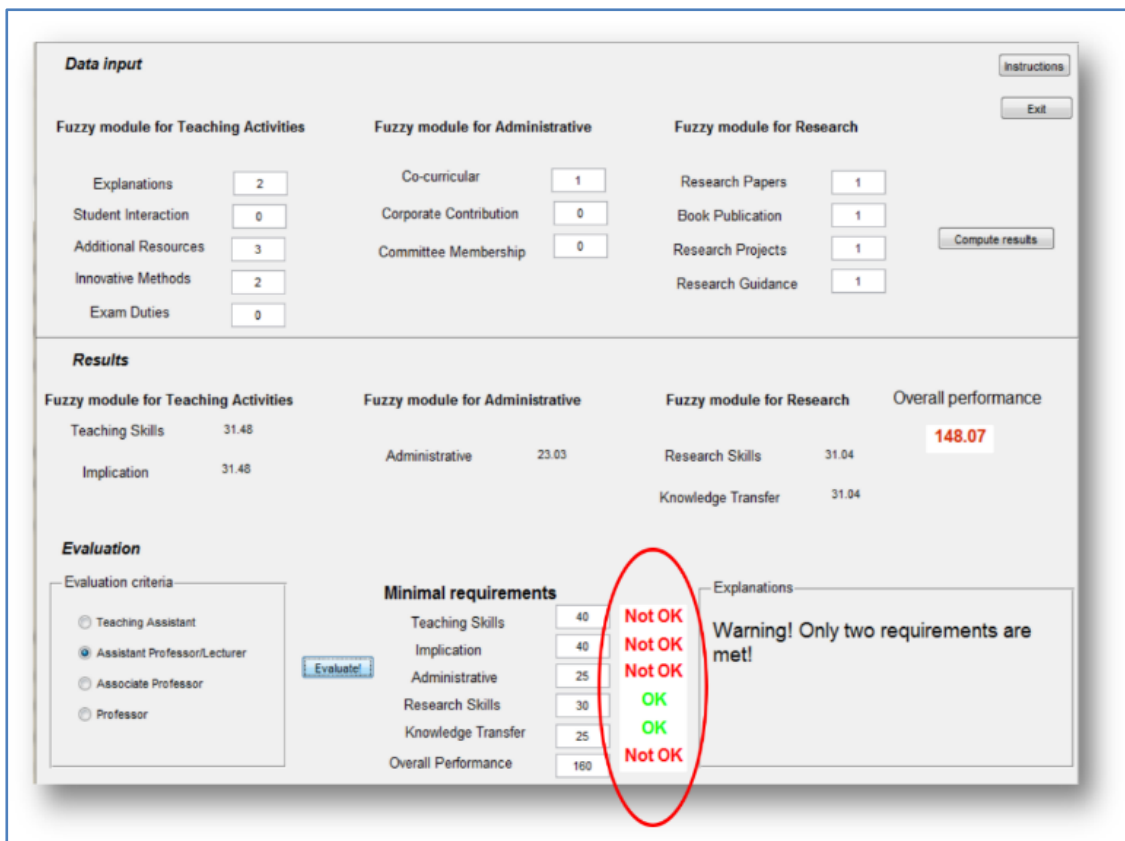


Figure 14. Situation when are NOT met the minimum requirements.



4. DISCUSSION and CONCLUSION

Fuzzy set theory is the most general theory of incompleteness yet formulated. Fuzzy logic offers the possibility of reasoning through general knowledge formulated in a general way and has therefore found its application in many fields. Fuzzy concepts and rules can be represented and manipulated by computer, a very valuable feature in the field of knowledge base engineering, where the knowledge of experts is usually formulated in ordinary language.

Applied fuzzy logic is indeed a computational technique that can be used to obtain more meaningful solutions than the classical exact methods when solving specific problems. At the same time, fuzzy systems work very well in the presence of uncertainty, imprecision and "noise". How well such fuzzy systems work is shown by their widespread application in recent years all over the world. There are already a number of well-known applications of fuzzy logic in various fields of science: in automatic control (temperature rules, speed control of the subway, autofocus of video cameras), in shape recognition (fuzzy classification algorithms), in measurements (information processing - sensors), in medicine (control of pacemakers), in economics (fuzzy decision methods).

The model of fuzzy logic applied to the evaluation of university professors is an approximation method that allows to formally model vague "knowledge" stored in a base of rules. The application of fuzzy logic in the evaluation of university personnel at different levels is due to the advantages it offers in the following specific situations:

- Enables modelling of non-linear, complex, or imprecisely known processes for evaluation of university personnel, according to level;
- allows the implementation of the human experience of the evaluators, in this case in the construction of the inference rules, using the linguistic variables explained in the theoretical part..

Fuzzy set theory is used in the evaluation of academics for the following purposes:

- Modeling: in this sense, fuzzy set theory is one of the methods that can be used to model different types of uncertainties related to the teacher's competencies in different circumstances;
- Generalization: classical models and methods are usually based on two-valued logic. This approach often does not adequately represent reality. Fuzzy set theory has been used mainly to relax the classical methods by introducing the gradual character;
- Simplification: fuzzy technology is used to reduce the complexity of data to an acceptable level through linguistic variables or fuzzy analysis

The present study brought as a novelty the application of fuzzy logic in the field of education, namely in the evaluation of university teachers. The evaluation of university teachers using fuzzy logic is a first stage of quantitative evaluation, which can be the basis for the final evaluation by the heads of departments. Why use fuzzy logic in faculty evaluation? Because it is a helpful complementary tool and has clear advantages such as: it is easy to understand and apply, it is flexible, it is tolerant of imprecise data, it can model complex functions with a high degree of accuracy, it can use expert knowledge, it can be combined with conventional control techniques.

The model of fuzzy logic applied to the evaluation of university professors is an approximation method that can formally model vague "knowledge" stored in a base of rules. The application of fuzzy logic in the evaluation of university personnel at different levels is due to the advantages it offers in the following specific situations:

- Enables modelling of non-linear, complex, or imprecisely known processes for evaluation of university personnel, according to level;

- allows the implementation of the human experience of the evaluators, in this case in the construction of the inference rules, using the linguistic variables explained in the theoretical part.

The authors consider that the results obtained with the methods proposed in this study can be used in other directions. For example, one of these directions can be the selection of a field of study of universities. In this case, the attitude of students closest to the maximum satisfaction level should be studied.

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

Authors are expected to present author contributions statement to their manuscript such as; **Vasile Florin Popescu**: Writing original draft, Methodology, Supervision, and Validation. **Marius Sorin Pistol**: Resources, Visualization, Software.

ORCID

Vasile Florin Popescu  <http://orcid.org/0000-0002-9972-9904>

Marius Sorin Pistol  <https://orcid.org/0000-0003-1172-3637>

5. REFERENCES

- Ahmed, F., & Toki, M. (2016). A Review on Washing Machine Using Fuzzy Logic Controller. *International Journal of Emerging Trends in Engineering*, 4(7), 64-67. <http://www.warse.org/IJETER/static/pdf/file/ijeter02472016.pdf>
- Bellman, R., & Zadeh, L. (1970). Decision Making in a Fuzzy Environment. *Management Sciences*, 17(4), 141-164. <http://dx.doi.org/10.1287/mnsc.17.4.B141>
- Chennakesava, R. (2008). *Fuzzy logic and neural networks*. Basic concepts & applications, New Age International Publishers, Darya Ganj, New Delhi-110 002, India.
- Chuen, L. (1990). Fuzzy Logic in Control Systems: Fuzzy Logic Controller – Part I. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2), 404-418. <http://ieeexplore.ieee.org/document/52551>
- Garrido, L. (2012). A Brief History of Fuzzy Logic. *Broad research in artificial intelligence and neuroscience*, 3(1), 71-77.
- Łukasiewicz, J., & Tarski, A. (1930). *Untersuchungen über den Aussagenkalkül (German)*. *Comptes rendus des séances de la Société des Sciences et des Lettres de Varsovie. CI III*, 23, 30–50. English translation: Investigations into the sentential calculus.
- Mamdani, E., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1), 1-13. [https://doi.org/10.1016/S0020-7373\(75\)80002-2](https://doi.org/10.1016/S0020-7373(75)80002-2)
- McCarthy, J. (1959). Programs with Common Sense at the Wayback Machine (archived October 4, 2013). In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, 756-91. Her Majesty's Stationery Office.
- Patjoshi R., & Mohapatra, K. (2010). Experimental Investigation on Microcontroller based Elevator Positioning Control System Using Fuzzy-Logic. *International Journal of Advanced Technology and Engineering Exploration*, 8(5), 88-94.
- Takagi, T., & Sugeno, M. (1985). Fuzzy Identification of Systems and Its Application to Modeling and Control. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15, 116-132. <http://dx.doi.org/10.1109/TSMC.1985.6313399>

- Subbulakshmi, L. (2014). Antilock-braking system using fuzzy logic. *Middle-East Journal of Scientific Research*, 20(10), 1306-1310, 2014, <http://dx.doi.org/10.5829/idosi.mejsr.2014.20.10.232>
- Vijayana, K., Srivastavaa, P.P., Raghunathb, M.K., & Saratchandraa, B. (2011) Enhancement of stress tolerance in mulberry. *Scientia Horticulturae*, 129(4), 511-519. <http://dx.doi.org/10.1016/j.scienta.2011.04.018>
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353. [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)
- Zadeh, L. (1996). Fuzzy logic=computing with words. *IEEE Transactions on Fuzzy Systems*, 4(2), 103–111. <https://doi.org/10.1109/91.493904>
- Zadeh, L. (1968). Fuzzy algorithms. *Information and Control*, 12(2), 94-102. [https://doi.org/10.1016/S0019-9958\(68\)90211-8](https://doi.org/10.1016/S0019-9958(68)90211-8)
- Zadeh, L. (1971). Quantitative fuzzy semantics. *Information Sciences*, 3(2), 159–176. [https://doi.org/10.1016/S0020-0255\(71\)80004-X](https://doi.org/10.1016/S0020-0255(71)80004-X)
- Zadeh, L. (1973). Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, 28-44. <https://doi.org/10.1109/TSMC.1973.5408575>