IJATE

Volume 7 Issue 4 2020

International Journal of Assessment Tools in Education

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

e-ISSN: 2148-7456

IJATE

e-ISSN 2148-7456

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

Volume 7 Issue 4 2020

Dr. Izzet 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 starting from June 2018 (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).

There is no submission or publication process charges for articles in IJATE.

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

IJATE is indexed in:

- Emerging Sources Citation Index (ESCI),
- Education Resources Information Center (ERIC),
- TR Index (ULAKBIM),
- European Reference Index for the Humanities and Social Sciences (ERIH PLUS),
- EBSCO
- Directory of Open Access Journals (DOAJ),
- Index Copernicus International
- SOBIAD,
- JournalTOCs,
- MIAR 2015 (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,
- ResearchBib,

ISSN: 2148-7456

English Language Editors

Copy & Language Editor

TABLE OF CONTENTS

Research Articles

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 522–534

https://doi.org/10.21449/ijate.793084

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Construction and Standardization of Examination Anxiety Scale for Adolescent Students

Nargis Abbasi[1,*], **Shilpi Ghosh**[2]

[1]Department of Education, Magrahat College, South 24 Parganas. Pin- 743355, West Bengal, India
[2]Department of Education, Vidya Bhavana, Visva-Bharati, Santiniketan, Pin-731235, West Bengal, India

ARTICLE HISTORY

Received: Apr. 07, 2020
Revised: Aug. 20, 2020
Accepted: Sep. 05, 2020

KEYWORDS

Examination anxiety,
Adolescent,
Scale construction,
Scale standardization

Abstract: This research paper describes the method of construction and standardization of a tool to measure examination anxiety of adolescent students. 2030 students belonging to the age group of 13-15 years from 19 schools under West Bengal Board of Secondary Education, participated in this research. The first draft of examination anxiety scale consisted 40 items. After reviewing the items and item analysis, the number of items were reduced to 38. EFA was carried out on obtained data. EFA revealed that total 21 items having factor loading greater than .40, are selected. They distributed under four factors such as, Bodily symptoms, Cognitive, Emotional reaction and behavioural reaction. CFA was executed on another sample group, consisted of 402 number of adolescent students of age group13-15. CFA results also supported the results of EFA. All the goodness of fit indices showed that the model is a good fit model. For concurrent validity, Examination Anxiety Scale made by researcher and Test Anxiety Inventory by Spielberger were administered on the same occasion on 110 school students of the age group13-15. Coefficient of correlation of two scales was estimated. The validity of Examination anxiety Scale is 0.71. The reliability coefficient of the examination anxiety scale using test-retest, split half and Cronbach's alpha methods were 0.801, 0.767, 0.764 respectively. Norms show that 16 percent of the students belong to the high examination anxiety group, 66 percent of the students in average examination anxiety group and 18 percent in low examination anxiety group.

1. INTRODUCTION

Zeidner (1992) corroborates, "contemporary society is best described as test-oriented and test consuming". In this context, famous psychologist Sarason (1959) implies, "We live in a test conscious, test-giving culture in which the lives of people are in part determined by their test performance". In present scenario test or examination is most prominent cause of anxiety among students. Generally, students feel the utmost fear of examination by anticipating their poor performances and failure, this causes examination anxiety. In fact, examination anxiety is an unpredictable worry about the consequence regarding performance, fear of being assessed, and the apprehension about the results. It also includes irrational thoughts, unnecessary demands and expectation, and catastrophic predictions. Examination anxiety is supposed to be "a major factor contributing to a variety of negative outcomes, including psychological distress,

CONTACT: Nargis Abbasi ✉ nargis.physics@gmail.com Department of Education, Magrahat College, Magrahat College, S. 24 Parganas, Pin-743355, West Bengal, India

ISSN-e: 2148-7456 /© IJATE 2020

academic underachievement, academic failure, and insecurity" (Hembree, 1988). A student with an optimum level of anxiety performs well in the examination but excessive level of examination anxiety deteriorates the performances in examination (Abbasi & Ghosh, 2020). According to Zeidner (1998) "many students have the ability to do well on exams, but perform poorly because of their debilitating levels of anxiety. Consequently, test anxiety may limit educational or vocational development, as test scores and grades influence entrance to many educational or vocational training programs in modern society". Wine (1971) implied that both self-relevant and task-relevant variables are attended by those people who have high test anxiety at the time of examination; on the other hand, those people who have low test-anxiety generally attend to task-relevant variables. In fact, those people have high test anxiety envisage in the time of examination.

Examination anxiety causes a couple of problems. However, each student has different symptoms, having different levels of intensity. Shukla (2013) categorized these symptoms under four dimensions.

i) Physical - nausea or diarrhea, extreme body temperature changes, dry mouth, headache, sweating, rapid heartbeat, shortness of breath, light-headedness,

ii) Emotional –feeling of helplessness, anger, excessive feelings of fear, uncontrollable crying, disappointment, depression,

iii) Behavioural – substance abuse, fidgeting, avoidance, pacing,

iv) Cognitive – negative thinking, the difficulty of organizing thought, negative self-talk, racing thoughts, comparing yourself to others, "going blank", difficulty concentrating, and feelings of dread.

According to McDonald (2010), 10 - 40 percent of all students are severely affected by examination anxiety. The percentage also increases in the case of the formal examination. The examination anxiety is a very serious problem of modern times; its consequences are found in several forms like trauma, psychological disorder, and suicide, as reported in the newspaper during the period of examination.

According to National Crime Bureau (NCB, 2015), there is a shocking report in the context of India that 2646 students, more than 7 per day in each year, are found to commit suicide due to failure in examinations. In 2014 the number was marginally lower- 2403 (NCB, 2014).

In present times, test anxiety measures are constructed to reveal a "bio-psychosocial model" of test anxiety, which hypothesizes the notion that test anxiety is revealed through 'behavioural, cognitive and physiological symptoms' (Lowe et al., 2008 & Embse et al., 2013).

Whereas Sarason et al. (1960), being earlier examination anxiety researchers interpreted the concept of test anxiety on the basis of one dimension, but later, Libert and Moris (1967) divided test anxiety into two components- one is "worry," and another is "emotionality". However, in the 1980s, more significantly the detailed definition of the dimension "worry" was proposed, such as "irrelevant thinking" and "worry" (Sarason, 1984), "worry" and "fear of failure" (Covington, 1985), and "distraction" and "low self-confidence" (Hodapp & Benson, 1997). Thus, it can be inferred from the past studies that the concept of examination anxiety has evolved into a multi-layered notion with several dimensions of responses, as reported by Zeidner (1998), that includes behavioral, physiological, and emotional and thinking components. Based on these deductions, this study focuses on the four major components of the examination anxiety scale. These are worry, emotional reaction, bodily symptoms, and behavioral reaction.

The first measurement instrument for examination anxiety was devised by Mandler and Sarason in 1952. This test anxiety questionnaire having 42 items aims at measuring the experience before and during intelligence test and course examination. Six years later, another Test Anxiety

Scale consisting of 21 items, was developed by Sarason (1958). Finally, in 1972 Sarason developed test anxiety scale with 37 items. Suinn (1969) developed Test Anxiety Behaviour Scale, another global measurement scale having 50 items, efficiently assesses the behavioral condition at the time of examination anxiety. Test Anxiety Inventory (TAI) was developed by Spielberger and its Associates (1980). The TAI, a standardized measurement scale for test anxiety, consists of 20 items that separate worry and emotionality and, at the same time, yields total score of examination anxiety. Actually, many more examination anxiety measures are found in the global perspective, but examination anxiety measures on the basis of Indian perspectives are rarely found. So this research aims at developing a scale that measures the examination anxiety of adolescents.

The main purpose of the study is to construct a standardized examination anxiety scale for adolescent students. The study also aims at computing reliability and validity of the scale.

2. METHOD

2.1. Population

The population of the study includes the adolescent students of the age group of 13-15 years, studying in schools under the West Bengal Board of Secondary Education.

2.2. Sample

2030 students of whom age group ranges 13-15 years of 19 schools of West Bengal Board of Secondary Education of West Bengal were selected as sample for the first participant group. The sample of the study was selected from Jalpaiguri, Coochbihar, Darjeeling, S. Dinajpur, South 24 Parganas, Burdwan, and Kolkata district of West Bengal. Boys and girls of class IX and X, from rural and urban areas of West Bengal, were selected.

For Confirmatory factorial analysis, like Akkus (2019) also did, a different sample was taken. In this present study the researchers also selected 402 numbers of students of whom age group ranges 13-15 years from 7 schools of West Bengal Board of Secondary Education as sample for Confirmatory factorial analysis.

2.3. Construction of Scale

The first step of the construction of the Examination Anxiety Scale (EAS) is to construct the items. The researchers constructed both positive and negative items reflecting examination anxiety of the students.

The researcher studied related literature on examination anxiety in order to collect and construct items. Teachers and students provided much information about the examination anxiety of the students. With the help of this information the researchers constructed items reflecting the examination anxiety of the students.

The researchers constructed the first draft of the Examination Anxiety Scale. The first draft of the Examination Anxiety Scale (EAS) consisted of 40 items. 28 items were positive items and 12 of the items were negative items.

The primary Examination Anxiety scale (EAS) is divided into four subscales or sub-points;

1. Bodily Symptoms Subscale
2. Cognitive (Worry Subscale)
3. Emotional Reaction Subscale
4. Behavioural Reaction Subscale

2.4. Evaluation by Experts and Reconstruction of the Tool

The items were prepared by the researchers and evaluated by the experts of the subject and language. The researchers followed their suggestions and made necessary modifications in the

Examination Anxiety Scale. This modified Examination Anxiety Scale (EAS) consisted of 40 items.

2.5. Scoring Key

The scale consists of a 5-point Likert scale. The five options given are strongly agree, agree, undecided, disagree, and strongly disagree. Table 1 and Table 2 show the scoring code of positive items and negative items respectively.

Table 1. *Scoring for positive items*

Types of rating	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
Score	5	4	3	2	2

Table 2. *Scoring for negative items*

Types of rating	Strongly agree	Agree	Undecided	Disagree	Strongly disagree
Scores	1	2	3	4	5

As each response weighted from 1 to 5, the minimum Examination Anxiety Scale (EAS) total score is 40, and the maximum total score is 200. The Examination Anxiety Scale (EAS) has four subscales which measure the four major components of examination anxiety. The subscales are Bodily Symptoms (EAS/Bo) Worry (EAS/W), Emotionality (EAS/E) and Behavioural Reaction (EAS/B). Worry refers to the cognitive side of anxiety (Sarason, 1984).

The items in EAS/Bo subscale are: 4, 6, 11, 15, 16, 18, 27, 32, 33, 35

The items in EAS/W subscale are: 1, 5, 12, 13, 14, 17, 22, 23, 24, 28, 37, 38, 39, 40

The items in EAS/E subscale are: 2, 3, 7, 8, 9, 19, 21, 29, 30, 34, 36

The items in EAS/ Be subscale are: 10, 20, 25, 26, 31.

2.6. Administration of the Scale

The draft Examination anxiety scale (EAS) consisting of 40 items was administered on a group of 2030 students belonging to the age group 13-15 years for item analysis and exploratory factor analysis. The sample was drawn from the 19 schools under West Bengal Board of Secondary Education. The schools included students from both rural and urban areas of the state. The researchers explained the purpose of the administration of the examination anxiety scale to the students. A clear instruction was given by the researchers regarding how to respond the items of the test. Then, the final draft of EAS was administered on the second participant group, which comprised of 402 number of school students for Confirmatory factor analysis.

3. RESULTS

3.1. Item Analysis

On the basis of the total score of each respondent, the researcher selected upper 27% cases of the whole group as a high score group and lower 27% cases of the whole group as a low score group. After that *t*- test value was calculated between two groups. *t* values of each item are shown in Table 3. The items which have *t* value of less than 1.96 have been rejected. According to Table 3, all the items except, 30 & 39 are significant. The researcher decided to select all the significant items which are significant and higher than 1.96 *t* value for the second draft of the scale.

Table 3. *t value of the items of Examination Anxiety Scale (Item discrimination index)*

Item no	*t* value	Item No.	*t* value	Item no.	Value
1	23.067	2	20.151	3	-22.377
4	24.005	5	-10.302	6	14.069
7	22.157	8	26.972	9	29.992
10	18.091	11	29.382	12	-22.788
13	22.482	14	-24.753	15	19.585
16	21.586	17	16.125	18	-24.036
19	-6.917	20	18.293	21	19.702
22	15.097	23	16.511	24	-10.141
25	17.711	26	20.451	27	20.403
28	17.829	29	-17.746	30	**-1.276**
31	19.116	32	16.843	33	25.896
34	-6.407	35	18.297	36	17.018
37	22.507	38	19.480	39	**-1.742**
40	-5.153				

3.2. Validity

Freeman (1960) interpreted validity as; "An index of validity shows the degree to which a test measures what it purposes to measure when compared with accepted criteria.".The validity of the Examination Anxiety Scale was determined by the following method.

3.2.1. *Construct Validity*

3.2.1.1. *Exploratory factor analysis*

The principal component analysis was carried out on the data for factor analysis. Varimax orthogonal technique was used for rotation. After analysis, 17 items are eliminated as they were distributed under multiple factors, and their factor loading less was than 0.4. Only 21 items were selected for the final draft. The 21 items were distributed under 4 factors.

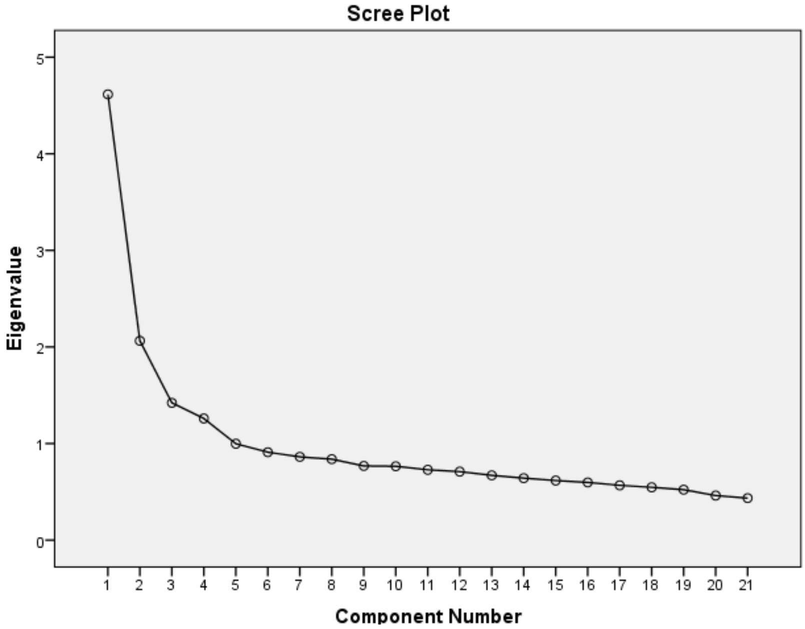


Figure 1. *Scree plot showing four factors.*

According to Figure 1, there was a sharp drop in the first four factors. They had a noteworthy contribution to variance explanation. According to Nancy et al. (2005), Kaiser-Meyer-Olkin (KMO) measures should be higher than 0.70. In our study, the KMO value was 0.851, which is greater than 0.70. It indicates that enough items are predicted by each factor. Bartlett's test of sphericity was computed as 8108.15, and p-value is 0.00, which is less than 0.05 at the 95% significant level. It indicates that the research sample was significantly suitable for the analysis of the study. The Scree plot was shown in Figure 1.

Table 4. *Factor loading after varimax rotation and extracted communalities and eigenvalues*

Items	Communality	Factor 1 (Bodily Symptoms)	Factor 2 (Cognitive)	Factor 3 (Emotional)	Factor 4 (Behavioral)
32	0.600	0.762			
4	0.441	0.649			
6	0.413	0.638			
16	0.377	0.583			
18	0.373	0.569			
11	0.422	0.562			
12	0.391		0.661		
14	0.436		-0.624		
24	0.463		-0.620		
13	0.482		0.613		
37	0.366		0.496		
8	0.470			0.652	
9	0.416			-0.635	
29	0.427			-0.596	
19	0.511			-0.525	
36	0.403			0.434	
20	0.588				.667
25	0.489				.597
10	0.461				.576
31	0.435				.458
26	0.389				0.410
Eigenvalue		4.615	2.064	1.422	1.259
Explained Variance		13.289	12.589	9.443	9.257
Total variance			44.570		

Table 4 presents the results of factor analysis. Factor analysis reported four strong factors with an eigen value greater than 1.00. The four factors were i) Bodily symptoms, ii) Cognitive (worry subscale), iii) Emotional reaction, and iv) Behavioural reaction

All item loading exceeded .40. 21 items were selected in final form of the scale out of which, six items are reversed items. All four factors together explain 44.165 % of total variance. The 1st 2nd, 3rd, and 4th factors explain 13.289, 12.589, 9.443, and 9.257% of total variances, respectively. When four factors were extracted, the highest communality is 0.607 for item 32, and the lowest communality is 0.365 for item 37.

3.2.1.2. *Confirmatory factor analysis*

Confirmatory factor analysis was done in order to determine the construct validity of EAS to verify that the items fit with four-factor model. For confirmatory factor analysis, data were collected from a separate sample. The sample comprised of 402 student studying class IX and X of West Bengal.

Confirmatory factor analysis was run through Amos 24.0 software. According to Browne & Cudeck (1932) a model is considered as a good fit if the $\chi^2 /df \leq 2$ (as cited in Akkus, 2019). Confirmatory factor analysis results showed that χ^2/df ratio is 1.823. It indicates that the exact fit hypothesis was accepted. The root means square error of approximation (RMSEA) value is 0.045. Goodness of fit index (GFI) value is 0.927, Adjusted goodness of fit index (AGFI) value is 0.908, Normal fit index (NFI) value is 0.709, and Comparative fit index (CFI) value is 0.890. All fit indices show that the model is a good fit model.

If RMSEA value is ≤ 0.05, then it means that the model is a good fit. And the RMSEA value ≤ 0.08 indicates the model fits well with reasonable error (MacCallum, Browne, & Sagawara, 1996). Most well fit model possesses GFI, AGFI, CFI value ≥ 0.9 for a strong model (Finch, Immekus, & French, 2016). Finally, it can be concluded that the EAS model is a good fit model. Confirmatory factorial analysis with standardized result is shown in Figure 2.

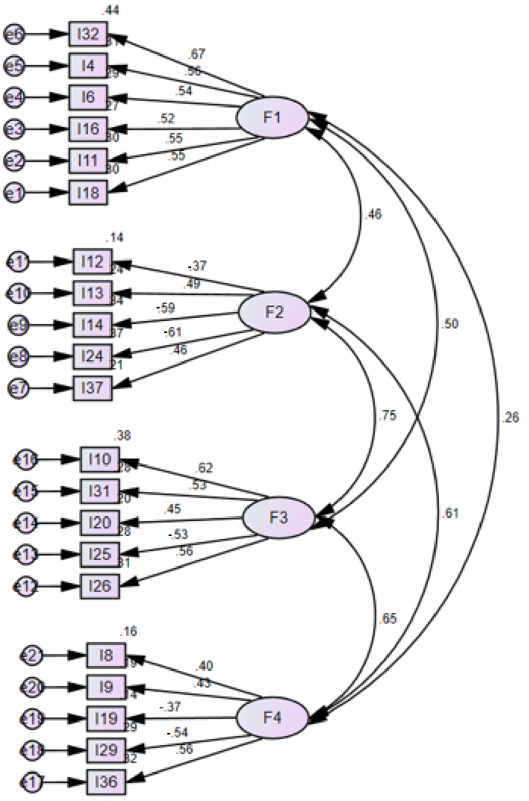


Figure 2. *Confirmatory Factor Analysis with Standardized Results*.

4.2.2. *Concurrent validity*

In the present study for concurrent validity, the Examination Anxiety Scale (EAS) made by the researcher and Test Anxiety Inventory (TAI) by Spielberger (1980) were administered on the same occasion on 110 school students of class 9 and 10 of Harirampur Betna High School of South Dinajpur district. Data were collected, and the coefficient of correlation of two scales was estimated. According to Table 5, the concurrent validity of Examination Anxiety Scale is 0.71. So the validity of the scale is good. Hence, the scale is valid.

Table 5. *Correlation coefficient between Examination Anxiety Scale (EAS) and TAI by Spielberger*

	Examination Anxiety Scale	Spielberger TAI
Examination Anxiety Scale Correlation	1	0.71
Spielberger TAI Correlation	0.71	1

4.3. Reliability

Anastasi and Ubrina (2005) have opined in context to reliability that "Reliability refers to the consistency of scores obtained by the same persons when they are re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions."

The researcher first administered the Examination Anxiety Scale (EAS) to secondary school students. After one month, the researcher administered the Examination Anxiety Scale to the same students.

4.3.1. *Correlation by test-retest*

It is clear from Table 6 that the correlation between test scores and retest scores is 0.801. It can be said that the reliability of the Scale is high. Hence, the Examination Anxiety Scale is reliable.

Table 6. *Correlation co-efficient by the test-retest method*

		Test	Retest
Test	Pearson correlation	1	0.801
	N	2030	2030
Retest	Pearson correlation	0.801	1
	N	2030	2030

Correlation is significant at the 0.01 level (2-tailed).

4.3.2. *Cronbach's alpha and split – half coefficient*

From Table 7, the reliability of the Examination Anxiety Scale (EAS) by split-half method is 0.767, and the reliability of the scale by Cronbach's Alfa method is 0.764. Hence, we can say the reliability of the examination anxiety is high.

Table 7. *Correlation coefficient by Cronbach's Alpha and Split-Half method*

Method	Reliability value
Split half	0.767
Cronbach's Alfa	0.764

4.4. Details of the Final Draft

Table 8 shows that 19 items are rejected, and 21 items are retained for the final draft of examination anxiety scale. Table 9 shows that 15 items out of 21 items were positive, and 06 items were negative. According to Table 9, the total number of positive item is 15 and total number of negative item is 6.

Table 8. *Distribution of selected or rejected items for the final draft of the examination anxiety scale*

S. No	Item number	*f*	Remarks
1	4,6,8,9,10,11,12,13,14,16,17,18,19,20,24,25,26,29,32,36,37	21	Selected
2	1,2,3,5,7,15,21,22,23,27,28,30,31,33,34,35,38,39,40	19	Rejected

Table 9. *Distribution of positive and negative items for the final draft*

Statement	Item number	Total
Positive	4,6,8,9,10,11,13,16,17,18, 20, 26,32,36,37	15
Negative	12,14,19,24,25,29	6
Total		21

Table 10 shows that 6 items were selected for bodily symptoms subscale, 5 items were selected for the cognitive subscale; 5 items were selected for Emotional reaction subscale. 5 items were chosen for the Behavioural reaction subscale.

Table 10. *Distribution of three subscales of examination anxiety scale*

Sl no	Subscale	Item no	Total items
1	Bodily Symptoms (EAS/Bo)	4,6,11,6,18,32	6
2	Cognitive (Worry Subscale EAS/W)	12,13,14,24,37	5
3	Emotional reaction (EAS/E)	8,9,19,29,36	5
4	Behavioural reaction (EAS/Be)	10,20,25,26,31	5
	Total		21

4.4. Standardization of Examination Anxiety Scale

For the calculation of norms, Z scores were calculated for each raw-score.

Table 11. *Z score for each raw score*

Serial no	Raw score	Z score	Serial no	Raw score	Z score	Serial no	Raw score	Z score
1.	33	-2.447	2.	34	-2.364	3.	35	-2.197
4.	36	-2.197	5.	37	-2.113	6.	38	-2.030
7.	39	-1.946	8.	41	-1.779	9.	42	-1.696
10.	43	-1.612	11.	44	-1.529	12.	45	-1.446
13.	46	-1.362	14.	47	-1.279	15.	48	-1.195
16.	49	-1.112	17.	50	-1.028	18.	51	-0.945
19.	52	-0.861	20.	53	-0.778	21.	54	-0.695
22.	55	-0.611	23.	56	-0.528	24.	57	-0.444
25.	58	-0.361	26.	59	-0.277	27.	60	-0.194
28.	61	-0.110	29.	62	-0.027	30.	63	0.056
31.	64	0.139	32.	65	0.222	33.	66	0.306
34.	67	0.389	35.	68	0.473	36.	69	0.556
37.	70	0.640	38.	71	0.723	39.	72	0.807
40.	73	0.890	41.	74	0.974	42.	75	1.057
43.	76	1.140	44.	77	1.224	45.	78	1.307
46.	79	1.391	47.	80	1.474	48.	81	1.558
49.	82	1.641	50.	83	1.725	51.	84	1.808
52.	85	1.891	53.	86	1.975	54.	87	2.058
55.	88	2.142	56.	89	2.225	57.	94	2.643

Table 11 shows the Z-score of each raw score of adolescent students. After calculating the Z-scores for all the raw scores, the range of Z-scores were divided into three levels, according to their corresponding raw scores, as shown in Table 12. According to Table 12, students having score >74 had high examination anxiety, students having score between 51 to 74, had avarage examination anxiety and students with score<51 had low examination anxiety. The finding of

Table 13 present that 16% of students belong the high examination anxiety category, 66% of students in the avarage examination anxiety category and 18% of students remains in the high examination anxiety category.

Table 12. *Norms for interpretation of Z score*

Sl no.	Range of raw score	Range of Z score	Level of examination anxiety.
1	Below 51	Below -1	Low examination anxiety
2	51 to 74	-1 to +1	Average examination anxiety
3	Above 74	Above 1	High examination Anxiety

Table 13. *Distribution of the sample in different levels of Examination Anxiety*

Sl no	Levels of anxiety	No of students	Percentage.
1.	high	325	16
2.	average	1340	66
3.	low	365	18

4. DISCUSSION and CONCLUSION

It is worthy to note that examination anxiety based research studies are often found in the case of children, but examination oriented research studies for an adolescent are rarely found especially in the context of West Bengal. This research successfully strives to measure the examination anxiety of the adolescents of the age group of 13-15 years and subsequently identify the students suffering from high examination anxiety. As the constructed Examination Anxiety Scale (EAS) having 21 items, possess high completion rate nearly 100 percent so it may be inferred that the scale may be administered easily with minimum supervision. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) have been conducted for determining construct validity. CFA has been applied to the different sample group, consisting of 402 number of adolescent students of the age group 13-15 years. Exploratory factor analysis revealed 4 factors, which was named as, Bodily Symptoms, Cognitive Dimension, Behavioural Reaction, and Emotional Reaction. EFA also yields 21 items. The final version of EAS, which was consisted of 21 items containing 15 positive items and 6 negative items. All the items have factor loading greater than .40. The final version of EAS was found to explain 44.16% of the total variance.

The final scale was applied, and confirmatory factor analysis has been executed on a sample of 402 adolescent students. All the fit indices such as, χ^2/Df, *RMSEA, GFI, AGFI, CFI* all were above acceptable values. CFA result showed that EAS has a good fit model, and it also confirmed the result of EFA. In this study, examination anxiety score of adolescent students is highly correlated with test anxiety inventory score by C.D Spielberger. The strong positive correlation between the constructed scale and Test Anxiety Inventory by C.D Spielberger which is considered as a standard established instrument is indicative of a high concurrent validity.

The constructed Examination Anxiety Scale (EAS) efficiently assesses the examination anxiety of adolescent students because it has high internal consistency reliability (Cronbach’s alpha correlation coefficient 0.764), high test-retest reliability (0.801) and excellent split-half reliability (0.767)

In this study, the norms show that 16% of students belong to the high examination anxiety group, 66% of students in the average examination anxiety group, and 18 % in the low examination anxiety group. This result stands nearly similar to the findings of Mary et al. (2014). In their study, they found that 8% of students remain in the high examination anxiety category, 74 % of students remain in the average examination anxiety category, and 18 % of students in the low examination anxiety category.

This research enables to identify the problems based on high examination anxiety, and at the same time, it unleashes the purview for the teachers to resolve the problems through guidance and counselling. This study also corroborates that high examination anxiety oriented issues can be resolved by creating an ambient and congenial environment among the family members. It provides a relevant indication towards the faulty evaluation system, which causes examination anxiety among adolescents**.**

Acknowledgements

Authors acknowledge the experts of language and subject, students, teachers, school authority and parents/ guardians of the participants for their support and contribution 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 author(s).

ORCID

Nargis Abbasi https://orcid.org/0000-0001-6267-7859
Shilpi Ghosh https://orcid.org/0000-0001-9076-9873

5. REFERENCES

Abbasi, N., & Ghosh, S. (2020). Effect of yoga on examination anxiety: A systematic review. *International Journal of Psychosocial Rehabilitation, 24*(8), 9113-9124.

Akkus, A. (2019). Developing a scale to measure student's attitude towards science. *International Journal of Assessment Tools in Education, 6*(4), 706-720.

Anastasi, A., & Ubrina, S. (2005). Psychological testing. New Delhi, India: Prentice Hall of India Private Limited.

Browne, B. M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In: Bollen, K. A. and Long, J. S. (Eds). *Testing structural equation models*, Beverly Hills, CA: Sage.

Cassady, J. C., & Johnson, R. E. (2001). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*, 270-295.

Covington, M. V. (1985). *Test anxiety: Causes and effects over time*. In H. M. van der Ploeg, R. Schwarzer, & C. D. Spielberger (Eds.), *Advances in Test Anxiety Research, 4*, 55-66. Lisse, The Netherlands: Swets & Zeitlinger

Embse, N. V., Kilgus, S., Segool, N., & Putwain, D. (2013). Test anxiety interventions of children and adolescents: A systematic review of treatment studies from 2000-2010. *Psychology in the Schools*, *50*, 57-71. https://doi.org/10.1002/pits.21660

Freeman, F. S. (1960). *Theory and practice of psychological testing*. New Delhi: Oxford & IBH Publishing Co. Pvt Ltd.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77. https://doi.org/10.2307/1170348

Hodapp, V., & Benson, J. (1997). The multidimensionality of test anxiety: A test of different models. *Anxiety, Stress and Coping, 10*(3), 219-244. https://doi.org/10.1080/10615809708249302

Leech, L.N., Barrett, C.K & Morgan, A.G. (2005). *SPSS for intermediate statistics*. London: Lawrance Erbaum Associates Publisher.

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports, 20*(3, PT. 1), 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. (2008). The Test Anxiety Inventory for Children and Adolescents (TAICA): Examination of the properties of a new multidimensional measure of test anxiety among elementary and secondary school students. *Journal of Psychoeducational Assessment*, *26*(3), 215-230. https://doi.org/10.1177/0734282907303760

MacCallum, R.C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modelling. *Psychological Methods*, *1*, 130-149

Mandler, G., & Sarason, S. B. (1952). A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, *47*, 166-173

Mary, R. A., Marslin, G., Franklin, G., & Sheeba, C. J. (2014). Test anxiety levels of board exam going students in Tamil Nadu. *Hindwai Publishing Corporation Biomed Research International*, 1-9. https://doi.org/10.1155/2014/578323

Mcdonald, S. A. (2010). The prevalence and effect of test anxiety in school children. *An International Journal of experimental educational psychology*, *21*(1), 89-101.

Sarason, I. G. (1958). Interrelationship among individual difference variables, behavior in psychotherapy, and verbal condition. *Journal of Abnormal and Social Psychology*, *56*, 339-344.

Sarason, I. G. (1959). Intellectual and personality correlates test anxiety. *The Journal of Abnormal and Social Psychology, 59*(2), 272-275. https://doi.org/10.1037/h0042200

Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., & Ruebush, B. K. (1960). *Anxiety in elementary school children: A report of research*. John Wiley & Sons Inc. https://doi.org/10.1037/14349-000

Sarason, I. G. (1984). Stress, anxiety and cognitive interference reactions to test. *Journal of Personality and Social Psychology*, *66*(4), 929-938.

Shukla, J. U. (2013). A study of examination Anxiety among secondary school students in the context of some variable. (Unpublished doctoral dessertation). Gujrat University, Ahmedabad, India.

Spielberger, C. D. (2010). *Test Anxiety Inventory*. New Jersey, USA: John Wiley & Sons, Inc.

Suinn, R. M. (1969). The STABS, a measure of test anxiety for behavior therapy: Normative data. *Behavior Research and Therapy*, *7*, 335-339.

Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76*(2), 92-104. https://doi.org/10.1037/h0031332

Zeidner, M. (1992). How to high school and college students cope with test situations? *British Journal of Educational Psychology, 66*, 115-128.

Zeidner, M. (1998). *Test Anxiety: The State of the art*. New York: Plenum Press.

6. APPENDIX

Table A1 displays the 21 items of standardized English verison of examination anxiety scale for adolescent students.

Table A1. *Examination Anxiety Scale for adolescent students (Standardized English version)*

Statements	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree	Total
1. When I sit for an important examination, I feel thrilled.						
2. During examination I frequently feel the urgency to go to toilet.						
3. While taking an examination I feel uneasy and upset.						
4. Even after preparing well for the examination I feel very nervous.						
5. I often look at the other people during exams.						
6. My hands often sweat and feel cold before and during examination.						
7. During examination I think that I will surely pass the examination and get promoted.						
8. Worry about the result of the examination interferes with my performance during examination.						
9. After examination I think most of my answers are right.						
10. Sometime I tremble before or during examination						
11. During an important examination I suffer from headache.						
12. I feel relaxed while taking an examination.						
13. During examination I often check the time.						
14. After an examination I say to myself, it is over and I did my best.						
15. I never play with my pencil or pen during an examination.						
16. My thought wander during examination.						
17. I feel very confident while I taking an examination.						
18. I think about current events during an examination.						
19. My mouth becomes dry before or during an important examination.						
20. I feel very jittery when talking an important examination.						
21. Before or during examination I think other students are brighter than me.						

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 535–548
https://doi.org/10.21449/ijate.647488
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Development of Language Awareness Scale Regarding Daily Life

Mustafa Erol [1,*], Ismail Karakaya [2]

[1]Primary School Education, Faculty of Education, Yıldız Technical University, İstanbul, Turkey
[2]Faculty of Education, Gazi University, Ankara, Turkey

ARTICLE HISTORY

Received: Nov. 15, 2019
Revised: May. 06, 2020
Accepted: Sep. 10, 2020

KEYWORDS

Turkish Consciousness,
Language Awareness,
Social Media,
Mass Media

Abstract: The aim of this study is to develop the language awareness scale regarding daily life. The study group consisted of 606 undergraduate students studying at a university in Istanbul. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were used in the study. EFA result indicates that the scale consists of 17 items and 4 factors and 67% of the total variance is explained. Factors were named as Individual Awareness, Social Media Awareness, Awareness Regarding Daily Life and Awareness in the Mass Media. As a result of the CFA, it was determined that the scale had 17 items and the fit indices of the structure were sufficient ($\chi 2$ / sd = 2.54, RMSEA = 069, SRMR =.07). Item-total correlations of the scale were found to range between 79 and 89. Cronbach Alpha internal consistency coefficient of the scale was found to be as 86. Based on these findings, it can be said that the scale can be used in a valid and reliable way to measure students' language awareness about daily life.

1. INTRODUCTION

Language is the most fundamental feature that distinguishes human from other living beings. Language helps the individual in many aspects such as being able to generate thoughts, expressing thoughts, acquiring information, remembering the past, living the day, directing the future, gaining personality, sustaining life, communicating and understanding (Ağca, 2001; Demir & Yılmaz, 2009; Yaman, 2015).

Factors such as economic, cultural and political relations between communities and nations; migrations, travels, scientific studies and the foreign language activities' becoming easier, the necessity/ desire to learn foreign languages have resulted in interaction between languages (Sarı, 2013; Zengin, 2017). This interaction has become more prominent in recent years with the development of technology (social media, television, smart phones, internet, etc.) (Zengin, 2017). Throughout history, like all the languages, Turkish language has both changed and branched off in all eras for various reasons, both in the form of changes that stem from the language's own natural structure and external factors such as various geographical distributions and relations with different socio-cultural environments (Özyetgin, 2006). Since the first known written documents (Chinese, Sanskritic, Mongolian, Arabic, Persian, Italian, Greek, Armenian, French, German and English, etc.), our language has exchanged words with various languages.

CONTACT: Mustafa Erol ✉ merol@yildiz.edu.tr Primary School Education, Faculty of Education, Yıldız Technical University, İstanbul, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

This interaction has mostly been in the form of word exchange and has not spoiled the sentence structure and functioning of our language. Because language mechanism that is the most resistant to change is syntax (Sarı, 2013). Today, the situation is completely different. It is seen that especially English affects our language in many ways and that it influences the sentence structure and functioning of Turkish in every field (Yaman, 2015). Receiving or using foreign words/letters although they have Turkish equivalents both affects the Turkish vocabulary negatively and ruins the beauty, naturalness and essence of Turkish (Akalın, 2000; Öner, 2006; Tosun, 2005; Ünalan, 2006).

Changes in the languages of nations with historical background can be considered normal; however, while taking words from other languages, also taking their rules and using it in ones' languages disrupts the structure, phonology, semantics, pronunciation, spelling and reading rules and traditions of the language in question, and since it causes disorder in the language, the language starts to corrupt (Tosun, 2005). The concept of corruption is not a problem related to the language itself, but a problem related to the users of the language (Buran, 2006; İpek, 2015). Because the preference of foreign elements in the language does not stem from the language itself but individuals' preferences (Gülsevin, 2006). Language awareness can be defined as "a conscious language usage sensitivity that the individual has developed aiming at the right and efficient use of language ranging from his/her choice of words in a way that s/he can control his/her own oral and written language use to morphological, syntactic and semantic structure accuracy, from spelling and punctuation rules to the ability to organize and transfer thoughts" (Büyükkantarcıoğlu, 2003; Carter, 2003). The term language awareness is used in the sense of consciousness, sensitivity and a gradually developing mental process developed by the individual regarding the characteristics and use of his or her own language (Ali, 2011; Büyükkantarcıoğlu, 2006). Ellis (2012) states that language awareness includes processes that can be obtained by looking at the accumulation of knowledge about language, from a conscious understanding of how languages work, how people learn and use them.

When the related literature was examined, many articles, books and declarations (Akalın, 2000; Aktaş & Şentürk, 2014; Alpay, 2015; Alyılmaz, 2010; Aslan & Kılıç, 2012; Bağcı-Ayrancı, 2017; Demir &Yapıcı, 2007; Erdoğan & Gök, 2009; Ersoylu, 2009; Girmen, Kaya, & Bayrak, 2010; Göçer, 2013; Gülsevin, 2006; İpek, 2015; Kolaç, 2008; Özçelik, 2006; Şenyuva, Ertüzün, Turhan, & Demir, 2017; Sever, 2001; Ulaş & Sevim, 2010; Yaman, 2015; Zengin, 2017) were accessed revealing the problems of Turkish language and solution offers regarding these problems, the extent to which individuals are aware of these problems and their awareness of these problems. Only one study that measures Turkish language awareness was accessed. This is the study developed by Yaman (2011), called "Turkish Consciousness Scale: Validity and Reliability Study". However, a study measuring the language awareness of individuals about daily life could not be reached. This study aims to measure individuals' awareness of language regarding their daily lives. Accordingly, the aim of the study can be specified as "to develop a language awareness scale related to daily life.

2. METHOD

2.1. Study Group

The research was conducted with 606 university students whose ages range between 18-32, studying at a university in Istanbul, Turkey. Within the scope of the study, initially, exploratory factor analysis (EFA) was performed with the data obtained from 310 students. Afterwards, confirmatory factor analysis (CFA) was performed with the data obtained from exploratory factor analysis (EFA) and 296 students. In addition, data were collected from 100 prospective teachers at three-week intervals and test-retest reliability was calculated. 51.8% of the university students in the study group were female (%) and 48.2% were male. The mean age of the university students participating in the study is 21.3. 20% of the teacher candidates were studying in the Primary School Teacher, 15.2% Social Studies Teaching, 9.2% Mathematics Teaching, 9.1% Science Teaching, 7.8% Computer and Instructional Technologies Teaching, 18.3% Turkish Language Teaching, 9.4% English Language Teaching and 11% Preschool Teaching.

2.2. Data Collection Tools

2.2.1. Personal Information Form

This form was prepared by the researcher(s) to find out the demographic information of the individuals in the study group. The form contains items aimed at determining some information about teacher candidates such as their ages and departments.

2.2.2. Stages of Developing the Scale of Language Awareness regarding Daily Life

Item pooling phase; In order to determine the items of the measurement tool, research studies in the literature and the developed measurement tools were examined (Akalın, 2000; Alyılmaz, 2010; Aslan & Kılıç, 2012; Bağcı-Ayrancı, 2017; Büyükkantarcıoğlu, 2006; Demir & Yapıcı, 2007; Erdoğan & Gök, 2009; Ersoylu, 2009; Girmen, Kaya & Bayrak, 2010; Gülsevin, 2006; İpek, 2015; Kolaç, 2008; Özçelik, 2006; Şenyuva, Ertüzün, Turhan, & Demir, 2017; Sever, 2001; Ulaş & Sevim, 2010) and 13 university students were asked five questions including the sub-dimensions of the scale and a pool of 46 items was formed.

The draft form with 46 items prepared consists of four sub-dimensions. In addition, whether the items were appropriate in terms of language and expression, their clarity and scientific appropriateness were examined and the necessary corrections were made. Negative items in the measurement tool are scored in reverse. The maximum score that can be obtained from the scale is 85 and the lowest score is 17. As a result of the additivity test, the analysis results regarding the scale's assessability, also on the basis of total score and sub-dimension were presented in the findings section.

Expert Opinion Stage (Scope and Appearance Validity); For the content and appearance validity, a 45-item draft was sent to two faculty members with a PhD in Turkish linguistics, two faculty members with a PhD in Turkish teaching, two faculty members with a PhD in classroom teaching, one faculty member with a PhD in assessment and evaluation and one faculty member who is an expert on the field of language validity. The experts were asked to evaluate the items in terms of "eligibility", "clarity" and "intelligibility" criteria and in terms of their appropriateness for the sub-dimension including the items. Experts evaluated each item considering the Lawshe analysis method according to three criteria: "appropriate", "partially appropriate", "inappropriate" and the content validity index was determined. Content validity index "(CVI) is obtained by 1 less than the ratio of the number of experts indicating the Required" opinion of any item to the total number of experts indicating the opinion of the article (Yurdugül, 2005).

In line with the opinions received from the experts, arrangements were made on the relevant items and the the measurement tool was given its final form. Among the items in the scale, the item "*We must protect our language just as we protect our flag*" was removed from the scale in line with the views of 5 of the 8 experts. It was decided that the items were capable of measuring the relevant structure in accordance with the feedback from experts regarding the validity of appearance. Based on the opinions of the experts regarding the items, the content validity index was calculated and the findings were presented in Table 1. 8 experts evaluated the pool of 45 items prepared according to Table 1 and the content validity index (CVI) was determined as 92.

Table 1. *Results regarding the content validity index*

Item Numbers	A	PA	I	CVI
Item 1	8	0	0	1.00
Item 2	8	0	0	1.00
Item 3	8	0	0	1.00
Item 4	8	0	0	1.00
…..	…	…	…	…
Article 32	5	2	1	0.25
….	…	…	…	…
Article 45	8	0	0	1.00
Number of Experts	8			
Content Validity Index (CVI)	0.92			

** A = Appropriate, PA = Partially Appropriate, I = Inappropriate, CVI = Content Validity Index*

2.3. Data Collection

The data of the study were collected in the 2018-2019 academic year. The data of the scale of language awareness regarding daily life were collected by the researchers. The data of the study were obtained with the help of Google forms from the researchers, and one-to-one from undergraduate students.

2.4. Data Analysis

Before starting the data analysis process, the data set was examined for missing data and extreme values (examined by Box-Plot graph) and the data containing the missing data and extreme values were removed from the data set. In the data set, seven data containing missing data and extreme values were removed from the data set. Afterwards, Kolmogorov-Smirnov test was performed for normality test and it was determined that the data set showed normal distribution (Z =.043. p =.200). Regarding normal distribution, Tabachnick and Fidell (2013) state that kurtosis and skewness values' being between -1.5 and +1.5 will meet the assumption of normality. Within the scope of the study, the kurtosis and skewness values were determined as -06 to -24. In addition, linear regression hypothesis was tested by scatter diagram (Kalaycı, 2016) and it was found that there was a linear relationship between dependent and independent variables. Following these procedures, the following operations were performed within the scope of the study.

Validity procedures: In order to reveal the structure of the scale, content and appearance validity (expert opinion. Content validity index), criterion validity-concurrence and predictive validity (Pearson moment-product correlation coefficient, regression analysis), exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were performed. The results regarding the validity processes are presented in the findings.

Reliability procedures: In order to determine the reliability of the scale, item analysis (Pearson product-moment correlation coefficient), analyzes aimed at the entire test (standard deviation, variance, standard error) and internal consistency coefficients (Cronbach α) techniques were used.

3. RESULTS / FINDINGS

The findings of this study, which aims to develop a Language Awareness Scale regarding Daily Life, were presented in two sub-headings as findings related to validity and reliability analyzes.

3.1. Findings regarding Validity Analyzes

3.1.1 Exploratory Factor Analysis (EFA)

Exploratory factor analysis was performed to determine the construct validity of the measurement tool. Exploratory factor analysis is a technique used to determine under how many sub-dimensions the items (variables) in a measurement tool prepared as a draft and applied will be gathered and to detect the type of relationship between these items (Seçer, 2015; Sönmez & Alacapınar, 2016). Below .40 item variance, 29 items below .50 which have overlapping characteristics were removed from the measuring device. When the findings of the remaining items were examined, the Kaiser-Mayer-Olkin (KMO) value of the scale was found to be .85 and Bartlett's Sphericity Test value was found to be .000 (*p* <.05). That KMO value is .85, which indicates that the data is suitable for factor analysis (Kalaycı, 2016). Common variance values of items in the scale range between .51 and .78. Factors with eigenvalues greater than 1 were considered to determine the scale's number of factors and scatter diagram was presented in Figure 1.

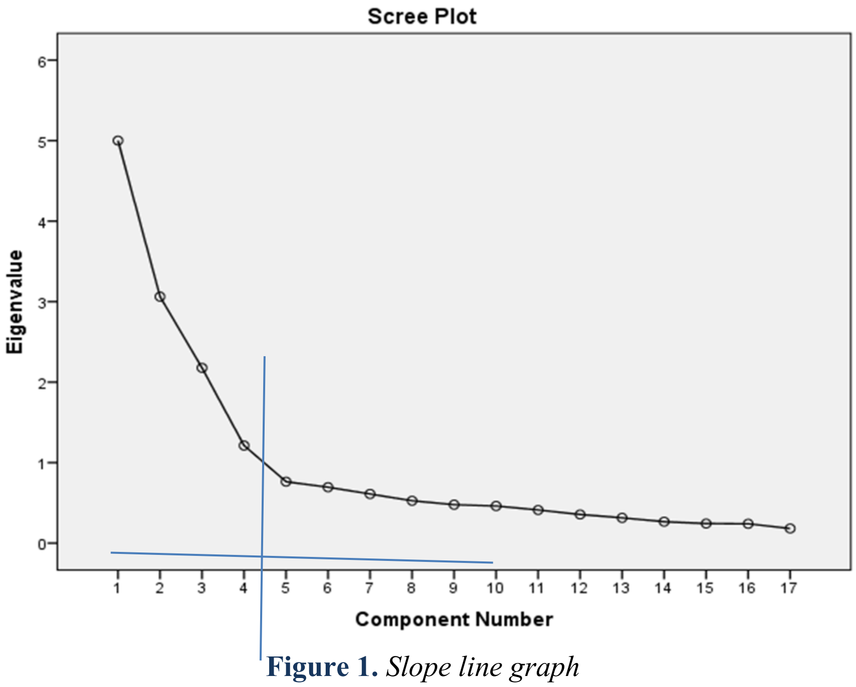


Figure 1. *Slope line graph*

According to Figure 1, it is possible to say that the scale is not separated with very strict lines after the fourth point and therefore consists of four factors. Detailed information on these components is provided in Table 2.

Table 2. *Number of factors related to eigenvalue statistics and explained variance ratio*

	Initial Eigenvalues			Sum of Square Loading		
Components	Total	Variance %	Collected%	Total	Variance %	Collected%
1. Component	5,001	29,417	29,417	4,061	23,887	23,887
2. Component	3,062	18,011	47,427	2,510	14,764	38,651
3. Component	2,177	12,806	60,234	2,455	14,440	53,091
4. Component	1,213	7,133	67,367	2,427	14,275	67,367

When Table 2 is analyzed, four factors with eigenvalues greater than 1 and the variance ratios explained by these factors are seen. It is recommended that, according to Kaiser Criterion, factors with eigenvalues above 1 be kept during factor extraction (Büyüköztürk, 2017). According to Özdamar (2017), determining the eigenvalues as much as the number of eigenvalues greater than one is the most commonly used factor determination criterion. The first factor explains 23.89% of the total variance, the second factor explains 14.76% of the total variance, the third factor explains 14.44% of the total variance, and the fourth factor explains 14.27%. Together, these four factors account for 67.38% of the total variance. As it was stated that this ratio needs to be at least 52% the obtained value was found sufficient (Henson & Roberts, 2006). The number of factors in the measurement tool can be interpreted after they are determined. In order to obtain meaningful factors and to determine the distribution of the items to the factors, verimax rotation was performed and the results were presented in Table 3 below.

Table 3. *Factor analysis results after varimax rotation*

		Factors			
	Items	1	2	3	4
Individual Awareness	1- I think that Turkish words should be derived to place non-Turkish words	**.844**			
	2- The use of our language with foreign word patterns damages our language (For example; **Cafe Sorgun**, **Otel The Yozgat** etc.)	**.840**			
	3- Speaking with only Turkish words and words translated into Turkish is an indication of backwardness	**.804**			
	4- I warn my friends who use foreign words despite having Turkish equivalents while having a conversation.	**.744**			
	5- When I come across a foreign word in a text I read, I look up its Turkish equivalent from the dictionary.	**.744**			
	6- I think that as individuals, we should speak Turkish properly in our daily lives.	**.743**			
	7- Wearing clothes with foreign words on them makes me uncomfortable.	**.728**			
	8- It bothers me if a text I read has foreign words used despite having Turkish equivalents	**.660**			
Social Media Awareness	9- The use of letters that are not in our alphabet (w, q, x) in social media bothers me (For example; **wadi** instead of **vadi** etc.)		**.839**		
	10- I warn my friends who misspell Turkish words on social media		**.759**		
	11- I approve the use abbreviated words (For example; **mrb** instead of **merhaba** etc.)		**.683**		
Awareness regarding Daily Life	12- I feel uncomfortable when I see foreign names given to the main roads and streets			**.877**	
	13- I am not bothered by seeing signs written with foreign words around me.			**.868**	
	14- I am bothered by seeing workplaces with foreign names around me.			**.862**	
Awareness in Mass Media	15- It is not important for me whether the language in the mass media is used in accordance with the rules of language				**.759**
	16- I feel uncomfortable that the Turkish pronunciation of foreign words used in mass media change from person to person in Turkish.				**.687**
	17- Programs with excessive use of local dialects should be expanded.				**.635**

When Table 3 is examined, the factor loadings of the individual awareness factor (8 items) of the scale are found to range between .66 and .84; the load values of social media awareness factor (3 items) range between .68 and .84; the load values of the awareness factor (3 items) regarding daily life range between .86 and .88, and the load values of the awareness factor in mass media (3 items) range between .64 and .76. The sub-factors of the scale were determined by scanning literature in the related field and experts.

3.1.2. *Confirmatory Factor Analysis (CFA)*

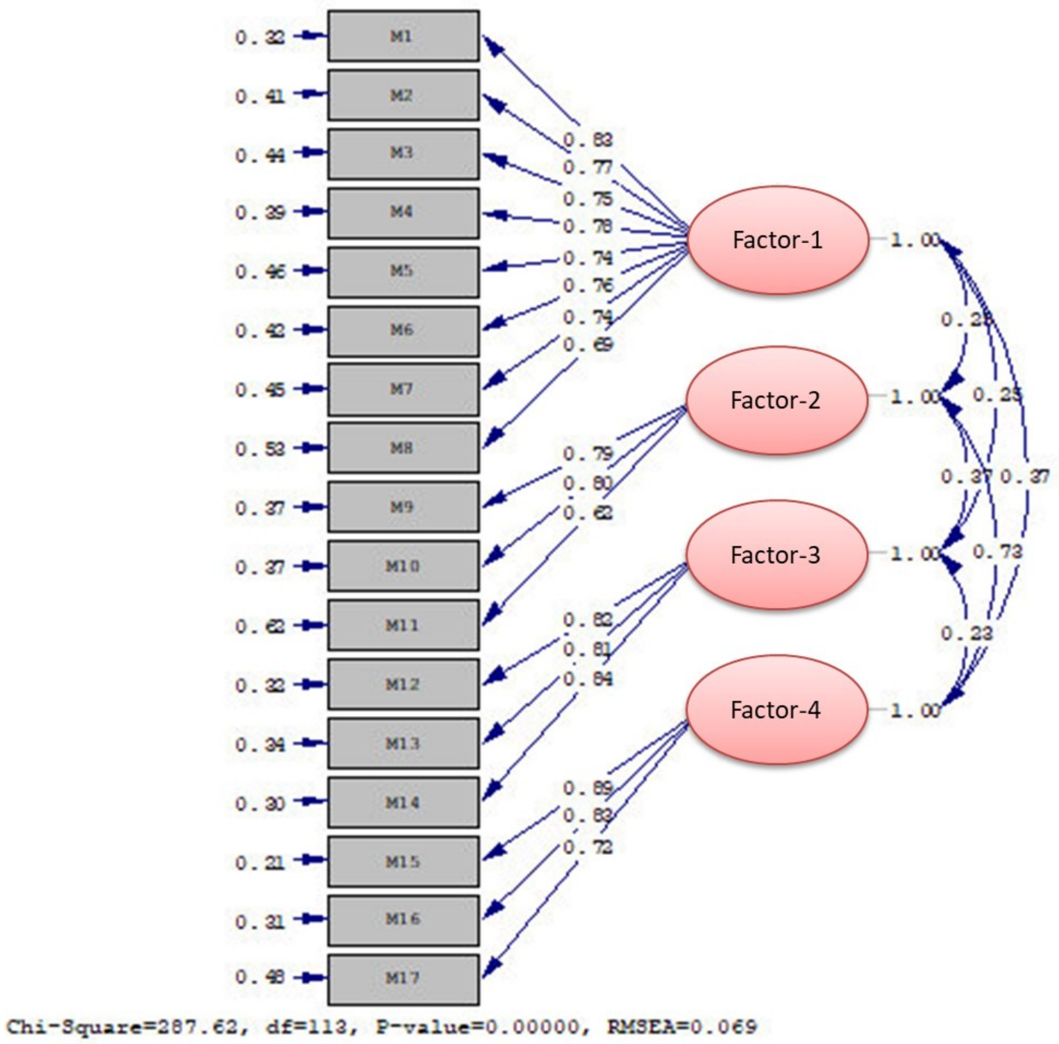
Confirmatory factor analysis was performed to test the model obtained as a result of exploratory factor analysis (EFA). Confirmatory factor analysis is the examination of whether the model formed as a result of exploratory factor analysis is validated (complies with the structure) (Özdamar, 2017; Seçer, 2015; Sönmez & Alacapınar, 2016). This analysis was conducted with a different study group than the group on which exploratory factor analysis was performed. The study group in which CFA was conducted consisted of 287 university students. In order to evaluate the results of the CFA, the fit indices were examined. At this point, fit indices such as the chi-square ratio divided by the degree of freedom (*χ2 / df*), *RMSEA* (Root Mean Square Error of Appropximation), *GFI* (Goodness of Fit Index), *AGFI* (Adjusted Goodness of Fit Index), *CFI* (Comperative Fit Index) and *SRMR* (Standardized Root Mean Square Residual) were calculated. The determined indices were interpreted with reference to the value ranges specified by Büyüköztürk, Şekercioğlu and Çokluk (2015). Statistical data of the fit indices are presented as Table 4.

Table 4. *Findings regarding fit indices*

Fit Indices	Perfect Fit Criterion	Good Fit Criterion	Value	Concordance Level
χ2 / sd	0 ≤ χ2 / sd ≤ 2	2 ≤ χ2 / sd ≤ 3	2.54	Good Fit
RMSEA	.00 ≤ RMSEA ≤ .05	.05 ≤ RMSEA ≤ .08	.069	Good Fit
AGFI	.90 ≤ AGFI ≤ 1.00	.85 ≤ AGFI ≤ .90	.87	Good Fit
GFI	.95 ≤ GFI ≤ 1.00	.90 ≤ GFI ≤ .95	.90	Perfect Fit
CFI	.95 ≤ CFI ≤ 1.00	.90 ≤ CFI ≤ .95	.97	Perfect Fit
NFI	.95 ≤ NFI ≤ 1.00	.90 ≤ NFI ≤ .95	.95	Perfect Fit
NNFI	.95 ≤ NNFI≤ 1.00	.90 ≤ NNFI≤.95	.96	Perfect Fit
RFI	.95 ≤ RFI ≤ 1.00	.90 ≤ RFI ≤.95	.94	Good Fit
IFI	.95 ≤ IFI ≤ 1.00	.90 ≤ IFI ≤ .95	.97	Perfect Fit
SRMR	.00 ≤ SRMR ≤.05	.05 ≤ SRMR ≤.10	.07	Good Fit
PNFI	.95 ≤ PNFI ≤ 1.00	.50 ≤ PNFI ≤.95	.80	Good Fit
PGFI	.95 ≤ PGFI ≤ 1.00	.50 ≤ PGFI ≤.95	.68	Good Fit

* The fit indices in Table 4 have been prepared with reference to Büyüköztürk, Şekercioğlu and Çokluk (2015).

As Table 4 shows, when the fit indices obtained from the confirmatory factor analysis are evaluated together, it is seen that the four-factor structure of the scale with 17 items has a good fit. The path diagrams and items structure parameters obtained from the first and second level confirmatory factor analysis are shown in Figure 2 and Figure 3 below.



Chi-Square=287.62, df=113, P-value=0.00000, RMSEA=0.069

(Factor1 = Individual Awareness, Factor2 = Social Media Awareness, Factor3 = Awareness regarding Daily Life, Factor4 = Awareness in Mass Media)

Figure 2. *Path chart obtained by correlated traits model confirmatory factor analysis.*

When the first and second level confirmatory factor analysis outputs in Figure 2 and Figure 3 were examined, it was determined that the standardized factor loadings between the items in the measurement tool and the structures that the items aimed to measure were statistically significant according to the *t* value. Therefore, it is seen that the scores of 17 items in the measurement tool measure the sub-dimensions that make up the structure of the language awareness skills scale related to daily life and factorial validity is provided.

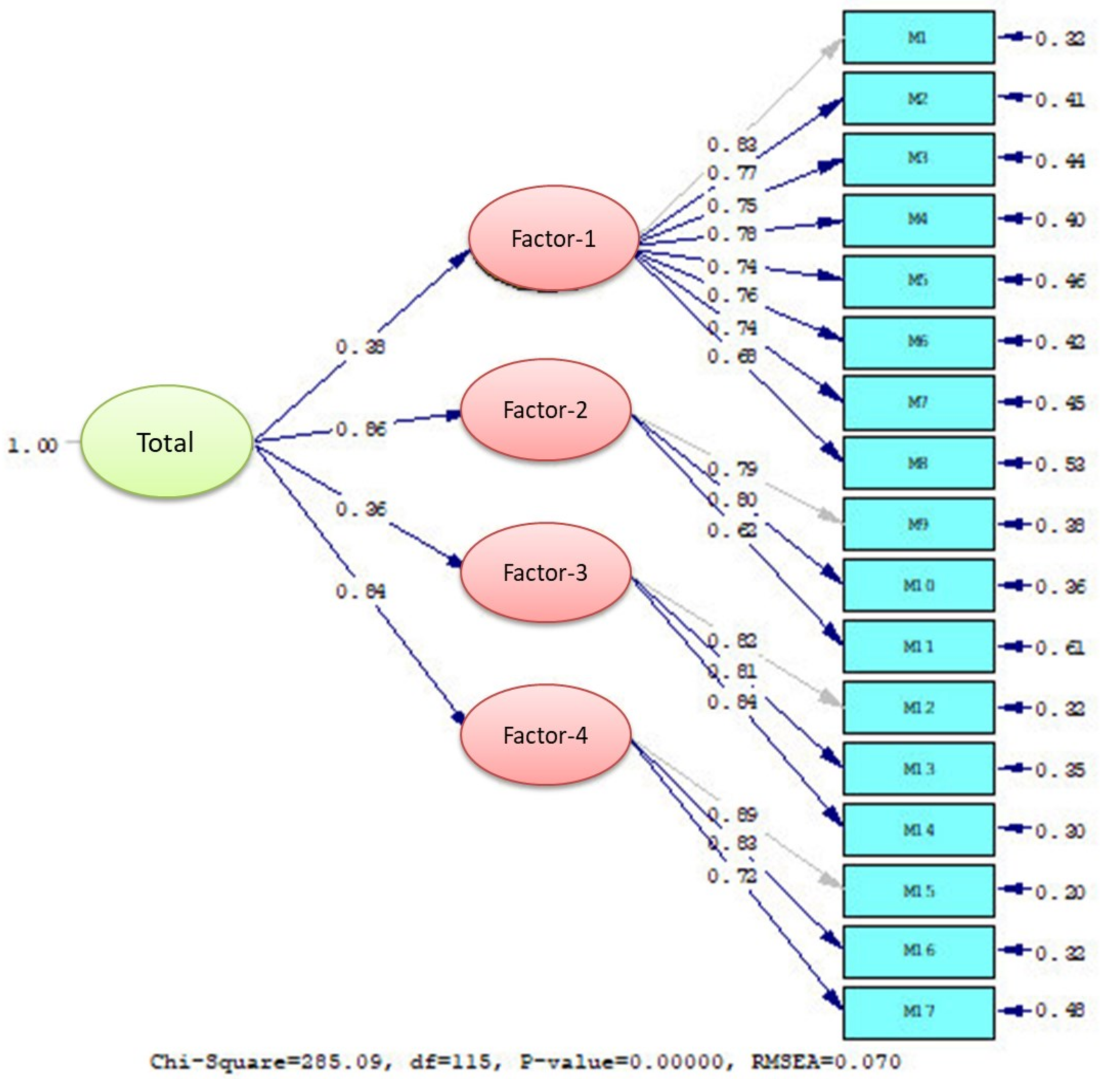


Figure 3. *Path chart obtained by second order factor analysis. (Factor1 = Individual Awareness, Factor2 = Social Media Awareness, Factor3 = Awareness regarding Daily Life, Factor4 = Awareness in Mass Media).*

3.2. Findings on Reliability Analyzes

Cronbach's Alpha value was calculated to determine the internal consistency coefficients of the scale and the results are given in the Table 5. When Table 5 is examined, the internal consistency (Cronbach's alpha) coefficient of the "Language Awareness Scale regarding Daily Life" is found to be .86 and internal consistency (Cronbach Alpha) coefficients regarding their sub-dimensions were found to range between .79 and .89. A reliability coefficient computed between .79-.86 for a test indicates that the test is reliable (Kalaycı, 2016; Özdamar, 2017). According to Bayram (2004) and Büyüköztürk (2017), a Cronbach's Alpha value above .70 can be regarded as appropriate in terms of reliability.

Table 5. *Findings on reliability coefficients*

Dimensions	Mean	Variance	Standard Deviation	Number of Items	Internal Consistency (Cronbach Alpha) Coefficient
Individual Awareness	37.1097	17.839	4.22363	8	.89
Social Media	12.7516	14.556	3.81526	3	.79
Daily Life	12.3065	6.213	2.49263	3	.86
Mass Media	15.3097	15.308	3.91258	3	.83

Table 6. *Findings regarding item statistics*

Item No.	Item Inference Test Average	Item Inference Test Variance	Adjusted Item - Total Correlation	Squared Multiple Correlation	Item Inference Cronbach Alpha Value	*t*-value	sd/*p*
M14	73.3581	89.519	.316	.555	.850	77.237	
M13	73.3645	89.572	.317	.580	.850	79.748	
M15	73.4032	88.377	.380	.583	.847	76.184	
M8	72.8710	90.572	.334	.444	.849	71.108	
M4	72.8613	88.696	.494	.580	.844	88.887	
M3	72.7097	90.867	.449	.534	.846	71.716	
M7	72.8645	90.454	.360	.432	.848	64.143	
M6	72.9935	88.667	.424	.468	.846	70.383	
M2	72.7355	90.635	.424	.649	.847	68.011	sd=167
M1	72.8290	89.857	.459	.698	.845	55.423	*p<.01
M5	72.8452	89.542	.442	.462	.845	51.931	
M11	74.1548	82.830	.560	.552	.839	53.611	
M16	73.5452	81.757	.632	.588	.835	33.552	
M10	73.5774	81.190	.657	.682	.834	40.682	
M12	73.7903	83.034	.575	.413	.838	38.209	
M9	74.5419	83.505	.510	.608	.842	31.586	
M17	73.8645	82.538	.521	.522	.841	45.886	

According to Table 6, total correlations of items in the scale are found to range between .32 and .66. Since the threshold value for the corrected-item total correlations is .30, it can be stated that the items under each component adequately measured the desired construct (Büyüköztürk, 2017).

4. DISCUSSION and CONCLUSION

As a result of the factor analysis conducted to determine the construct validity of the language awareness scale regarding daily life; the factors, the slope line graph and the eigenvalues of which were higher than 1 were examined and the scale was found to have a four-factor structure. These four factors explain 67% of the total variance. When the distribution of items is examined, it is observed to fall under Individual Awareness, Social Media Awareness, Awareness regarding Daily Life, Mass Media Awareness factors. Load values of the first factor of the scale .66 and .84; load values of the second factor range between .68 and .84; the load values of the third factor range between .86 and .88 and the load values of the fourth factor range between .64 and .76. Factor loadings should be above .30 and factor loadings above .50 are accepted to be quite good (Kalaycı, 2016). When the factor loadings of the language awareness scale related to daily life are examined they appear to be over .60. When these results are taken into consideration, it can be said that the results of exploratory factor analysis (EFA) of daily language awareness scale are within acceptable limits. In addition, when the results of the first and second level confirmatory factor analysis (CFA) are examined, it is seen that the sub-dimensions that form the structure of the 17-item daily life awareness scale were measured and factorial validity was obtained.

In the study, the reliability coefficient of the scale (Cronbach's Alpha) was found to be .86 and its sub-dimensions were found to range between .79 and .89. When these results are taken into consideration, it is seen that the scale meets the reliability criteria (Büyüköztürk, 2017; Kalaycı, 2016). When the findings of the study are examined in terms of these criteria, it can be said that the whole measurement instrument developed is in a very reliable range.

Total item correlation values of the scale are found to range between .32 and .66. According to item-total correlation results in the measurement tool, it was determined that there were no items with a value less than .30. According to Büyüköztürk (2017), total correlations of the items should not be less than .30. Besides when *t* (*p* <.01) values are examined, it is seen that the items forming the scale are distinctive. When all the results of the study are evaluated together, it is seen that the scale will be used in a valid and reliable way to measure the language awareness regarding daily life.

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.

ORCID

Mustafa Erol https://orcid.org/0000-0002-1675-7070
İsmail Karakaya https://orcid.org/0000-0003-4308-6919

5. REFERENCES

Ağca, H. (2001). *Türk dili*. Ankara: Gündüz Eğitim ve Yayıncılık.

Akalın, Ş. H. (2000). *Bilişim çağı ve Türkçenin sorunları.* Erişim Tarihi 21 Aralık 2018, http://turkoloji,cu,edu,tr/DIL%20SORUNLARI/03.php

Aktaş, E., & Şentürk, L. (2014). *Türkçe eğitiminde kuramsal ve uygulamalı çalışmalar, Sosyal medyada Türkçenin kullanımı üzerine nitel bir çalışma*. Ankara: Pegem Akademi.

Ali, S. (2011). Critical Language awareness in pedagogic context. *English Language Teaching, 4*(4), 28-35. https://doi.org/10.5539/elt.v4n4p28

Alpay, N. (2015). *Türkçe sorunları kılavuzu.* İstanbul: Metis Yayıncılık.

Alyılmaz, C. (2010). Türkçenin öğretiminin sorunları. *Turkish Studies-International Periodical For The Languages, Literature and History of Turkish or Turkic, 5*(3), 729-749. https://doi.org/10.7827/TurkishStudies.1629

Aslan, A, & Kılıç, Y. (2012). Türkçe öğretmenleri ile öğretmen adaylarının Türkçe bilinç düzeyleri (Ağrı ili Örneklemi). *Turkish Studies-International Periodical For The Languages, Literature and History of Turkish or Turkic, 7(*4), 799-806. https://doi.org/10.7827/TurkishStudies.4078

Bağcı-Ayrancı, B. (2017). Öğretmen adaylarına göre Türkçenin güncel sorunları. *International Journal of Language Academy, 5*(2), 63-78. https://doi.org/10.18033/ijla.3550

Bayram, N. (2004). *Sosyal bilimlerde spss ile veri analizi*. Bursa: Ezgi Kitapevi.

Buran, A. (2006). Yozlaşma dilin kullanımıyla ilgilidir" konulu röportaj. *Bizim Külliye Dergisi, 30*, 40-44.

Büyükkantarcıoğlu, N. (2003). *Dil farkındalığı ve işlevsel dil kullanımı bağlamında anadilimiz: gözlemler, öneriler.* Cumhuriyetimizin 80, Yılında Türkçemiz, Ankara: Ankara Ticaret Odası ve Anadolu Çağdaş Eğitim Vakfı, 19-26.

Büyükkantarcıoğlu, N. (2006). *Toplumsal gerçeklik ve dil*. İstanbul: Multilingual Yayınları.

Büyüköztürk, Ş. (2017). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum* (23. Baskı). Ankara: Pegem Akademik Yayıncılık.

Büyüköztürk, Ş., Şekercioğlu, Ç., & Çokluk, Ö. (2015). *Sosyal bilimler için çok değişkenli istatistik: spss ve lisrel uygulamaları.* Ankara: Pegem Akademi Yayıncılık.

Carter, R. (2003). Language awareness. *ELT Journal, 57*(1), 64-65.

Demir, C., & Yapıcı, M. (2007). Ana dili olarak Türkçenin öğretimi ve sorunları. *Sosyal Bilimler Dergisi, 9*(2), 177-192.

Demir, N., & Yılmaz, E. (2009). *Türk dili*. Ankara: Grafiker Yayınları.

Ellis, E, M. (2012). Language awareness and its relevance to TESOL. *University of Sydney Papers in TESOL*, *7*, 1-23.

Erdoğan, T., & Gök, B. (2009). Türkçenin ana dili olarak öğretiminde karşılaşılan sorunlar ve bu sorunların giderilmesine yönelik öneriler: Ankara Örneği. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi, 3*(36), 1-16.

Ersoylu, H. (2009). *Türkiye Türkçesinin çağdaş sorunları üzerine incelemeler.* Ankara: Ötüken Neşriyat.

Girmen, P., Kaya, M. F., & Bayrak, E. (2010). *Türkçe eğitimi alanında yaşanan sorunların lisansüstü tezlere dayalı olarak belirlenmesi.* 9, Ulusal Sınıf Öğretmenliği Eğitimi Sempozyumu (20 -22 Mayıs 2010), Bildiriler, 133-138, Elazığ.

Göçer, A. (2013). Türkçe öğretmeni adaylarına göre Türkçenin güncel sorunları. *Adıyaman Üniversitesi Sosyal Bilimler Enstitüsü Dergisi Türkçenin Eğitimi Öğretimi Özel Sayısı. 6*(11), 491-515. https://doi.org/10.14520/adyusbd.466

Gülsevin, G. (2006). Dil kirliliği sorunu. In Gülsevin, G., & Boz, E. (Eds.), *Türkçenin çağdaş sorunları*. Ankara: Gazi Kitabevi.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practise. *Educational and Psychological Measurement, 66*(3), 393-416. https://doi.org/10.1177/0013164405282485

İpek, B. (2015). Bireyde dil bilinci. *Journal of Turkish Language and Literature, 1*(2), 33-44.

Kalaycı, Ş. (2016). *SPSS uygulamalı çok değişkenli istatistiktik teknikleri* (7. Baskı). Ankara: Asil Yayıncılık.

Kolaç, E. (2008). Sınıf öğretmeni adaylarının ana dilimizin yaşadığı sorunlara ilişkin farkındalıkları, görüş ve önerileri. *Uluslararası Sosyal Araştırmalar Dergisi. 1*(15), 441-455.

Öner, T. (2006). Bilişimde özenli Türkçenin önemi. *Bilişim ve Bilgisayar Mühendisliği Dergisi, 1*(1).

Özçelik, S. (2006). Türkçe ve bilim dili sorunu. In Gülsevin, G., & Boz, E. (Eds.), *Türkçenin çağdaş sorunları.* Ankara: Gazi Kitabevi.

Özdamar, K. (2017). *Ölçek ve test geliştirme yapısal eşitlik modellemesi*. Eskişehir: Nisan Kitapevi.

Özyetgin, M. A. (January, 2006). *Tarihten bugüne Türk dili alanı.* (Paper Presented at Chinese Academy of Social Science, Sino-Foreign Relationship Department of Institute of History, Beijing, China. Abstract retrieved from www.eurasianhistory.com.

Sarı, İ. (2013). Dil etkileşimi bağlamında ses-anlam eşlemesi ve Türkçedeki örnekleri. *Türk Kültürü Araştırmaları Dergisi, 1*, 1-27.

Seçer, İ. (2015). *SPSS ve LİSREL ile pratik veri analizi*. Ankara: Anı Yayıncılık.

Sever, S. (2001). Öğretim dili olarak Türkçenin sorunları ve öğretme öğrenme sürecindeki etkili yaklaşımları. *Ankara Üniversitesi Eğitim Fakültesi Dergisi, 34*, 11- 22.

Sönmez, V., & Alacapınar, F. G. (2016). *Sosyal bilimlerde ölçme aracı hazırlama*. Ankara: Anı Yayıncılık.

Şenyuva, E., Ertüzün, F., Turhan, K., & Demir, N. (2017). Türk diline ilişkin sorunlar, çözüm önerileri ve Türkçe bilinci: kuşaklararası karşılaştırma. *Uluslararası Türkçe Edebiyat Kültür Eğitim Dergisi Sayı, 6*(3), 1384-1397. https://doi.org/10.7884/teke.3959

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Boston: Pearson.

Tosun, C. (2005). Dil zenginliği, yozlaşma ve Türkçe. *Journal of Language and Linguistic Studies, 1*(2), 136-153.

Ulaş, A. H., & Sevim, O. (2010). İnternet ortamındaki Türkçenin genel durumu. *Ekev Akademi Dergisi, 14*(44), 185-192.

Ünalan, Ş. (2006). *Dil fukaralığı: söz varlığından söz darlığına.* Ankara: Ankara İl Milli Eğitim Müdürlüğü Yayınları.

Yaman, E. (2015). *Türkçe bilinci.* Ankara: Açağ Yayınları.

Yaman, H. (2011). Türkçe bilinci ölçeği: geçerlik ve güvenirlik çalışması. *Türk Eğitim Bilimleri Dergisi, 9*(1), 151-167.

Yurdugül, H. (Eylül, 2005). *Ölçek geliştirme çalışmalarında kapsam geçerlilik indekslerinin kullanımı*. Paper Presented at XIV Ulusal Eğitim Bilimleri Kongresi, Pamukkale Üniversitesi Eğitim Fakültesi, Denizli.

Zengin, E. (2017). Türkçenin diğer dillerle etkileşimi ve sonuçları. *Uluslararası Sosyal Araştırmalar Dergisi, 10*(52), 293-299. https://doi.org/10.17719/jisr.2017.1892

6. APPENDIX

Items of the scale according to sub-dimensions: English-Turkish version

Sub-Factors	Items	Alt Boyutlar	Maddeler
Individual Awareness	1- I think that Turkish words should be derived to place non-Turkish words. 2- The use of our language with foreign word patterns damages our language (**For example; Cafe Sorgun, Otel The Yozgat etc.).** 3- Speaking with only Turkish words and words translated into Turkish is an indication of backwardness. 4- I warn my friends who use foreign words despite having Turkish equivalents while having a conversation. 5- When I come across a foreign word in a text I read, I look up its Turkish equivalent from the dictionary. 6- I think that as individuals, we should speak Turkish properly in our daily lives. 7- Wearing clothes with foreign words on them makes me uncomfortable. 8- It bothers me if a text I read has foreign words used despite having Turkish equivalents	**Bireysel Farkındalık**	1- Türkçeleşmemiş kelimelerin yerine Türkçe kelime türetilmesi gerektiğini düşünüyorum. 2- Dilimizin yabancı kelime kalıpları ile kullanılması dilimize zarar vermektedir (Örneğin; **Cafe Sorgun**, **Otel The Yozgat** vb.). 3- Sadece Türkçe ve Türkçeleşmiş kelimeler kullanarak konuşmak geri kalmışlığın göstergesidir. 4- Sohbet ederken Türkçe karşılığı olduğu halde yabancı kelime kullanan arkadaşlarımı uyarırım. 5- Okuduğum bir metinde yabancı bir kelime ile karşılaştığımda sözlükten Türkçe karşılığını ararım. 6- Birey olarak günlük yaşantımızda güzel bir Türkçe ile konuşmamız gerektiğini düşünüyorum. 7- Üzerinde yabancı kelime yazan giysiler giymek beni rahatsız eder. 8- Okuduğum bir metinde yabancı kelimelerin Türkçe karşılıkları olduğu halde kullanılması beni rahatsız eder.
Social Media Awareness	9- The use of letters that are not in our alphabet (w, q, x) in social media bothers me (For example; wadi instead of vadi etc.). 10- I warn my friends who misspell Turkish words on social media. 11- I approve the use abbreviated words (**For example; mrb instead of merhaba etc.).**	**Sosyal Medya Farkındalığı**	9- Sosyal medyada alfabemizde olmayan (w, q, x) harflerin kullanılması beni rahatsız eder (Örneğin; vadi yerine wadi vb.). 10- Sosyal medyada Türkçe kelimeleri yanlış yazan arkadaşlarımı uyarırım. 11- Kelimelerin kısaltılarak kullanılmasını doğru buluyorum (**Örneğin; merhaba yerine mrb vb**.).
Awareness regarding Daily Life	12- I feel uncomfortable when I see foreign names given to the main roads and streets. 13- I am not bothered by seeing signs written with foreign words around me. 14- I am bothered by seeing workplaces with foreign names around me.	**Günlük Hayata İlişkin Farkındalık**	12- Cadde ve sokaklara yabancı isimler koyulmasından rahatsız olurum. 13- Çevremde yabancı kelimelerle yazılmış tabelalar olmasından rahatsız olmam. 14- Çevremde yabancı isimli iş yerleri görmek beni rahatsız eder.
Awareness in Mass Media	15- It is not important for me whether the language in the mass media is used in accordance with the rules of language. 16- I feel uncomfortable that the Turkish pronunciation of foreign words used in mass media change from person to person in Turkish. 17- Programs with excessive use of local dialects should be expanded.	**Kitle İletişim Araçlarındaki Farkındalık**	15- Kitle iletişim araçlarında dilin kurallarına uygun kullanılıp kullanılmadığı benim için önemli değildir. 16- Kitle iletişim araçlarında kullanılan yabancı kelimelerin Türkçe söyleniş biçimlerinin kişiden kişiye değişmesinden rahatsız olurum. 17- Yerel ağızların aşırı kullanıldığı programlar yaygınlaştırılmalıdır.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 549–569

https://doi.org/10.21449/ijate.725701

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Comparison of Teacher Training Programs in terms of Attitudes towards Teaching Profession and Teacher Self-Efficacy Perceptions: A Meta-Analysis

Ismail Yelpaze [1,*], **Levent Yakar** [2]

[1]Kahramanmaraş Sütçü İmam University, Faculty of Education, Department of Educational Sciences, Turkey

ARTICLE HISTORY

Received: March 19, 2020
Revised: July 3, 2020
Accepted: Sep. 10, 2020

KEYWORDS

Teacher training program,
Teacher self-efficacy perception,
Teaching profession attitudes

Abstract: The aim of this study is to conduct a meta-analysis of studies comparing teacher training programs in terms of attitude towards teaching profession and perception of teacher self-efficacy. For this purpose, the results of the study comparing the faculty of education (FE) and other teacher training programs/faculties were searched and recorded separately for both subjects. A total of 36 studies were recorded in accordance with the criteria, and 27 of these studies were used for the attitude towards teaching profession and 24 for the teacher self-efficacy perceptions. According to the results of the meta-analysis conducted according to the random effect model, teacher candidates in FE have more negative attitude and lower self-efficiency than ones in other teacher training faculties/programs. The difference in both subjects was found to be weak but not statistically significant. The effect size of most common comparison, FE-Pedagogical Formation Certificate Program comparisons in the literature is similar to the general effect. It is concluded that faculties of education whose main purpose is to train teachers do not increase these features of their students sufficiently.

1. INTRODUCTION

The main purpose of education is to prepare human beings for life. In order to achieve this goal in formal education, all stakeholders in education need to work effectively. The most important factor for the education system -which includes students, teachers, administrators, education programs, family, other personnel, buildings, equipment and environment- is the teacher (Kartal, Temelli, & Şahin, 2019), and its influence is higher than other factors (Çapa & Çil, 2000). The teacher, is also the most important player in the creation of qualified manpower for the development of the country, preparing the individual for life and ensuring social peace (Kaya, 2001; Özden, 1999). All citizens of the country are necessarily included in the education system and spend time together with teachers for 12 years period when they are most open to learning and self-development. This shows the important role of the teacher in the training of individuals.

There are a lot of studies investigating that the teacher has an important effect on the success of the student (Canales & Maldonado, 2018; Çelik, Örenoğlu Toraman, & Çelik, 2018). Teachers responsible for the training of qualified individuals are expected to have various qualifications, too. These qualifications can only be gained by a planned education. Therefore, the importance of training qualified teachers is quite clear. The Ministry of National Education

*CONTACT: İsmail YELPAZE ✉ ismailyelpaze@gmail.com Kahramanmaraş Sütçü İmam University, Faculty of Education, Department of Educational Sciences, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

(MoNE, 2017) has identified three main competence of the teaching profession; professional knowledge, professional skill and attitude and values. As a result, teachers are expected to have knowledge and skills related to their field and adopt the values of the profession and gain a positive attitude.

It is known that teacher candidates' attitudes affect their professional success as well as their professional knowledge and skills (Doğan, 2013; İlter, 2009). Attitude is defined as the tendency of the individual to like-dislike an event, situation or object (Kenrick, Neuberg, & Cialdini, 2005) or the individual's emotion-thought-behavioral tendencies towards an object (Kağıtçıbaşı, 2013). Since attitudes include behavioral tendencies towards attitude objects (Sakallı, 2001), a strong and positive attitude can direct the behavior of an individual. Thus, it is seen that attitudes of teachers towards their professions are very determinant in directing their professional behaviors (Özkan, 2012), and a positive attitude provides success and satisfaction in the profession (Recepoğlu, 2013). Teachers who have a positive attitude towards their profession commit with a passion to profession and are more motivated to fulfill the requirements of the profession (Durmuşoğlu, Yanık, & Akkoyunlu, 2009). For this reason, it is important to determine the attitudes of teacher candidates towards the profession in predicting their success and satisfaction in the profession.

Along with attitude towards the profession, self-efficacy perception about the profession is also another factor affecting the quality of the teacher. The concept of self-efficacy, first introduced by Bandura in 1977, is defined as the subjective perception of the individual that he can successfully overcome this challenge (as cited in Senemoğlu, 2012). Since the perception of teaching self-efficacy is a more subjective topic, its definition is defined as the subjective assessment of the teachers that they have the skills to perform the tasks related to the teaching needs specific to their fields (Tschannen-Moran, Woolfolk Hoy, & Hoy, 1998). Self-efficacy perception not only determines the way of thinking, emotions and behavior of individuals but also affects their resilience in the face of difficulties (Bandura, 1997). For this reason, teachers' perceptions of professional efficacy affect their success and professional satisfaction (Karabıyık & Güvenlikaz, 2014).

Although the reasons such as low socioeconomic status and negative individual characteristics of students affect learning negatively (Kartal, Temelli, & Şahin, 2019), teachers with high teaching self-efficacy can turn students' learning in a positive way (Tucker et al., 2005). Because teachers with high teaching self-efficacy are eager to plan and implement their plans, they are open to new thoughts to meet the needs of students, and they try and research new methods (Gülebağlan, 2003). On the other hand, if teacher candidates have low teacher self-efficacy they have more difficulties when they begin the profession (Arastaman, 2013; Brown, Lee, & Collins, 2015). As a result, it seems important that teacher training programs should aim to ensure that teacher candidates not only have academic knowledge but also gain positive attitudes towards the profession and have realistically high teacher self-efficacy.

The majority of teacher training programs in Turkey are located in the education faculties of universities. However, teacher candidates who graduated from the Non-Thesis Master Program in Secondary Education Teaching (NTMP) were able to become a teacher until 2008. Nowadays, those who graduate from faculties other than education faculty can become teachers if they have graduated from the pedagogical formation certificate program (PFCP) that is a follow-up of NTMP (Akdemir, 2013). In addition, students of the faculty of theology (FT) and the faculty of science and literature (FSL) can become teachers if they take pedagogical formation courses. In addition to these programs, graduates from the faculty of technical education (FTE) and teaching programs in physical education and sports school (PESS) can also become teachers. Programs mentioned above will be referred to as teacher training programs hereinafter. Since teacher training programs play an important role in shaping the teacher candidates' beliefs and attitudes towards the profession (Hong & Greene, 2011), it can be thought that the professional attitudes and teacher self-efficacy perceptions of the teacher candidates studying in different programs will also differ.

Since the importance of attitudes towards teaching profession and perceptions of teacher self-efficacy is realized by researchers, there are a lot of studies in the literature comparing teacher self-efficacy perceptions and attitudes towards the profession of teacher candidates studying in different programs. However, these studies investigated different results. For example, in the study of Dadandı, Kalyon, and Yazıcı (2016), teacher candidates in the faculty of education have a more positive attitude towards the profession, while in the study of Bağçeci, Yıldırım, Kara, and Keskinpalta (2015), students at pedagogical formation certificate program have a more positive professional attitude. Similarly, in the study of Yaşar-Ekici (2017), students at pedagogical formation certificate program have higher teacher self-efficacy perception, while in Çetin’s (2017) study, students of education faculty have higher teacher self-efficacy perception. These different results cause confusion and uncertainty. Thus, it is necessary to reveal which group is in favor of attitudes and self-efficacy perceptions. Because singular studies are carried out with limited samples and at limited times, they have limitations to provide comprehensive information about which teacher candidate's professional attitude and self-efficacy perception is more positive. Besides, even if the difference is significant, interpretation of the results only on statistical significance can be misleading (Cohen, 1990) and it is necessary to investigate whether statistical significance represents practical significance (Ellis, 2010).

Meta-analysis is one of the methods to overcome these limitations. Through meta-analysis, studies with different findings can be presented in an effective and holistic way, taking into account the sample sizes and undergoing a systematic evaluation (Lipsey & Wilson, 2001). In fact, there are meta-analysis studies examining the professional attitudes of teachers. When the related meta-analysis studies are examined, it is seen that they are related to comparison of attitude towards teaching profession in terms of gender (Erdamar, Aytaç, Türk, & Arseven, 2016; Tuncer, 2016), comparison of teacher self-efficacy in terms of gender (Çelik, Koç-Erdamar, & Toraman, 2016) and comparison of attitude towards teaching profession in terms of education and other faculties (Atalmış & Köse, 2018). When these studies are analyzed, it is seen that there is no detailed comparison of teacher self-efficacy and attitude towards the teaching profession in terms of each teacher training program. With this study, a step is taken to fill these gaps in the literature. By this means, it is expected that the attitudes towards teaching profession and perceptions of teacher self-efficacy are in favor of which group is more reliable, so there will be some light shed on this uncertainty. Last but not least, teacher training programs will be able to see their own levels of teacher candidates' perception of self-efficacy and attitude towards the teaching. Also, they can benefit from the findings of this study in assessing their educational outcomes.

Consequently, the aim of this study is to conduct meta-analysis of studies comparing teacher training programs in terms of teacher candidates’ attitude towards teaching profession and perception of teacher self-efficacy.

2. METHOD

This study aims to present the findings in a holistic way by bringing together the relevant studies in the field. In order to achieve this aim, the research was carried out by meta-analysis. Meta-analysis, which means collecting analyzes, is a method based on achieving a general result by combining the results obtained from different studies (Dinçer, 2014). From another perspective, meta-analysis can be seen as a literature review based on quantitative data. Accordingly, meta-analysis can be expressed as the statistical analysis of the results of numerous analyzes obtained from individual studies in order to integrate the findings (Glass, 1976).

2.1. Data Collection Procedure

Since the study focused on two main issues related to teaching such as attitude towards teaching profession and teacher self-efficacy perception, the data were collected in each subject in separate processes. The processing steps for each subject are as follows; Keywords of "öğretmenlik mesleğine yönelik tutum karşılaştırma" (comparison of attitude towards teaching profession), "öğretmenlik tutum karşılaştırma" (comparison of attitude of teaching), "öğretmen öz-yeterlik algısı karşılaştırma" (comparison of teacher self-efficacy perception), "öğretmen öz-yeterlik karşılaştırma" (comparison of teacher self-efficacy), "attitude towards teaching comparison" and "teacher self-efficacy comparison" were scanned in databases of Google Scholar, Turkish Academic Network And Information Center (Ulakbim) and Turkish Council of Higher Education (YÖK) National thesis center, in January 2018 - August 2019, two times. Among the reached published and unpublished studies which meet the following criteria were included in the meta-analysis process.

1. The study should be related to teacher candidates' attitude towards teaching profession and / or perception of teacher self-efficacy. (Studies with general self-efficacy perception level were not taken into consideration.)

2. The study must include data from at least one of the relevant subjects of pre-service teachers who graduated or studying at the Faculty of Education and pre-service teachers who graduated or studying at the other departments

3. The study must be conducted on pre-service teachers. (Studies on teachers, appointed candidate teachers or undergraduate students or alumni who are not pre-service teachers were not taken into consideration.)

4. Study must present sample sizes, averages and standard deviations/ t score or U values and sample sizes values that can be transformed to effect size.

As a result of the surveys, a total of 40 studies which meet the above mentioned criteria were reached. As the U-values were given only for the subscales in one of the studies, and only the item analyzes were given in two of them, they were excluded from the final study since they could not be used in the meta-analysis when the re-detailed examinations. And, one study was excluded from the final study list due to lack of validity and reliability evidence of the used scale and analysis was conducted with 36 studies.

Data were recorded independently by both researchers. After the data were recorded, cross-checks were made and agreed data was decided to take final study list. The name of the study, the name of the researchers, the year of publication, the type of publication, the publisher, the name of the programs being compared, as well as the corresponding scale mean scores, sample sizes and standard deviations or t values of the pre-service teachers in each program were recorded. For some studies, sample sizes and U values were recorded. The number and sample size of the studies according to the subject, type of publication and comparison units are presented in Table 1.

In Table 1, the number of the studies which were performed in the meta-analysis and general characteristics of studies are presented. A total of 24 (17 + 7) studies were about attitudes towards teaching profession and 19 (12 + 7) studies were about teacher self-efficacy perception. In one of these studies, there were 4 FE / Other comparisons that contain research results on both subjects. There were 2 FE / Other comparisons in 2 studies which were about teacher self-efficacy study. Thus, 27 comparisons were made for attitudes towards the teaching profession and 24 comparisons in teacher self-efficacy perception were subjected to meta-analysis.

Table 1. *The number of studies involved in Meta-analysis*

Subject	Comparison Program	Article	Thesis	Book Section	Total
Attitude towards Teaching Profession	PFCP (a)	8(2679)	1(1116)		9(3795)
	FSL (b)	3(403)			3(403)
	FT (c)	1(273)			1(273)
	PESS (d)	2(523)			2(523)
	NTMP	2(513)			2(513)
	Total	16(4391)	1(1116)		17(5507)
Teacher Self-Efficacy Perception	PFCP	6(2035)			6(2035)
	FSL	1(338)			1(338)
	PESS	1(411)			1(411)
	FTE	1(495)			1(495)
	NTMP		1(496)		1(496)
	PFCP&FSL	1(407)	1(854)		2(1261)
	Total	10(3686)	2(1350)		12(5036)
Both	PFCP	3(1084)		1(452)	4(1536)
	PESS	1(411)			1(411)
	(a), (b), (c), (d)	1(786)			1(786)
	FTE	1(250)			1(250)
	Total	6(2531)		1(452)	7(2983)
Total		32(10608)	3(2466)	1(452)	36(13526)

PFCP (a): Pedagogical Formation Certificate Program, FSL (b): Faculty of Science&Letter, FT (c): Faculty of Theology, PESS (d): Physical Education and Sports Scholl, NTMP: Non-Thesis Master Program, FTE: Faculty of Technical Education

Sample sizes of the studies used in meta-analysis are given in brackets in Table 1. When the sample sizes are examined, 8490 (5507 + 2983) pre-service teachers constitute the total sample in attitudes towards teaching profession studies, while 8019 (5036 + 2983) pre-service teachers constitute the total sample in the teacher self-efficacy studies.

2.2. Data Analysis

The data were analyzed by meta-analysis. Comprehensive Meta-Analysis Version 3 (CMA) software designed for meta-analysis was used for data analysis. In order to analyze the data, some processes were performed before analysis. In the literature search, some studies reported U value in comparison findings. Since the U value cannot be used directly in the used software, the U value was converted to Cohen's effect size d (Lenhard & Lenhard, 2016). Another process performed before the data analysis was to combine the subscale findings. There are many scales for attitudes towards teacher profession and perception of teacher self-efficacy used in the studies and these scales contain different names and number of dimensions. In order to bring the studies together, only the scale total scores were taken into consideration. In the studies which did not include the scale total score, the subscale values were combined and statistics related to the scale total score were calculated. In this study, a meta-analysis was performed for each of the two subjects.

The most important statistic is effect size in meta-analysis (Dinçer, 2014). It is stated that the effect size must be reported together with the p value that reveals the difference of the effect size statistics from zero (Sullivan & Feinn, 2012) and even the effect size is more important than the p value (Borenstein, Hedges, Higgins, & Rothstein, 2009). For this reason, in the expression and interpretation of the findings in the study, firstly the effect size and then the p value were taken into consideration.

There are different effect sizes such as Cohen's d, Hedge's g or Glass's Δ for the overall effect that results from meta-analysis. In some cases, these statistics may have superiorities to another. It is stated that Cohen's d statistic provides accurate result if the number of samples in the studies is over 20. (Lipsey & Wilson, 2001) Cohen's d statistic was used in this study because the samples of all the studies included were above 20. Cohen's effect size ranges and their meanings are given in Table 2 (Cohen, Manion, & Morrison, 2007).

Table 2. *Cohen's Effect Size Ranges and Their Meanings*

Cohen's d	Meaning
0-0.20	Weak Effect
0.21-0.50	Modest Effect
0.50-1.00	Moderate Effect
>1.00	Strong Effect

Another case that should be decided before presenting the meta-analysis findings is which of the fixed effects and random effects models will be used in the calculation of the overall effect. The fixed effect model is based on the assumption that all the studies analyzed have the same effect and the difference between the results of studies is due to the sampling error in the studies. The biggest difference of the random effect model from the fixed effect model is that it is based on the assumption that the studies may have different effects (Üstün & Eryılmaz, 2014). Providing the source of variance correctly will help in choosing the right model. Considering the years of the studies, the university in which they were conducted, and the units which comparison was made, the use of random effects method was found to be a more appropriate option in the analysis for overall effect calculation. The results of the heterogeneity test were also taken into consideration in the model selection (Yıldırım, Çırak-Kurt, & Şen, 2019). If the Q statistic is greater than the chi-square value of p = 0.05 at the relevant degree of freedom, it indicates heterogeneity. Another statistic in this area is I^2, when I^2 value is 25, 50, 75; it indicates low, medium and high heterogeneity, respectively (Higgins, Thompson, Deeks, & Altman, 2003). In the analyzes in which the effect of the sub-groups was analyzed via sub-group variable, random effect model was used between sub-groups and a fixed effect model was used within group. This method is called a mixed effect model (Borenstein, Hedges, Higgins, & Rothstein, 2013).

Another statistic that comes to the fore in meta-analysis is publication bias. Publication bias can be expressed as the literature bias due to the fact that the probability of publication of studies that are not statistically significant or have low effect size is lower than the probability of publication of studies with significant differences or have large effect size (Borenstein et al., 2013). There are several reasons of publication bias that should be considered in meta-analysis. In this study, detailed scans were carried out to prevent bias that may occur in the literature review, and the studies to be analyzed were selected with consensus. In addition, unpublished master's thesis and doctoral dissertation were also included to minimize the effect of publication bias on the results of meta-analysis. Publication bias analysis was carried out with funnel plot, Duval and Tweedie's (2000a; 2000b) trim and fill method and Egger's linear regression methods (Egger, Davey Smith, Schneider, & Minder, 1997).

In order to investigate the possible publication bias in the literature, the funnel graph in which individual studies represented as a point, was examined. The points in the funnel graph are shown at the intersection of the horizontal plane corresponding to the effect size of the individual study and the vertical plane corresponding to its standard error. The funnel graph has a vertical line extending from the overall effect size calculated in the analysis. The fact that the points representing the individual studies have a symmetrical appearance around this vertical line contains an opinion that there is no publication bias.

Another statistical method used in publication bias is the trim and fill method of Duval and Tweedie (2000a, 2000b). In this method, the effect size is calculated again by subtracting the point representing the individual study located far from the funnel graph. Trim and fill method is a repetitive process based on the funnel graph becoming symmetrical around the new effect size. Subtraction also reduces the variance while regulating the effect size. In order to prevent this situation, the studies are re-added to the analysis and a mirror image is added to the funnel for these studies (Bakioğlu & Göktaş, 2018). The small difference between the original effect size and the effect size obtained with the trim and fill method of Duval and Tweedie indicates that there is no publication bias.

From the point of view of revealing evidence of bias, the funnel plot method is based entirely, and the trim and fill method of Duval and Tweedie is partially based on visual evidence. Therefore, it can produce subjective results. To overcome this limitation, Egger's linear regression method, which examines statistical bias, was also used. This method is based on the model that contains the regression of the standard normal deviation of the studies against the certainty of this value. For a symmetrical funnel plot according to the model, the regression line obtained from the studies is expected to extend linearly through the origin. The non-significance of the *p* value obtained from the method (> 0.05) indicates that the studies are linearly aligned and there is no bias (Egger, Davey Smith, Schneider, & Minder, 1997).

In the literature search, it was seen that there were six different teacher training programs compared with FE. The fact that the results of the meta-analysis are generally considered as FE / Other comparison without considering these differences will not be sufficient to elaborate the results. In order to fully understand the direction and strength of the possible differences in different program comparisons, each teacher training program was used as a sub-groups variable. Thus, a process for explaining a source of variance between studies was performed (Borenstein, Hedges, Higgins, & Rothstein, 2010). Due to the low number of studies in some subgroups, the number of individuals included in the studies was low, and the effect size of these subgroups caused the standard error to be high. Therefore, findings accompanied by standard errors are discussed.

3. FINDINGS

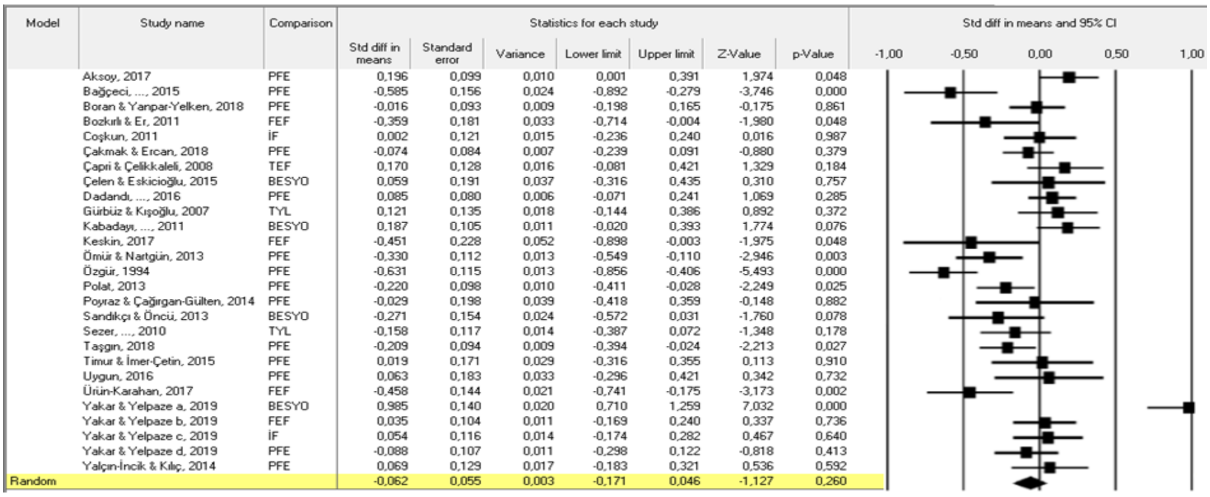
In this section, first of all, the findings of the meta-analysis for attitude towards teaching profession will be given. The overall effect size and heterogeneity values of all studies related to attitude towards teaching profession included in the research were calculated. The results are presented in Table 3.

According to the fixed effect model given in Table 3, Q statistics were found as 143.243 from the heterogeneity results obtained. Since this value is higher than 38.885 Q value in 26 degrees of freedom in the Chi-square table, it can be said that the studies contain heterogeneity. When the I^2 statistics are examined, high heterogeneity is observed. When these statistics are taken into consideration, it can be said that the studies are highly heterogeneous. In addition, since the distribution of the effect sizes in Figure 1 shows the difference between the studies, it was decided that the random effect model is suitable for the attitude towards to teaching profession.

Table 3. *Overall effect sizes and heterogeneity results for attitude towards teaching profession*

Model	N	ES	SE	95% CI.		Z	*p*	df	*Q*	I^2
				Lower	Upper					
Fixed	27	-0.051	0.023	-0.096	-0.006	-2.221	0.026	26	143.243	81.84
Random	27	-0.062	0.055	-0.171	0.046	-1.127	0.260			

The overall effect size of the random effect model for the attitude towards teaching profession is -0.062. This value appears to indicate weak effect. Since value is close to 0, it can be said that the overall effect against FE is negligible in FE/Other comparisons. When the significance of the effect size, which is of secondary importance after the effect size, was considered, the effect was not significant (p> 0.05). The forest graph showing the effect size of individual study and its weight in all studies is presented in Figure 1.



Model	Study name	Comparison	Statistics for each study						
			Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
	Aksoy, 2017	PFE	0,196	0,099	0,010	0,001	0,391	1,974	0,048
	Bağçeci, ..., 2015	PFE	-0,585	0,156	0,024	-0,892	-0,279	-3,746	0,000
	Boran & Yanpar-Yelken, 2018	PFE	-0,016	0,093	0,009	-0,198	0,165	-0,175	0,861
	Bozkırlı & Er, 2011	FEF	-0,359	0,181	0,033	-0,714	-0,004	-1,980	0,048
	Coşkun, 2011	İF	0,002	0,121	0,015	-0,236	0,240	0,016	0,987
	Çakmak & Ercan, 2018	PFE	-0,074	0,084	0,007	-0,239	0,091	-0,880	0,379
	Çapri & Çelikkaleli, 2008	TEF	0,170	0,128	0,016	-0,081	0,421	1,329	0,184
	Çelen & Eskicioğlu, 2015	BESYO	0,059	0,191	0,037	-0,316	0,435	0,310	0,757
	Dadandı, ..., 2016	PFE	0,085	0,080	0,006	-0,071	0,241	1,069	0,285
	Gürbüz & Kışoğlu, 2007	TYL	0,121	0,135	0,018	-0,144	0,386	0,892	0,372
	Kabadayı, ..., 2011	BESYO	0,187	0,105	0,011	-0,020	0,393	1,774	0,076
	Keskin, 2017	FEF	-0,451	0,228	0,052	-0,898	-0,003	-1,975	0,048
	Ömür & Nartgün, 2013	PFE	-0,330	0,112	0,013	-0,549	-0,110	-2,946	0,003
	Özgür, 1994	PFE	-0,631	0,115	0,013	-0,856	-0,406	-5,493	0,000
	Polat, 2013	PFE	-0,220	0,098	0,010	-0,411	-0,028	-2,249	0,025
	Poyraz & Çağırgan-Gülten, 2014	PFE	-0,029	0,198	0,039	-0,418	0,359	-0,148	0,882
	Sandıkçı & Öncü, 2013	BESYO	-0,271	0,154	0,024	-0,572	0,031	-1,760	0,078
	Sezer, ..., 2010	TYL	-0,158	0,117	0,014	-0,387	0,072	-1,348	0,178
	Taşgın, 2018	PFE	-0,209	0,094	0,009	-0,394	-0,024	-2,213	0,027
	Timur & İmer-Çetin, 2015	PFE	0,019	0,171	0,029	-0,316	0,355	0,113	0,910
	Uygun, 2016	PFE	0,063	0,183	0,033	-0,296	0,421	0,342	0,732
	Ürün-Karahan, 2017	FEF	-0,458	0,144	0,021	-0,741	-0,175	-3,173	0,002
	Yakar & Yelpaze a, 2019	BESYO	0,985	0,140	0,020	0,710	1,259	7,032	0,000
	Yakar & Yelpaze b, 2019	FEF	0,035	0,104	0,011	-0,169	0,240	0,337	0,736
	Yakar & Yelpaze c, 2019	İF	0,054	0,116	0,014	-0,174	0,282	0,467	0,640
	Yakar & Yelpaze d, 2019	PFE	-0,088	0,107	0,011	-0,298	0,122	-0,818	0,413
	Yalçın-İncik & Kılıç, 2014	PFE	0,069	0,129	0,017	-0,183	0,321	0,536	0,592
Random			-0,062	0,055	0,003	-0,171	0,046	-1,127	0,260

Figure 1. *The forest graph showing the effect size of individual study about attitude towards teaching profession*

The squares on the right in Figure 1 represent the effect size of the individual study and the line adjacent to the square represents the upper and lower limits of the effect size in the 95% confidence interval. The weight of the individual work in the overall effect size is represented by the area of the square concerned. The rhombus under the graph shows the overall effect size of the studies.

When the values of the individual studies on the left side of Figure 1 are examined, the effect sizes of the studies vary between -0.631 and 0.985. Positive values indicate the effect size in favor of FE. Accordingly, the effect size of 27 studies is 13 in favor of FE and 14 in against FE. Subgroups effect sizes for attitude towards the teaching profession for each program compared with FE in order to investigate the source of the variance seen between studies are given in Table 4.

Table 4. *Attitudes towards teaching profession in terms of compared programs*

Program	N	ES	SE	95% CI		Z	*p*	df	*Q*	*p*
				Lower	Upper					
PESS	4	0.244	0.262	-0.269	0.758	0.93	0.351			
FSL	4	-0.282	0.145	-0.566	0.002	-1.95	0.052			
FT	2	0.029	0.084	-0.135	0.194	0.35	0.727			
PFCP	14	-0.125	0.063	-0.248	-0.001	-1.98	0.048			
FTE	1	0.170	0.128	-0.081	0.421	1.33	0.184			
NTMP	2	-0.027	0.139	-0.299	0.246	-0.19	0.847			
Total	27	-0.049	0.042	-0.132	0.033			5	9.1	0.105

Analysis results which were analyzed according to the mixed effect model shown in Table 4; it is seen that the effect sizes calculated for 6 different programs vary between -0.282 and 0.244. Three of these effect sizes are in favor of FE and 3 of them are against. In FE / PFCP comparison, the effect size was found to be weak and statistically significant in favor of PFCP. The effect sizes seen in other comparisons were not statistically significant. In the comparison program which is the mediator variable, FE / FSL comparison with the highest effect size was also against the FE and the effect size was low and very close to the statistical significance level. When the FE / PESS comparisons are examined, it was found that the pre-service teachers who are trained in FE have a more positive attitude towards teaching profession, the effect size of this difference is low, but it is not statistically significant.

When the effect of using different programs as sub-group variable on the heterogeneity it is found that Q value does not reach the value of 11.07 which is X^2 value of 5 degrees of freedom. Therefore, the comparison of different programs did not make a significant contribution to the variance (p> 0.05).

Figure 2a shows the funnel graph, while 2b shows funnel graph for trim and fill method for revealing publication bias in the attitudes towards teaching profession studies included in the meta-analysis.

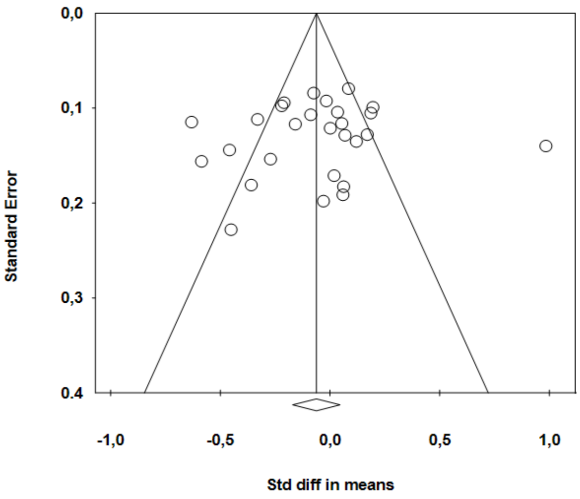


Figure 2a. *Attitude towards teaching profession funnel graph*

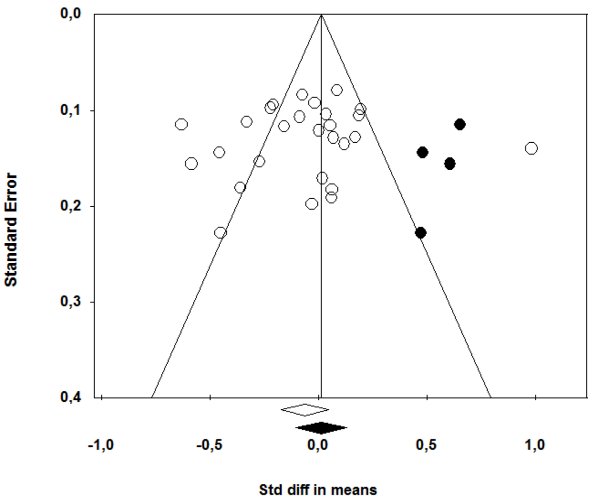


Figure 2b. *Attitude towards teaching profession funnel graph for trim and fill method*

In Figure 2a, the effect sizes of the studies about the attitude towards the teaching profession and their standard errors are represented by circles. It is seen that the expected symmetrical distribution is partially achieved. It is seen that the effect sizes of the studies which favor of FE are smaller and closer to each other compared to the studies against FE, which prevents the appearance of full symmetry. For this reason, Figure 2a cannot provide complete information about publication bias.

The black dots in Figure 2b represent the publications that must be added to achieve full symmetry according to the trim and fill method of Duval and Tweedie (2000a, 2000b). The black equilateral quadrangle shows the overall effect size that will occur with the trim and fill method. When Figure 2b is analyzed, if 4 publications, in favor of FE, are added to the analysis for corresponding coordinates the funnel plot will be fully symmetrical. The only number to be added to make the 27 studies fully symmetrical is only 4, which is an indication that the current situation is close to symmetrical distribution. If these 4 studies are added, the overall effect will increase from -0.062 to 0.010. The fact that the difference is small indicates that there is no publication bias in the studies. Egger's regression method (Egger, Davey Smith, Schneider, & Minder, 1997), which is another method used in publication bias test, has *t* value of 0.49 and *p* value of 0.62. A statistical significance value of *p*> 0.05 indicates that there is no bias.

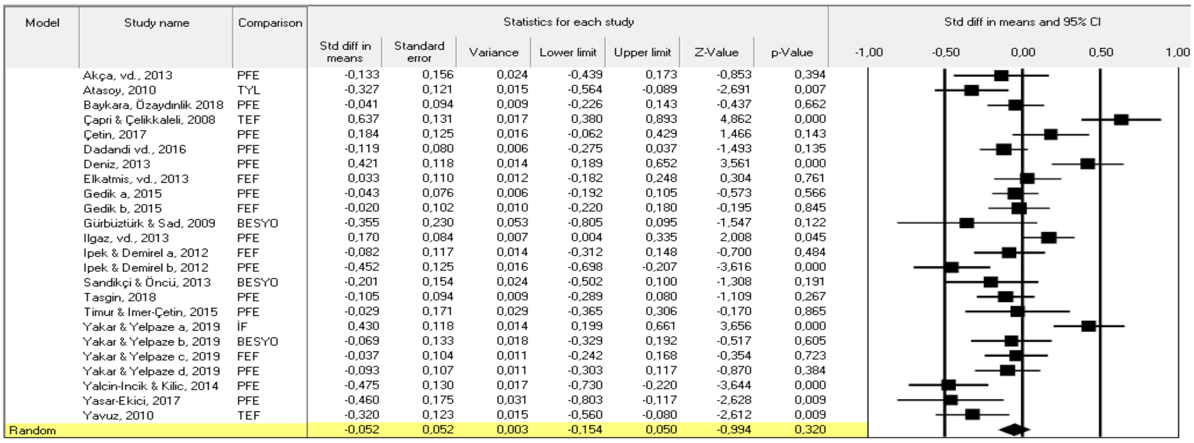
The findings related to the meta-analysis study for teacher self-efficacy perception are given below. Firstly, homogeneity analysis was made to decide which method to use in calculating the effect size. Findings related to this analysis and the overall effect size calculations are presented in Table 5.

Table 5. *Overall effect size and heterogeneity results for teacher self-efficacy perception*

Model	N	ES	SE	95% CI		Z	*p*	df	*Q*	I^2
				Lower	Upper					
Fixed	4	-0.035	0.023	-0.081	0.010	-1.535	0.125	23	111.443	79.36
Random	4	-0.052	0.052	-0.154	0.050	-0.994	0.320			

According to the fixed effect model given in Table 3, *Q* statistics were found as 111.443 from the heterogeneity results obtained. Since this value is higher than 35.172 *Q* value in 23 degrees of freedom in the Chi-square table, it can be said that the studies contain heterogeneity. When the I^2 statistics are examined, high heterogeneity is observed. When these statistics are taken into consideration, it can be said that the studies are highly heterogeneous. In addition, since the distribution of the effect sizes in Figure 3 shows the difference between the studies, it was decided that the random effect model is suitable for the teacher self-efficacy perception.

The overall effect size of the random effect model for teacher self-efficacy perception is -0.052. This value appears to indicate weak effect. Since value is close to 0, it can be said that the overall effect against FE is negligible in FE / Other comparisons. When the significance of the effect size, which is of secondary importance after the effect size, was considered, the effect was not significant (*p*> 0.05). The forest graph showing the effect size of individual study and its weight in all studies is presented in Figure 3.



Model	Study name	Comparison	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
			Statistics for each study						
	Akça, vd., 2013	PFE	-0,133	0,156	0,024	-0,439	0,173	-0,853	0,394
	Atasoy, 2010	TYL	-0,327	0,121	0,015	-0,564	-0,089	-2,691	0,007
	Baykara, Özaydınlık 2018	PFE	-0,041	0,094	0,009	-0,226	0,143	-0,437	0,662
	Çapri & Çelikkaleli, 2008	TEF	0,637	0,131	0,017	0,380	0,893	4,862	0,000
	Çetin, 2017	PFE	0,184	0,125	0,016	-0,062	0,429	1,466	0,143
	Dadandi vd., 2016	PFE	-0,119	0,080	0,006	-0,275	0,037	-1,493	0,135
	Deniz, 2013	PFE	0,421	0,118	0,014	0,189	0,652	3,561	0,000
	Elkatmis, vd., 2013	FEF	0,033	0,110	0,012	-0,182	0,248	0,304	0,761
	Gedik a, 2015	PFE	-0,043	0,076	0,006	-0,192	0,105	-0,573	0,566
	Gedik b, 2015	FEF	-0,020	0,102	0,010	-0,220	0,180	-0,195	0,845
	Gürbüztürk & Sad, 2009	BESYO	-0,355	0,230	0,053	-0,805	0,095	-1,547	0,122
	Ilgaz, vd., 2013	PFE	0,170	0,084	0,007	0,004	0,335	2,008	0,045
	Ipek & Demirel a, 2012	FEF	-0,082	0,117	0,014	-0,312	0,148	-0,700	0,484
	Ipek & Demirel b, 2012	PFE	-0,452	0,125	0,016	-0,698	-0,207	-3,616	0,000
	Sandikçi & Öncü, 2013	BESYO	-0,201	0,154	0,024	-0,502	0,100	-1,308	0,191
	Tasgin, 2018	PFE	-0,105	0,094	0,009	-0,289	0,080	-1,109	0,267
	Timur & Imer-Çetin, 2015	PFE	-0,029	0,171	0,029	-0,365	0,306	-0,170	0,865
	Yakar & Yelpaze a, 2019	İF	0,430	0,118	0,014	0,199	0,661	3,656	0,000
	Yakar & Yelpaze b, 2019	BESYO	-0,069	0,133	0,018	-0,329	0,192	-0,517	0,605
	Yakar & Yelpaze c, 2019	FEF	-0,037	0,104	0,011	-0,242	0,168	-0,354	0,723
	Yakar & Yelpaze d, 2019	PFE	-0,093	0,107	0,011	-0,303	0,117	-0,870	0,384
	Yalcin-Incik & Kilic, 2014	PFE	-0,475	0,130	0,017	-0,730	-0,220	-3,644	0,000
	Yasar-Ekici, 2017	PFE	-0,460	0,175	0,031	-0,803	-0,117	-2,628	0,009
	Yavuz, 2010	TEF	-0,320	0,123	0,015	-0,560	-0,080	-2,612	0,009
Random			-0,052	0,052	0,003	-0,154	0,050	-0,994	0,320

Figure 3. *The forest graph showing the effect size of individual study about teacher self-efficacy*

As it can be seen in Figure 3, the effect sizes of the studies vary between -0.475 and 0.637. Positive values indicate effect sizes in favor of FE. Accordingly, when the effect sizes of 24 studies are analyzed separately, it is seen that there are results in favor of FE in 6 studies and against FE in 18 studies.

Subgroups effect sizes for teacher self-efficacy perception for each program compared with FE in order to investigate the source of the variance seen between studies are given in Table 6.

Table 6. *Effect sizes of teacher self-efficacy perception*

Program	N	ES	SE	95% CI		Z	*p*	df	Q	*p*
				Lower	Upper					
PESS	3	-0.162	0.092	-0.343	0.018	-1.763	0.078			
FSL	4	-0.025	0.054	-0.130	0.081	-0.458	0.647			
FT	1	0.430	0.118	0.199	0.661	3.656	0.000			
PFCP	13	-0.079	0.065	-0.206	0.048	-1.224	0.221			
FTE	2	0.157	0.478	-0.780	1.095	0.329	0.742			
NTMP	1	-0.327	0.121	-0.564	-0.089	-2.691	0.007			
Total	24	-0.044	0.034	-0.111	0.024			5	23.9	0.00

When the analysis results made according to the mixed effect model in Table 6 are examined, the effect sizes calculated for six different programs vary between -0.327 and 0.430. Two of the effect sizes appear to be in favor of FE, while four appear against FE. The FE / PFCP comparison, with the greatest number of studies, showed that the effect size was weak in favor of PFCP and not statistically significant. All of the effect sizes of programs which have more than individual study were found to be weak and were not found statistically significant. It was seen that there was only one study in the FE / FT and FE / NTMP comparisons, and they have highest effect size. Compared to pre-service teachers who are trained in FE, while pre-service teachers who are trained in FT consider themselves inadequate, pre-service teachers who are trained in NTMP education consider themselves more sufficient in terms of professional competence. However, since there was only one study in these subgroups, they were found to have slightly higher standard errors except FTE comparison.

When the effect of using different programs as subgroup variable on the heterogeneity is seen Q value exceeds the value of 11.07 which is X^2 value of 5 degrees of freedom. Therefore, the comparison of different programs made a significant contribution to the variance (p< 0.01).

Figure 4a shows the funnel graph, 4b shows funnel graph for trim and fill method for revealing publication bias in the teacher self-efficacy perception studies included in the meta-analysis.

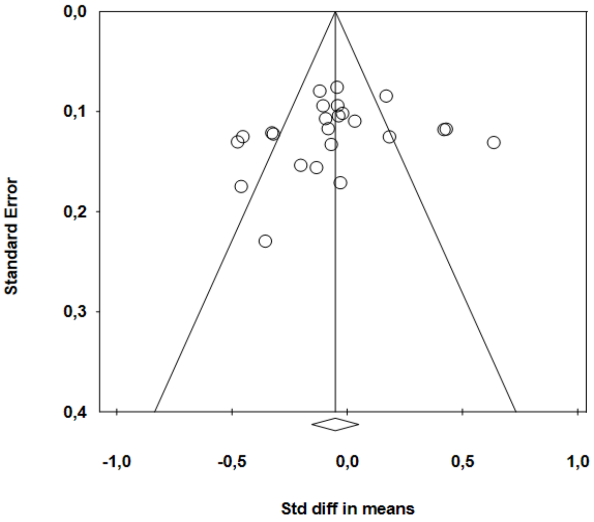


Figure 4a. *Teacher self-efficacy perception funnel graph*

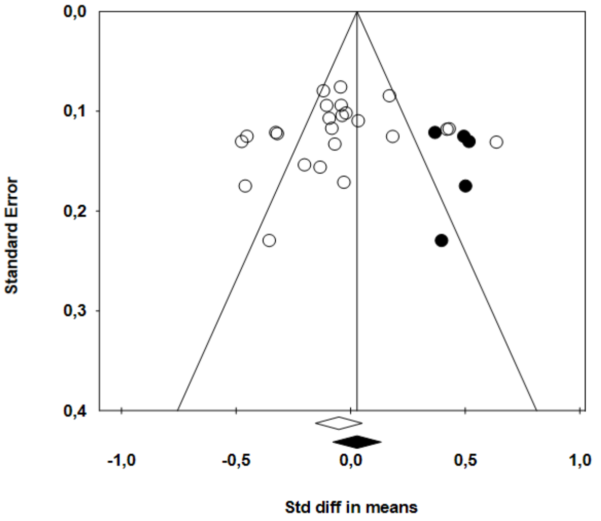


Figure 4b. *Teacher self-efficacy perception funnel graph for trim and fill method*

In Figure 4a, the effect sizes of the studies about teacher self-efficacy perception and their standard errors are represented by circles. It is seen that the expected symmetrical distribution is partially achieved. It is seen that the effect sizes of the studies in favor of FE are smaller and closer to each other compared to the studies against FE, which prevents the appearance of full symmetry. For this reason, Figure 4a cannot provide complete information about publication bias.

The black dots in Figure 2b represent the publications that must be added to achieve full symmetry according to the trim and fill method of Duval and Tweedie (2000a, 2000b). The black equilateral quadrangle shows the overall effect size that will occur with the trim and fill method. When Figure 4b is analyzed, if 5 publications, in favor of FE, are added to the analysis for corresponding coordinates the funnel plot will be fully symmetrical. The only number to be added to make the 24 studies fully symmetrical is only 5, which is an indication that the current situation is close to symmetrical distribution. If these 5 studies are added, the overall effect will increase from -0.052 to 0.031. The fact that the difference is small indicates that there is no publication bias in the studies. Egger's regression method (Egger, Davey Smith, Schneider, & Minder, 1997), which is another method used in publication bias test, has *t* value of 0.83 and *p* value of 0.41. A statistical significance value of *p*> 0.05 indicates that there is no bias.

4. DISCUSSION and CONCLUSION

The aim of this study was to examine the singular study findings which comparing teacher candidates' attitudes towards teaching profession and teacher self-efficacy perceptions in terms of teacher training programs by using meta-analysis method. In accordance with this purpose, firstly, publication bias of studies, then the effect sizes related to attitudes towards the teaching profession and teacher self-efficacy perception in terms of teacher training programs are discussed.

Firstly, detailed scans were carried out to prevent bias that may occur in the literature review and the studies to be analyzed were selected with consensus. In the examination of the funnel graph method, there was no clear conclusion about the publication bias regarding both the attitude towards the teaching profession and the teacher self-efficacy perception studies. It was concluded that there was no publication bias for the studies according to trim and fill method of Duval and Tweedie, and the linear regression methods of Egger.

When FE students are compared with the students in other programs, there are modest effect size findings that show that FE students have a more positive attitude than students in various programs and have a more negative attitude than others. When analyzed as a whole, it was found that FE students have a more negative attitude compared to other program students, the effect size is weak and not statistically significant. Similarly, in the meta-analysis study conducted by Atalmış and Köse (2018), it is stated that the attitudes of FE students do not differ significantly from the attitudes of PFCP and FSL students. Considering the above findings, it is seen that the attitudes of FE students towards the profession are not different from the students in the other programs, because those who have a more positive attitude and those with a more negative attitude balance each other. Although FE students' negative attitudes are not significantly different from other students' attitudes, it needs to be investigating deeply, since FE students have more negative attitudes towards the teaching profession, even though they choose their teaching programs to become teachers. It also points out that the adequacy of the selection application of students for FE teacher training programs and the content of the curriculum should be questioned. As a matter of fact, it is stated in some studies that FE freshman students have a more positive attitude than senior students (Çakmak & Ercan, 2018). These findings indicate that FE teacher training programs may be inadequate in making students love the profession and gain a positive attitude.

When FE students' attitudes towards teaching profession are compared for each program, it is seen that the attitudes of faculty of education (FE) students are more positive than the students in physical education and sports teaching (PESS), faculty of theology (FT) and faculty of technical education (FTE). Considering Cohen's (Cohen, Manion, & Morrison, 2007) effect size classification, it was found that the effect size of the difference was low for PESS and weak for FT and FTE, but all three effect sizes were not statistically significant. As with FE teaching programs, the purpose of PESS and FTE programs is to train teachers. Therefore, it is thought that PESS and FTE students prefer these programs to become teachers. Likely, the responsibility of training teachers for religious culture and moral knowledge teaching, and vocational courses in religious vocational high schools was given to FT. For this reason, FT students may also have preferred FT programs in order to become teachers. As a result, students in four programs may have similar attitude levels since they prefer their programs with the awareness of being and desire to become a teacher.

Several factors may have been influential in the attitude of PESS, FTE and FT students being more negative than FE students. For example, theology faculty students have other options besides being teacher, such as religious staff, teaching Quran (Korukçu, 2011). Similarly, PESS students work in various sports related professions other than teaching profession (Şaşmaz Ataçocuğu & Zelyurt, 2017). Therefore, being a teacher is not seen as an indispensable option for these students. Lastly, negative attitudes of teacher candidates in FTE may be due to their low appointment probability (Çapri & Çelikkaleli, 2008).

As a result of comparing the attitudes, it was found that attitudes towards the profession of teacher candidates at the Faculty of Science and Letters (FSL), Pedagogical formation education program (PFCP) and non-thesis master's degree program (NTMP) were more positive than the FE students. Considering the effect size classification of Cohen (Cohen, Manion, & Morrison, 2007), the effect size was found to be meaningful only for PFCP students, and it has weak effect size. PFCP certificate program students consist of individuals who have graduated from undergraduate programs that do not have the aim to train teachers. Participants of studies took pedagogical formation lessons in order to be a teacher during the research studies period. This means that they decided to become teachers because they would have difficulties in finding a job related to their own graduation areas or they thought they made the wrong choice and they turned towards this path. As a matter of fact, it is stated that among the reasons for PFCP application, ease of finding a job and love of profession are the two most prominent reasons (Kiraz & Dursun, 2015). PFCP students may be making more informed decisions because they are graduated and their age is getting older. In this regard, since they may be more willing and determined to be a teacher than FE students at the undergraduate level, they may have adopted a more positive attitude towards the teaching profession.

The teacher self-efficacy perceptions of FE students are more positive than some programs (FTE) and more negative than some programs (PFCP) and effect size is modest. When FE students were compared with all other programs as a whole, it was found that FE students' teacher self-efficacy perceptions were more negative, but the effect size of this difference was weak and statistically insignificant due to the different results in the opposite direction balancing each other. Since this finding will cause many data to be lost, a comparison was made on the basis of programs.

When FE and other programs were compared separately, it was found that teacher self-efficacy perceptions of FT and FTE students were lower than FE students. Considering the effect size classification of Cohen (Cohen, Manion, & Morrison, 2007), the effect size of the difference seen between FE and FT was modest and significant, while the effect size of the difference between FE and FTE was weak and not significant.

Considering the related research sample shows that FT students have not taken the teaching practice course yet. They may feel inadequate for an unknown job, as they have not yet experience with the requirements of the profession. In addition, since the main purpose of FE teaching programs is to train teachers, it is stated that it is more likely to gain professional competence than FT students who aim to train theologians (Coşkun, 2011). In addition, many field courses at FE are carried out by associating them with the teaching profession. For example, the content of the community service course can be mostly student oriented. In short, FE instructors may have motivated their students to feel more proficient as they will have a better command of teaching profession lessons. Lastly, the fact that there is only one study in the FT sample included in the meta-analysis within the scope of this study requires a more cautious evaluation of the results.

On the other hand, self-efficacy perceptions of PESS, FSL, PFCP and NTMP students were found to be higher than FE students. Considering the effect size classification of Cohen (Cohen, Manion, & Morrison, 2007), the effect size of the difference is significant only with NTMP and it was found to be modest.

When the findings are examined, there is only one study investigating NTMP students with higher self-efficacy perception than FE students, and it is noteworthy that this study was conducted with candidates of music teachers. For this reason, these features should be taken into consideration while generalizing. Despite this, various explanations can be made about why NTMP students feel more adequate. Firstly, since NTMP students are graduated, considering the date of the study maybe they performed their professions. It is stated that the teacher candidates - who are graduated - are currently teaching in various institutions (Baykara Özaydınlık, 2018). Therefore, it is possible that they will feel more competent because they have more experience. Another explanation is about whether the participants consider themselves realistic. As a matter of fact, PESS students who have a negative attitude towards the profession also consider themselves more adequate. In this context, since FE students evaluate themselves more realistically (Yalçın-İncik & Kılıç, 2014), they may find themselves inadequate.

As a result, it is seen that FE students' attitudes towards teaching profession and their perceptions of teacher self-efficacy do not differ significantly than teacher candidates who are trained in other programs. However, when analyzed in detail on the basis of programs, it is seen that FE students' attitudes towards the profession are significantly lower than PFCP students and the effect size of this difference is weak. The teacher self-efficacy perceptions of FE students were significantly higher than FT students, but significantly lower than NTMP students. It is concluded that faculties of education whose main purpose is to train teachers do not increase these features of their students sufficiently. In the light of these findings, the following suggestions can be presented to researchers and practitioners:

Since there are not enough studies in the programs such as FT and PESS that continue to train teachers, more comparison studies can be conducted on this subject.

Although a holistic result has been revealed with meta-analysis, qualitative studies can be conducted to provide detailed data on different research results.

Interviews with FE students can be made and factors affecting their attitudes towards profession and perception of teaching self-efficacy can be determined.

Lecturers at FE can use various methods to help their students gain more teacher self-efficacy and more positive attitudes towards the profession.

This study has several strengths and limitations. The literature was scanned by two different researchers and the studies reached were carefully examined. In addition to the overall comparison of FE with other programs totally, extensive findings were presented by examining the comparisons FE with each program. The study has limitations since studies included are from Turkey sample. In addition, there is only one study in some programs, thus this is not suitable for the purpose of meta-analysis.

Acknowledgements

This research study was supported by Kahramanmaras Sutcu Imam University as a research project with 2018/2-48 number and extracted from this supported project.

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).

ORCID

İsmail YELPAZE https://orcid.org/0000-0003-4428-0502

Levent YAKAR https://orcid.org/0000-0001-7856-6926

5. REFERENCES

Akdemir, A. S. (2013). Türkiye'de öğretmen yetiştirme programlarının tarihçesi ve sorunları. *Electronic Turkish Studies*, *8*(12) 15-28.

Arastaman, G. (2013). Eğitim ve fen edebiyat fakültesi öğrencilerinin öz-yeterlik inançları ve öğretmenlik mesleğine karşı tutumlarının incelenmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi (KEFAD), 14*(2), 205-217.

Atalmış, E., & Köse, A. (2018). Türkiye'deki öğretmen adaylarının öğretmenlik mesleğine yönelik tutumları: Bir meta-analiz çalışması. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 393-413.

Bağçeci, B., Yıldırım, İ., Kara, K., & Keskinpalta, D. (2015). Pedagojik formasyon ve eğitim fakültesi öğrencilerinin öğretmenlik mesleğine yönelik tutumlarının karşılaştırılması. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi, 17*(1), 307-324.

Bakioğlu, A., & Göktaş, E. (2018). Bir eğitim politikası belirleme yöntemi: Meta analiz. *Medeniyet Eğitim Araştırmaları Dergisi, 1*(2), 35-54.

Bandura, A. (1997). *Self- efficacy: The exercise of control*. New York: Freeman

Baykara Özaydınlık, K. (2018). A comparative analysis of preservice teachers' metacognitive learning strategies and teacher self-efficacy perceptions. *Hacettepe University Journal of Education, 33*(1), 125-143. https://doi.org/10.16986/HUJE.2017028409

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis.* UK: John Wiley & Sons, Ltd.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed effect and random effects models for meta analysis. *Res. Synth. Method, 1*, 97-111. https://doi.org/10.1002/jrsm.12

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2013). *Meta-analize giriş.* (Çev. S. Dinçer). Ankara: Anı Yayıncılık.

Brown, A. L., Lee, J., & Collins, D. (2015). Does student teaching matter? Investigating pre-service teachers' sense of efficacy and preparedness. *Teaching Education, 26*(1), 77-93.

Canales, A., & Maldonado, L. (2018). Teacher quality and student achievement in Chile: Linking teachers' contribution and observable characteristics. *International Journal of Educational Development, 60*, 33-50.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). New York, NY: Routledge.

Çakmak, M., & Ercan, L. (2018). Öğretmen adaylarının öğretmenlik mesleğine ilişkin tutumları ve problem çözme becerilerinin incelenmesi. *Ufuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 7*(13), 29-43.

Çapa, Y., & Çil, N. (2000). Öğretmen adaylarının öğretmenlik mesleğine yönelik tutumlarının farklı değişkenler açısından incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 18*(18), 69-73.

Çapri, B., & Çelikkaleli, Ö. (2008). Öğretmen adaylarının öğretmenliğe ilişkin tutum ve mesleki yeterlik inançlarının cinsiyet, program ve fakültelerine göre incelenmesi. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 9*(15), 33–53.

Çelik, Ö. C., Koç Erdamar, G., & Toraman, Ç. (2016). Gender differences in teachers and student teachers' self-efficacy beliefs: A Meta-Analysis. In E. Atasoy, R. Efe, I. Jażdżewska, & H. Yaldır (Eds.), *Current advances in education* (pp: 587-602). Sofya: St. Kliment Ohridski University Press.

Çelik, S., Örenoğlu Toraman, S., & Çelik, K. (2018). Öğrenci başarısının derse katılım ve öğretmen yakınlığıyla ilişkisi. *Kastamonu Eğitim Dergisi, 26*(1), 209-217. https://doi.org/10.24106/kefdergi.378129

Çetin, O. (2017). An investigation of pre-service science teachers' level of efficacy in the undergraduate science teacher education program and pedagogical formation program. *Journal of Education and Practice, 8*(12), 22-32.

Dadandı, İ., Kalyon, A., & Yazıcı, H. (2016). Eğitim fakültesinde öğrenim gören ve pedagojik formasyon eğitimi alan öğretmen adaylarının öz-yeterlik inançları, kaygı düzeyleri ve öğretmenlik mesleğine karşı tutumları. *Bayburt Eğitim Fakültesi Dergisi, 11*(1), 253-269.

Dinçer, S. (2014). *Eğitim bilimlerinde uygulamalı meta-analiz*. Pegem Akademi.

Doğan, S. (2013). *Sınıf öğretmenlerinin öz yeterlik algısı ve öğretmenlik mesleğine yönelik tutumlarının incelenmesi*. Unpublished master's thesis. Erzincan Üniversitesi, Sosyal Bilimler Enstitüsü.

Durmuşoğlu, M. C., Yanık, C., & Akkoyunlu, B. (2009). Türk ve Azeri öğretmen adaylarının öğretmenlik mesleğine yönelik tutumları. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 36*, 76-86.

Duval, S., & Tweedie, R. (2000a). A nonparametric" trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ (Clinical research ed.), 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the ınterpretation of research results*. Cambridge, UK: Cambridge University.

Erdamar, G., Aytaç, T., Türk, N., & Arseven, Z. (2016). The effects of gender on attitudes of preservice teachers towards the teaching profession: A meta-analysis study. *Universal Journal of Educational Research, 4*(2), 445-456.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher, 5*(10), 3-8.

Gülebağlan, C. (2003). *Öğretmenlerin işleri son ana erteleme eğilimlerinin mesleki yeterlilik algıları, mesleki deneyimleri ve branşları bakımından karşılaştırılmasına yönelik bir araştırma.* Unpublished master’s thesis. Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical research* ed.), *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hong, J., & Greene, B. (2011). Hopes and fears for science teaching: The possible selves of preservice teachers in a science education program. *Journal of Science Teacher Education, 22*, 491-512.

İlter, İ. (2009). *Öğretmen adaylarının öğretmenlik mesleğine ilişkin tutumlarının bazı değişkenler açısından incelenmesi* (Unpublished master’s thesis). Fırat Üniversitesi, Sosyal Bilimler Enstitüsü, Elazığ.

Kağıtçıbaşı, Ç. (2013). *Günümüzde insan ve insanlar*. İstanbul: Evrim.

Karabıyık, B., & Korumaz, M. (2014). Relationship between teacher's self-efficacy perceptions and job satisfaction level. *Procedia-Social and Behavioral Sciences, 116*, 826-830.

Kartal, O. Y., Temelli, D., & Şahin, Ç. (2018). Ortaokul matematik öğretmenlerinin bilişim teknolojileri öz-yeterlik düzeylerinin cinsiyet değişkenine göre incelenmesi. *Journal of Theoretical Educational Science*, *11*(4), 922-943.

Kaya, Z. (2001). *Bir meslek olarak öğretmenlik ve öğretmenlik mesleğine giriş*. Ankara: Pegem Akademi.

Kenrick, D. T., Neuberg, S. T., & Cialdini, R. B. (2005). *Social psychology: Unraveling the mystery* (3th ed.). Boston, MA: Pearson.

Kiraz, Z., & Dursun, F. (2015). Pedagojik formasyon eğitimi alan öğretmen adaylarının aldıkları eğitime ilişkin algıları. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 11*(3), 1008-1028. https://doi.org/10.17860/efd.37544

Korukçu, A. (2011). İlahiyat fakültesi son sınıf öğrencilerinin yaygın din eğitimine bakışları. *Değerler Eğitimi Dergisi, 9*(21), 55-97.

Lenhard, W., & Lenhard, A. (2016). Calculation of effect sizes. Retrieved from https://www.psychometrica.de/effect_size.html. Dettelbach (Germany): *Psychometrica.*

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. California: Sage Publications

Milli Eğitim Bakanlığı, (2017). *Öğretmenlik mesleği genel yeterlikleri*. MEB Öğretmen Yetiştirme ve Eğitimi Genel Müdürlüğü. Ankara

Özden, Y. (1999). *Eğitimde dönüşüm eğitimde yeni değerler*. Ankara: Pegem A Yayınları

Özkan, H. H. (2012). Öğretmenlik formasyon programındaki öğretmen adaylarının öğretmenlik mesleğine ilişkin tutumlarının incelenmesi (SDU Örneği). *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi* (KEFAD), *13*(2), 29-48.

Recepoğlu, E. (2013). Öğretmen adaylarının yaşam doyumları ile öğretmenlik mesleğine ilişkin tutumları arasındaki ilişkinin incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, Special Issue (1)*, 311-326.

Sakallı, N. (2001). *Sosyal etkiler: Kim, kimi nasıl etkiler*? Ankara: İmge.

Senemoğlu, N. (2012). *Gelişim, öğrenme ve öğretim kuramdan uygulamaya*. Ankara: Pegem Akademi.

Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the p value ıs not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Şaşmaz Ataçocuğu, M., & Zelyurt, M. K. (2017). Spor bilimleri fakülteleri mezunlarının işsizlik deneyimleri üzerine nitel bir araştırma. *Sportif Bakış: Spor ve Eğitim Bilimleri Dergisi, SI* (1), 70-97.

Tschannen-Moran, M., Woolfolk-Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*, 202-248.

Tucker, C. M., Porter, T., Reinke, W. M., Herman, K. C., Ivery, P. D., Mack, C. E., & Jackson, E. S. (2005). Promoting teacher efficacy for working with culturally diverse students. *Preventing School Failure: Alternative Education for Children and Youth, 50*(1), 29-34.

Tuncer, M. (2016). Evaluation of the attitude towards teaching profession in terms of gender through use of meta analysis method (A case study of Turkey). In E. Atasoy, R. Efe, I. Jażdżewska, & H. Yaldır (Eds.), *Current Advances in Education* (pp. 587-602). Sofya: St. Kliment Ohridski University Press.

Üstün, U., & Eryılmaz, A. (2014). Etkili araştırma sentezleri yapabilmek için bir araştırma yöntemi: Meta-analiz. *Eğitim ve Bilim, 39*(174), 1-32.

Yalçın-İncik, E., & Kılıç, F. (2014). Attitudes regarding the teaching profession, professional efficacy beliefs and vocational self-esteem of teacher canditates enrolled at education faculties and pedagogic formation programmes. *International Journal of Social Science and Education, 4*(2), 380-391.

Yaşar Ekici, F. (2017). Okul öncesi öğretmen adayları ile pedagojik formasyon eğitimi alan öğretmen adaylarının öğretmenliğe yönelik öz yeterlik inançlarının karşılaştırılması. *Journal of the Human and Social Sciences Researches, 6*(5), 3003-3022.

Yıldırım, I., Çırak-Kurt, S., & Şen, S. (2019). The effect of teaching "learning strategies" on academic achievement: A meta-analysis study. *Eurasian Journal of Educational Research, 79*, 87-114. https://doi.org/10.14689/ejer.2019.79.5

6. APPENDIX 1: Studies involved in Meta-Analysis

Akça, F., Demir, S., & Yılmaz, T. (2015). Öğretmen adaylarının özyeterlik algıları ile akademik kontrol odaklarının karşılaştırılması. *Journal of Innovative Human Sciences, 2*(1), 01-09. http://sproc.org/ojs/index.php/IJIRE

Aksoy, E. (2017). Turkish student teachers' attitudes toward teaching in university-based and alternative certification programs in Turkey. *Asia Pacific Education Review, 18*(3), 335-346.

Atasoy, M. U. (2010). *Lisans ve tezsiz yüksek lisans öğrenimi görmekte olan müzik öğretmeni adaylarının genel öğretmenlik öz yeterlik algılarının incelenmesi* (Unpublished Doctoral Thesis). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Bağçeci, B., Yıldırım, İ., Kara, K., & Keskinpalta, D. (2015). Pedagojik formasyon ve eğitim fakültesi öğrencilerinin öğretmenlik mesleğine yönelik tutumlarının karşılaştırılması. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi, 17*(1), 307-324.

Baykara Özaydınlık, K. (2018). A comparative analysis of preservice teachers' metacognitive learning strategies and teacher self-efficacy perceptions. *Hacettepe University Journal of Education, 33*(1), 125-143. https://doi.org/10.16986/HUJE.2017028409

Boran, M., & Yanpar Yelken, T. (2018). Eğitim fakültesi ile pedagojik formasyon öğrencilerinin öğretmenlik mesleğine yönelik duyarlılıklarının ve etkili öğretmen özelliklerinin incelenmesi. *Uluslararası Sosyal Bilimler Eğitimi Dergisi, 4*(2), 144-164.

Bozkırlı, K. Ç., & Er, O. (2011). Türkçe ve Türk dili ve edebiyatı öğretmeni adaylarının öğretmenlik mesleğine ilişkin tutumlarının çeşitli değişkenler açısından incelenmesi (Kafkas Üniversitesi örneği). *Turkish Studies International Periodical for the Languages, Literature and History of Turkish or Turkic, 6*(4), 457-466.

Coşkun, M. K. (2011). Din kültürü öğretmen adaylarının öğretmenlik mesleğine yönelik tutumları: İlahiyat-Eğitim DKAB karşılaştırması. *EKEV Akademi Dergisi, 15*(48), 269-279.

Çakmak, M., & Ercan, L. (2018). Öğretmen adaylarının öğretmenlik mesleğine ilişkin tutumları ve problem çözme becerilerinin incelenmesi. *Ufuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 7*(13), 29-43.

Çapri, B., & Çelikkaleli, Ö. (2008). Öğretmen adaylarının öğretmenliğe ilişkin tutum ve mesleki yeterlik inançlarının cinsiyet, program ve fakültelerine göre incelenmesi. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 9*(15), 33–53.

Çelen, A., & Eskicioglu, Y. (2015). özel yetenek sınavı ile öğrenci alan öğretmenlik bölümlerinde öğrenim gören öğrencilerin mesleğe yönelik tutum ve durumluk-sürekli kaygı düzeylerinin incelenmesi. *Route Educational and Social Science, 2*(3), 1-18.

Çetin, O. (2017). An investigation of pre-service science teachers' level of efficacy in the undergraduate science teacher education program and pedagogical formation program. *Journal of Education and Practice, 8*(12), 22-32.

Dadandı, İ., Kalyon, A., & Yazıcı, H. (2016). Eğitim fakültesinde öğrenim gören ve pedagojik formasyon eğitimi alan öğretmen adaylarının öz-yeterlik inançları, kaygı düzeyleri ve öğretmenlik mesleğine karşı tutumları. *Bayburt Eğitim Fakültesi Dergisi, 11*(1), 253-269.

Deniz, S. (2013). Öğretmen adaylarının öğrenme stilleri ve öğretmen öz-yeterlik algı düzeylerinin bazı değişkenler açısından incelenmesi. *International Online Journal of Educational Sciences, 5*(3), 667-684.

Elkatmış, M., Demirbaş, M., & Ertuğrul, N. (2013). Eğitim fakültesi öğrencileri ile formasyon eğitimi alan fen edebiyat fakültesi öğrencilerinin öğretmenlik mesleğine yönelik öz yeterlik inançları. *Pegem Eğitim ve Öğretim Dergisi, 3*(3), 41-50.

Ilgaz, G., Bülbül, T., & Çuhadar, C. (2013). Öğretmen adaylarının eğitim inançları ile öz-yeterlik algıları arasındaki ilişkinin incelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 13*(1), 50-65.

İpek, C., & Demirel, İ. N. (2012). Sınıf öğretmenliği ve pedagojik formasyon programı öğretmen adaylarının öğretmen özyeterlik inançları. *Bayburt Üniversitesi Eğitim Fakültesi Dergisi, 7*(1), 54-67.

Gedik, Z. G. (2015). *Öğretmen özyeterlik algısı ölçeğinin psikometrik özelliklerinin eğitim fakültesinden mezun olan ve olmayan öğretmen adayları gruplarında incelenmesi* (Unpublished Master Thesis). Abant İzzet Baysal Üniversitesi Eğitim Bilimleri Enstitüsü, Bolu.

Gürbüz, H., & Kışoğlu, M. (2007). Tezsiz yüksek lisans programına devam eden fen edebiyat ve eğitim fakültesi öğrencilerinin öğretmenlik mesleğine yönelik tutumları (Atatürk Üniversitesi Örneği). *Erzincan Eğitim Fakültesi Dergisi, 9*(2), 71–83.

Gürbüztürk, O., & Şad, S. N. (2009). Student teachers? beliefs about teaching and their sense of self efficacy: A descriptive and comparative analysis. *İnönü Üniversitesi Eğitim Fakültesi Dergisi, 10*(3), 201-226.

Kabadayı, M., Bostancı, Ö., Atan, T., Yazıcı, M., & Evli, F. (2011). Eğitim fakülteleri ile beden eğitimi ve spor yüksek okullarında okuyan öğretmen adaylarının öğretmenlik mesleğine karşı tutumlarının incelenmesi. *Türkiye Kickboks Federasyonu Spor Bilimleri Dergisi, 4*(2). Retrieved from http://edergi.kickboks.gov.tr/media/Dosya/yayin/07/2.doc

Keskin, Y. (2017). Coğrafya öğretmen adaylarının öğretmenlik mesleğine yönelik tutum ve kaygı düzeyleri (Erzurum Örneği). *e-Kafkas Eğitim Araştırmaları Dergisi, 4*(2), 43-57.

Ömür, Y. E., & Nartgün, Ş. S. (2013). Öğretmen adaylarının öğretmenlik mesleğine ilişkin tutumları ile güdülenme düzeyleri arasındaki ilişki. *Eğitimde Politika Analizi Dergisi, 2*(2), 41-55.

Özgür, N. F. (1994). *Öğretmenlik mesleğine kaşı tutum* (Unpublished Doctoral Thesis), Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul

Polat, S. (2013). Pedagojik formasyon sertifika programı ve eğitim fakültesi öğrencilerinin öğretmenlik mesleğine yönelik tutumlarının incelenmesi. e-*Uluslararası Eğitim Araştırmaları Dergisi, 4*(2), 48-60.

Poyraz, C., & Çağırgan Gülten, D. (2014). Pre-service mathematics teachers' attitudes towards the profession of teaching. *International Online Journal of Educational Sciences,6*(3), 558-569.

Sandıkçı, M., & Öncü, E. (2013). Beden eğitimi ile diğer alanlardaki öğretmen adaylarının öğretmenlik mesleğine ilişkin yeterlik algıları ve tutumlarının belirlenmesi ve karşılaştırılması. *Pamukkale Journal of Sport Sciences, 4*(1), 135-151.

Sezer, A., Pınar, A., & Yıldırım, T. (2010). Coğrafya öğretmeni adaylarının bazı profil özellikleri ve öğretmenlik mesleğine yönelik tutumlarının incelenmesi. *Marmara Coğrafya Dergisi*, (22), 43-69.

Taşgın, A. (2018). Öğretmen adaylarının öğretmenlik mesleğine ilişkin tutumları ile mesleki özyeterlikleri arasındaki ilişkinin incelenmesi. In A. İşcan, (Ed.), *Eğitim Bilimlerinde Örnek Araştırmalar* (pp. 87-105). Ankara: Nobel.

Timur, B., & İmer Çetin, N. (2017). Examining self-efficacy beliefs and attitudes of pre-service science teachers' and pedogogical proficiency students' towards science teaching profession. *International Journal of Active Learning*, *2*(2), 15-27.

Uygun, S. (2016). Pedagojik formasyon ve eğitim fakültesi öğrencilerinin öğretmenlik mesleğine yönelik duyarlıklarının karşılaştırılması. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi* (KEFAD), *17*(1), 313-330.

Ürün Karahan, B. (2017). Türkçe öğretmenliği ve Türk dili ve edebiyatı bölümü öğrencilerinin yaşam boyu öğrenme eğilimlerinin mesleğe yönelik tutumları ile ilişkisi. *e-Kafkas Eğitim Araştırmaları Dergisi, 4*(3), 30-44.

Yakar, L., & Yelpaze, İ. (2019). Öğretmen yetiştiren programlara kayıtlı öğrencilerin öğretmenlik mesleğine yönelik tutumları ve öğretmen öz-yeterlik algıları. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 47, 107-129. https://doi.org/10.9779/pauefd.473678

Yalçın-İncik, E., & Kılıç, F. (2014). Attitudes regarding the teaching profession, professional efficacy beliefs and vocational self-esteem of teacher canditates enrolled at education faculties and pedagogic formation programmes. *International Journal of Social Science and Education, 4*(2), 380-391.

Yaşar Ekici, F. (2017). Okul öncesi öğretmen adayları ile pedagojik formasyon eğitimi alan öğretmen adaylarının öğretmenliğe yönelik özyeterlik inançlarının karşılaştırılması. *Journal of the Human and Social Science Researches, 6*(5), 3003-3022.

Yavuz, M. (2010). An analyze of teacher candidate students' perception of self efficacy. *Procedia-Social and Behavioral Sciences, 2*(2), 1394-1398.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 570–587

https://doi.org/10.21449/ijate.716177

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Development of a Computerized Adaptive Version of the Turkish Driving Licence Exam

Nukhet Cikrikci[1,*], **Seher Yalcin**[2,*], **Ilker Kalender**[3], **Emrah Gul**[4], **Cansu Ayan**[2], **Gizem Uyumaz**[5], **Merve Sahin-Kursad**[2], **Omer Kamis**[2]

[1]Faculty of Education, İstanbul Aydın University, İstanbul, Turkey
[2]Faculty of Education, Ankara University, Ankara, Turkey
[3]Faculty of Education, Bilkent University, Ankara, Turkey
[4]Faculty of Education, Hakkari University, Hakkari, Turkey
[5]Faculty of Education, Giresun University, Giresun, Turkey

ARTICLE HISTORY

Received: Apr. 09, 2020

Revised: Sep. 3, 2020

Accepted: Sep. 19, 2020

KEYWORDS

Computerized adaptive testing,

Driving licence exam,

Hybrid simulation

Abstract: This study tested the applicability of the theoretical Examination for Candidates of Driving License (ECODL) in Turkey as a computerized adaptive test (CAT). Firstly, various simulation conditions were tested for the live CAT through an item response theory-based calibrated item bank. The application of the simulated CAT was based on data from e-exams administered by the Ministry of National Education (MoNE). Results of the first stage of the study were used to determine the rules for starting, continuing, and terminating the live CAT exam for ECODL. Secondly, the live CAT exam was applied according to the results of the simulation. Candidate drivers (n = 280) who had taken the ECODL as an e-test participated in the second stage. Thirdly, the opinions of the individuals who took the computer-based test towards the computer-based testing application were mapped. In the termination rule of the CAT-based ECODL, testing with a fixed number of questions yielded the smallest estimated measurement error. We also found that when ECODL was implemented as CAT, it could reliably differentiate among testers in terms of competence of theoretical knowledge of driving and provide basis for accurate decisions regarding their proficiency. According to the findings obtained on the candidates' opinions on the computer-based testing application, it was seen that they considered computer-based application more practical an easier in terms of testing.

1. INTRODUCTION

In today's world, training of drivers is of crucial importance because traffic safety depends primarily on driver's skills besides several other factors. Driving a vehicle is a complex process in which cognitive, affective and psychomotor skills are involved (Young, Regan & Hammer, 2007). Drivers who have not adequate skills may cause traffic accidents with death and or severe injuries or damage to public properties. Furthermore, such traffic accidents affect not

*CONTACT: Nükhet ÇIKRIKÇI ✉ nukhetdemirtasli@aydin.edu.tr 🖃 İstanbul Aydın University, Faculty of Education, İstanbul, Turkey and Seher YALÇIN, yalcins@ankara.edu.tr, Ankara University, Faculty of Education, Ankara, Turkey.

ISSN-e: 2148-7456 /© IJATE 2020

only parties involving the accidents; it also affects pedestrians, families, social life, economy of the country (Mulkat-Ajibola, 2015).

This is why being an adult is a legal prerequisite for being eligible to drive a motor vehicle (Miser, 1999). Mental health condition and physical abilities are also considered. Any individual who meets these criteria may apply for driving license examination. Passing this examination provides individuals with an official certificate which entitles people to drive.

In general, driving license examinations involve assessment of two traits: (i) theoretical aspects regarding driving (i.e. traffic rules, traffic signs, first-aid, etc.) and (ii) practice (i.e. driving a vehicle). If different examples in the world are considered, this common pattern can be observed. For example, in England, first phase of driving license examination has two sections (Bozkurt, 2003). One of the sections include multiple-choice items about traffic signs, rules and regulations, etc., whereas the other section is a hazard perception test. This test measures examinees' reactions for dangerous situation that can be faced while driving. Multiple-choices test is given as computerized, while perception test is administered by watching some scenarios. Germany gives examinees a computerized test including items about traffic signs, rules, etc. Examinees are also given several items specially designed for the type of vehicle they want to drive (AngloInfo the Global Expat Network Berlin, 2015). Item are randomly selected from an item pool and administered in a computerized medium. Australia uses a similar testing procedure. Items are selected from a pool (AngloInfo the Global Expat Network Berlin, 2015). Finland and Belgium have also similar procedures (Ministry of Transport and Communications, 2015). In all of these countries, completing training programs of is a prerequisite to take the driving license examination.

1.1. Turkey Case

In Turkey, causes of traffic accidents can be grouped into three: problems regarding (i) individuals (as pedestrians, passengers, and drivers), (ii) road and (iii) vehicles. Figures announced by the state agencies indicate that 86.2% of the traffic accidents were caused by drivers. The rest is shared by passengers (0.47%), pedestrians (9.38%), roads (0.95%), and vehicles (0.58%) (Turkish Statistical Institute, 2018).

As figures point out, drivers are the leading factor in traffic accidents. Turkish governments have been implementing some policies regarding traffic accidents. These policies can be group into three: (i) security measures including penalties, (ii) training to increase security level of pedestrians and drivers, and (iii) selection of individuals who are fully eligible to drive.

Although there are several rules and regulations regarding the first two groups of policies implemented by Turkish bodies, the last item (selection procedures) seems to be neglected. For example, a common finding indicated by the literature is that content, length, and objectives of the training programs of learner drivers are not enough and appropriate (Balkız, 1999; Çakır, 2006; İnal, 2001; Kuyumcu, 2001; Tanrıkulu, 2002; Türkoğlu, 2002; Tütüncü, 2001; Vursavaş, 2004). Another finding reported in the literature is that training of learner drivers provided a significant increase in drivers' knowledge of driving (Bozkurt, 2003; Çakır, 2006; Gülecen, 1998; Kaçmaz-Omak, 2012; Türkoğlu, 2002). Elander, West and French (1993) found that risk-taking behaviour of drivers have a direct relationship between traffic accidents. Despite the large body of research in the literature regarding driving licenses, there is no study examining the selection procedure of drivers in Turkey. However, certification process of individuals who are eligible to drive is as important as other components such as rules, regulations, and quality of training programs.

1.2. The Turkish Examination for Candidates of Driving License (ECODL)

In Turkey, all candidates should complete a training program including both theoretical and practical aspects of driving before applying for the driving license exam. These training programs are offered by private institutions certified by Turkish authorities. The training program includes sixty-three hours for theory and 20 hours of practice: 42% traffic and environment, 15% first-aid, 24% driving, 19% motor and vehicle techniques and driving technique (Çakır, 2006). Those who complete the training program successful become eligible to take the driving license examination.

The Examination for Candidates of Driving License (ECODL), the standard driving exam in Turkey, has also been offered as a computer-based exam, or e-exam, since 2011 in addition to paper-and-pencil tests. The first part of the driving license examination includes tests measuring different theoretical aspects of the driving. ECODL, driving licence exam has three sections: (i) first aid, (ii) traffic and environment, and (iii) motor and vehicle technique. The items in the bank were included in multiple-choice questions, each with four options.

The ECODL theoretical examination in Turkey currently administered by MoNE has some problems, such as test items being used without a prior pilot study, an excessive number of items testing knowledge of Bloom's taxonomy, limitations with respect to measurement techniques, and the questions being leaked to the public beforehand. Nearly identical questions are used in successive ECODL test sessions. Furthermore, although offered as e-examinations, these computer-based tests are restricted in terms exploiting the benefits of the technology (for example, item assignment suitable to the level of capability of the candidate).

1.3. Development of Computerized Adaptive Version of ECODL

By considering these issues, this study develops a valid and reliable ECODL test that can reliably reflect candidates' proficiency in a computerized environment by using the advantages of this technological development. Accordingly, it examines the applicability of the theoretical examination of ECODL as a computerized adaptive test (CAT). In the present study, development of computerized adaptive version The Turkish ECODL was examined. Some examples of e-exams in various disciplines across the world are the GMAT (Graduate Management Admission Test), GRE (Graduate Record Examination), MAP (Measures of Academic Progress), NCLEX (National Council Licensure Examinations), ASVAB (Armed Services Vocational Aptitude Test Battery), and STAR (Math, Reading, and Early Literacy) (Economides & Roupas, 2007).

The aim of computerized adaptive tests (CATs) is to effectively predict the candidate's proficiency using shorter tests than paper-and-pencil exams (Weiss, 2004). CAT-based tests are more reliable and useful as they match the difficulty of questions with the ability/competence of the test taker. Items in the test are chosen to provide the most information about the candidate's ability (Orcutt, 2002; Weiss & Kingsbury, 1984). A significant benefit of this is that it allows the use of an item bank with estimated parameter values according to the IRT (Item Response Theory), which defines capability parameters and item parameters on the same scale (theta scale). Therefore, it can predict the level of difficulty of questions that a respondent can correctly answer based on his/her ability predicted using the answers to previous questions (Weiss, 2004).

The greatest advantage of CAT is that it provides functional testing through uncompromising adherence to the validity and reliability of measurement. In other words, it provides a difficulty-based question display for candidates depending on their previous responses. In any CAT application, if a candidate correctly answers a question with a mean difficulty, which is shown on the screen, the computer displays a more challenging question next. If the candidate responds incorrectly, the system chooses an easier item next. Thus, respondents with high proficiency levels are assigned more difficult items whereas those with lower levels are presented easier items (Davey, 2011; DeMars, 2010; Weiss, 1982). As a result, the test ensures the optimal question assignment that can best reveal the candidate's proficiency according to performance.

This case allows respondents to answer shorter tests in less time, without the need to assign to them questions that are too difficult, and thus above the proficiency, or too easy items or below their proficiency (Linden & Glas, 2002; Wainer & Mislevy, 1990; Weiss & Kingsbury, 1984). This enhances the usefulness of CAT.

Another advantage of CAT is the low likelihood of cheating on it. That is, the probability of cheating is eliminated because every respondent is given a test consisting of different questions in accordance with his/her proficiency, which enhances the security of CAT (Thompson & Weiss, 2011). Moreover, this ensures that the test results are generated more quickly (Linacre, 2006; Orcutt, 2002; Rudner, 1998; Weiss, 2004; Weiss & Kingsbury, 1984). CAT also has some limitations: the necessity of a large item bank; the respondent's inability to go back to a previous question, check it, and change answers; and test bias due to familiarity with computers (Bugbee & Bernt, 1990; Economides & Roupas, 2007; Sutton, 1991).

Apart from the well-accepted advantages of CAT application, it still has some discussable disadvantages. For instance, it is claimed that reading a text from a computer screen takes longer than reading from a printed material. Moreover, solving mathematics questions is possible to be problematic since the test takers will not be able to underline the text, which is a strategy while answering an item in the test (Bugbee & Bernt, 1990). Not being able to go back to the questions and to check, to edit or to change them during the implementation is also discussed to be a limitation (Linacre, 2000). Another highlighted problem is the concern for a possible increase in bias in tests. It is stated that computer familiarity (hence some variables such as socio-economic status, or gender) may affect the test results as a source of bias. It is thought that this may be an extra source of anxiety for the ones who are not familiar with computers (Linacre, 2000; Sutton, 1991). On the other hand, while this situation would be more problematic in the past, it is expected to decrease today since the computer and smart phone technologies have entered into our lives at such a pace.

The first set of studies on CAT applications in Turkey started with the assessment of academic achievement in science and mathematics (Kaptan, 1993; Köklü, 1990; Yaşar, 1999). The findings showed that there was no significant difference in the academic achievement of students as reflected in CAT and traditional paper-and-pencil tests. In numerous subsequent studies, the applicability of CAT to academic tests for transition to secondary and higher-education examinations was explored (İşeri, 2002; Kalender, 2011). From the findings, it was clear that the examinations can be applied as CATs on the condition that the item bank was enriched with a larger number of specific questions.

Subsequent studies on CAT examined the applicability of different proficiency tests that were not administered by the federal government but were taken by large groups (academic achievement, foreign language proficiency, computer literacy) according to different CAT strategies (Gökçe, 2012; Kezer & Koç, 2014; Özbaşı & Demirtaşlı, 2015). Some studies were conducted in this vein in medicine (Öztuna, 2012), and that long surveys given to patients for accurate diagnoses can be employed as CATs. These studies encouraged the development of CAT software (Kalender, 2011). In general, the above-mentioned common findings show that a variety of tests in different domains can be offered as CATs. A sufficiently large item bank is also needed such that it can ensure the reliability of tests, consisting of specific questions while addressing a wide range of capabilities. Regarding such a popular testing application whose both advantages and disadvantages are discussed, it is considered to be important to get the opinions of the test takers right after a live CAT application in order to be able to research if it is possible to switch into this application, to make suggestions for a possible implementation and to develop methods regarding the precautions to be taken. Powers and O'Neill (1993) stated that the opinions of the study group changed positively and greatly after taking the real application and mentioned the importance of taking opinions after the implementation.

1.4. Research Questions

(1) What are the methods of estimating capability, and what is the termination rule with the minimum error according to the simulation for a live CAT application?

(2) As a result of the application of the live CAT under the selected ideal conditions,

a. What was the testing time per person?

b. What were the distributions of standard error of the test information values of estimation, proficiency level, and estimated proficiency level?

(3) What are the opinions of the individuals who took the ECODL exam on the computer-based application?

2. METHOD

As stated earlier, this study describes the development of computerized adaptive version of The Turkish ECODL. The development of a CAT-based ECODL starting by creating an item bank for first aid, motor and vehicle technique, and traffic and environment. Items were pre-tested before including into the item bank. Then several CAT simulations were run in order to determine optimal CAT design before live CAT. Then, a group of learner drivers were participated to live CAT version of ECODL. Participants were also asked to report their opinions about the live CAT stage. The rest of this section details the methodology of the study.

2.1. Development of the Item Bank

In the item writing phase, academician who have expertise in item development, trainers of learner drivers, and staff from Ministry of National Education worked in collaboration. A teacher who teaches traffic gave a training for item developers. Separate test plans for each subject were prepared prior to the item writing process to specify the basic and advanced driver proficiencies at three cognitive levels (knowledge, comprehension, and problem solving). The number and distribution of the items employed for measurement were also specified in the plan. The items were written by the authors using a checklist. Each item writer sent his/her items to field experts who independently checked them (item fit to measurement, grammar, spelling, punctuation, and scientific accuracy). Scientific accuracy of items inspected by teachers who traffic, teach first-aid and health classes. In addition, a drawer was employed to draw visuals in the items. Items were revised more than once based on suggestions made by the parties involved. Item development phase completed between November 2017 and March 2017. At the end, a total of 913 multiple-choice items with 4 alternatives were compiled to be included in the pilot study.

2.2. Pilot Study for the Items

The item bank was pilot-studied on a sample of 1787 university students from different cities of Turkey. Also, between June and August in 2017, items were pilot-studies on learner drivers when they took computerized (not adaptive) ECODL. Test-takers were given regular items as well as items developed for the CAT version. At that stage, 90% of the item bank were tested by around 250 learner drivers.

Another phase of close inspection was conducted. Items were once again revised scientific accuracy, measurement quality, grammar, clarity. Some items which included large visuals could not the displayed on the computer screen properly. These items were excluded from the item bank. Finally, a total of 892 items (417 for traffic and environment, 270 for first-aid and 192 for motor and vehicle technique) were kept in the item bank.

2.3. Item Calibration

In a period of 8 weeks, the 892 items were used for learner drivers who took actual driver license examination. Some of items were not responded by sufficient number of learner drivers (at least 200 responses for each item). As a result, 235 items for traffic and environment, 201 for first-aid and, 192 items for motor and vehicle techniques were left in the item bank.

Analyses revealed that 3PL model had the best item-model fit. Some of the items in the item bank indicated poor item fit and excluded from the study. 193 items for traffic and environment, 139 for first-aid and, 167 items for motor and vehicle techniques were kept. Table 1 presents the items parameters for each domain.

Table 1. *Descriptive of the domains in the item bank.*

	Traffic and Environment			First-Aid			Motor and Vehicle Techniques		
	a	b	c	a	b	c	a	b	c
Min	0.30	-4.00	0.10	0.30	-4.00	0.18	0.30	-4.00	0.17
Max	1.08	4.00	0.38	1.18	4.00	0.30	1.88	4.00	0.43
Mean	0.51	0.90	0.24	0.48	1.35	0.25	0.57	0.35	0.25
Median	0.45	1.00	0.25	0.45	2.28	0.20	0.54	0.32	0.25

2.4. Simulations Prior to Live CAT

Prior to live CAT administration, as suggested in the literature (Thompson & Weiss, 2011), a series of post-hoc simulations were conducted to define optimal live CAT design. At that phase of the project, different CAT strategies were employed. Starting rule and item selection procedure were kept constant. At the beginning of each simulation, examinees were assigned an ability score of 0. Items were selected using Fisher's method (Weiss, 1982). On the other hand, two ability estimation methods were used: Bayesian Expected a Posteriori and Maximum Likelihood Estimation. Also, two different test termination methods were defined: a fixed number of items and SE-threshold. For SE-based test termination, SEs of 0.30 and 0.40 were used as threshold. Fixed number of items were defined as 50, number of items given in ECODL.

At that stage, it was noted that since not every item in the item bank developed for CAT was administered during the pilot study, there were missing values in the item-person matrix (Weiss & Guyer, 2012). As a solution, the hybrid simulation method was adopted (Nydick & Weiss, 2009). In the hybrid simulation approach, the available item set answered by every respondent was used to predict the ability of each respondent on all questions that were not answered by him/her. Estimation was then employed to assign responses to the omitted and unanswered items using a suitable IRT model and the Monte Carlo simulation. An item response matrix with all data for every respondent that could be employed in a post-hoc simulation was thus obtained (Weiss & Guyer, 2012). Thus, six different CAT (three test termination rules x two ability estimation methods) strategies were simulated.

2.5. Live CAT

The live CAT test was performed with an item bank of a total of 628 items. A computer software developed by Kalender (2011) was used to conduct live CAT. The software was installed 11 laptop computers. The live CAT administrations, which were applied between September and November 2017, were conducted just after the learner derivers took ECODL in testing centers. Learner drivers were invited to voluntarily participate to the live CAT sessions. A small gift (a first-aid kit) was given to each participant. After the purpose of the study was explained and the consent were received, the participants were administrated a live CAT session.

A total of 280 learner drivers (age mean: 21 years, s.d.: 5.19) were participated in this study. The descriptive statistics of the participants are given in Table 2. As can be seen, most of the participants are at least high school graduates. More than 85% of the participants have experience with computers more than 5 years. Almost 90% of the participants preferred ECODL exam in computerized format instead of paper-and-pencil format.

Table 2. *Demographic information of live CAT participants.*

Variable	Value	N	%	Total
Gender	Female	121	43.2	
	Male	159	56.8	280
Educational background	Primary school	5	1.8	
	Secondary school	8	2.9	
	High school	89	31.8	
	Undergraduate	168	60.0	
	Postgraduate	10	3.6	280
Any experience with computers	None	5	1.8	
	1-2 Years	11	3.9	
	3-5 Years	22	7.9	
	More	242	86.4	280
Any experience with computerized tests	Yes	118	42.1	
	No	161	57.5	279
Preferred test format	Computerized	250	89.3	
	Paper-and-pencil	29	10.4	279

2.6. Opinions after the Live CAT

Just after the live CAT administrations, participants were also given a questionnaire to report their opinions. The questionnaire consisted of two parts. The first part included some questions on participants' demographic information (gender, education level) and their experience with computers (computer usage experience, computer-based testing experience). The second part included 18 items inquiring their opinions towards the comparison of computerized tests and pen-and-paper tests and the techniques and usefulness of computerized tests. Each item was presented as an opinion statement and the participants were asked to mark their agreement level among the suitable reflective categories which ranged from *Completely agree* (1) to *Completely disagree* (4).

2.7. Data Analysis

Data analysis was carried out in three phases. In the first, IRT-based analysis of each item included in the pilot study was performed to determine ones that were in the item bank and predict their related parameters. Before IRT analyzes, the unidimensionality assumption was examined separately for each booklet on the basis of every subject area. In this context, exploratory factor analysis based on the tetrachoric correlation matrix was done with the R programming. It has been observed that there was one dominant dimension in all subject areas. The provision of the unidimensionality assumption shows that the local independence assumption is also met (Hambleton & Swaminathan, 1985). Also, the test completion time of the participants shows that the test was not a speed test. The responses of the candidate drivers were analyzed according to the IRT scaling with best model–data fit. As a result of the item analyses carried out using the 3PL model that showed the best model–data fit, the difficulty of the questions/items varied in the range (-4.00 and +4.00) and item discrimination in the range (0.30 and 1.88). The probability of correct response by those with low capability was estimated at 0.25. Questions with the expected level of measurement quality were selected and included in the item bank according to the test subjects.

For sub-questions a and b, in accordance with the second research sub-question of the study, descriptive statistics of the participants' response times, the final proficiency estimation, standard error of estimation, and test information were calculated. The final proficiency and the standard error in measurement obtained from the CAT application were analyzed to find the distribution of proficiency estimation and values representing the reliability of the estimation. Test information for each respondent obtained from the overall ECODL test and the total test information value for all participants in the live CAT application for each ECODL were calculated. In IRT-based test development, reliability is the amount of test information at a certain theta/proficiency level (Hambleton, Swaminathan & Rogers, 1991). For the third sub-question, the opinions of the group were made suitable to interpret on the item level by taking the frequencies and percentages in the reflective categories for each item.

3. RESULT

According to the results of the simulative CAT of ECODL analyses in accordance with the first research sub-question, the estimated ability of all test subjects yielded the smallest error when the EAP method was employed. With regard to the rule for test termination, the application with fixed questions was found to yield the smallest estimated error. The initial level of theta was set to zero in all conditions. The results of an analysis of the simulation conditions of the CAT are presented in Table 3.

Table 3. *Analysis of tested conditions in CAT simulation on three ECODL subjects.*

Test Subject	Condition No.	Estimated Ability on Method	Test Termination Rule	Item Selection Method	Mean Measurement Error of the Item Bank	Mean Measurement Error of Proficiency Estimations	Mean Number of Items (Minimum and Maximum Numbers of Questions)
Traffic and Environment	1	2	Se (0.30)	1	0.887	0.887	190 (190–190)
	2	3	Se (0.30)	1	0.396	0.396	187.967 (44–190)
	3	2	Fixed question (23)	1	0.885	1.253	23 (23–23)
	4	3	Fixed number of items 23)	1	0.396	0.487	23 (23–23)
	5	2	Se (0.40)	1	0.891	0.891	190 (190–190)
	6	3	Se (0.40)	1	0.396	0.410	123.932 (28–190)
Motor and Vehicle Technique	1	2	Se (0.30)	1	1.010	1.010	139 (139–139)
	2	3	Se (0.30)	1	0.477	0.477	139 (139–139)
	3	2	Fixed question (12)	1	1.017	1.592	12 (12–12)
	4	3	Fixed number of items (12)	1	0.477	0.606	12 (12–12)
	5	2	Se (0.40)	1	1.018	1.018	139 (139–139)
	6	3	Se (0.40)	1	0.477	0.477	138 (37–139)
First Aid	1	2	Se (0.30)	1	0.745	0.745	170 (170. 170)
	2	3	Se (0.30)	1	0.328	0.339	130.703 (31. 170)
	3	2	Fixed question (9)	1	0.735	1.488	9 (9. 9)

(Continued)

Table 3. Continued

Test Subject	Condition No.	Estimated Ability on Method	Test Termination Rule	Item Selection Method	Mean Measurement Error of the Item Bank	Mean Measurement Error of Proficiency Estimations	Mean Number of Items (Minimum and Maximum Numbers of Questions)
First Aid	4	3	Fixed number of items (9)	1	0.328	0.553	9 (9. 9)
	5	2	Se (0.40)	1	0.744	0.744	170 (170. 170)
	6	3	Se (0.40)	1	0.329	0.398	37.403 (18. 170)

Note: 1: Item selection method based on maximum information, 2: Maximum likelihood estimation, proficiency estimation method, 3: Expected a posteriori estimation, proficiency estimation method

The CAT results in Table 3 show that the best strategy in the item bank for all three test subjects was condition 4. Accordingly, when questions in live CAT were given, the test needed to be terminated with a fixed number of questions. It was thus decided that a fixed number of questions, parallel to questions used by MoNE, would be best: 23 for traffic and environment, 12 for first aid, and nine for motor and vehicle technique. When the EAP method was used for proficiency estimation, the result was reliable. In this case, the item bank of all three test subjects was arranged according to these strategies to be employed in live CAT. Descriptive statistics of the participants' response times in the live CAT performed under ideal conditions, defined in question "a" of the second sub-question of the research, are presented in Table 4.

Table 4. *Descriptive statistics of participants' response times in live CAT application of ECODL.*

	Response Time (min)	
Test subject	Mean	Median
Traffic and Environment	7.97	7.85
First Aid	4.97	4.87
Motor and Vehicle Technique	2.11	2.05

Table 4 shows that the mean response time ranged from two to eight minutes according to test subject. Although the ECODL as a Computerized Adaptive Test (ECODL_CAT) application had a fixed number of questions, it was practical in terms of time taken for application. The distributions of the predicted proficiency levels, test information levels, and standard errors according to the live CAT application performed under the ideal conditions defined in question "b" of the second sub-question of the research were examined. At the end of the ECODL tests, the proficiency level of each participant (theta) and standard error of measurement (Sem) values of estimated proficiency were obtained. As a result of the IRT-based CAT application of ECODL, test information values calculated in light of the participants' predicted proficiency levels (values) and the measured standard errors in estimated proficiency were considered the criteria of reliability (Embretson & Reise, 2000; Hambleton et al., 1991), and were obtained for each proficiency level. The levels of theta of the participants' estimated proficiency, and the related estimated errors and distribution of test information are summarized in Table 5.

Table 5. *Descriptive statistics of distributions of test information values, estimated proficiency, and errors in estimated proficiency of participants in live CAT application of ECODL.*

Statistics	Distribution of estimated proficiency (θ) of the participants	Distribution of test information values ($I_i(\theta)$) of estimated proficiency of participants	Distribution of measured standard errors (Sem) in estimated proficiency of participants
Mean	0.38	6.31	0.40
Median	0.52	6.47	0.39
Minimum and Maximum	-3.00–2.96	3.16–8.21	0.35–0.56
Standard Deviation	1.28	1.00	0.03

Table 5 shows that individual proficiency levels in the ECODL_CAT application were moderate in terms of theta, in the range of -4.00 < ϴ < +4.00 with a standard deviation of 1.00 and an IRT theta scale mean of zero. The value of theta of participants was scattered over a wide range around and at the moderate level of ability (mean: 0.38; range: -3.00 and 2.96). Theta was distributed in two extremes and over a wide range. Respondents who showed a more homogeneous distribution of scores (min: 56, max: 98, mean: 81.65, median: 82, sd: 7.34) in the ECODL e-examination conducted by MoNE, displayed a more heterogeneous distribution of ability in the CAT application. This might have obtained because the item bank used in the ECODL_CAT application consisted of items at different levels of difficulty and a sufficient number of parameter values to discriminate among them. Items in the bank had good psychometric quality, and were employed after being tested on the candidates for the driver's exam. Thus, an ECODL_CAT application based on these items provides a valid and reliable measurement of individual differences. The measured distribution of standard error (Sem) of the participants' estimated proficiency ranged from 0.35 to 0.56. The calculated test information values ranged from 3.16 to 8.21. In the scope of the third sub-question of the research, after grouping the opinions given on the questionnaire items according to their common points, the results were summarized in Table 6.

Table 6. *Frequency and percentages about questionnaire items.*

		Response Categories					
	Items	Statistic	I Completely agree	I Agree	I don't agree	I never agree	Total
Opinions about comparison of CAT and paper-pencil applications	i1	Frequency	174	76	23	6	279
		%	62.4	27.2	8.2	2.2	100.0
	i2	Frequency	157	75	36	10	278
		%	56.5	27.0	12.9	3.6	100.0
	i5	Frequency	161	82	18	17	278
		%	57.9	29.5	6.5	6.1	100.0
	i9	Frequency	46	43	87	90	266
		%	17.3	16.2	32.7	33.8	100.0
	i10	Frequency	143	112	15	4	274
		%	52.2	40.9	5.5	1.5	100.0
	i16	Frequency	199	56	14	5	274
		%	72.6	20.4	5.1	1.8	100.0

(continued)

Table 6. Continued

	Items	Statistic	Response Categories I Completely agree	I Agree	I don't agree	I never agree	Total
Opinions about technical application properties of CAT application	i3	Frequency	186	78	10	4	278
		%	66.9	28.1	3.6	1.4	100.0
	i4	Frequency	107	95	44	31	277
		%	38.6	34.3	15.9	11.2	100.0
	i12	Frequency	169	98	5	2	274
		%	61.7	35.8	1.8	.7	100.0
	i13	Frequency	23	9	71	171	274
		%	8.4	3.3	25.9	62.4	100.0
	i14	Frequency	109	100	38	24	271
		%	40.2	36.9	14.0	8.9	100.0
Opinions about visual properties	i6	Frequency	200	55	10	7	272
		%	73.5	20.2	3.7	2.6	100.0
	i8	Frequency	138	97	32	6	273
		%	50.5	35.5	11.7	2.2	100.0
	i18	Frequency	12	31	105	125	273
		%	4.4	11.4	38.5	45.8	100.0
Opinions about content of the items used in CAT application	i7	Frequency	98	108	61	7	274
		%	35.8	39.4	22.3	2.6	100.0
	i11	Frequency	152	104	14	2	272
		%	55.9	38.2	5.1	.7	100.0
	i15	Frequency	48	82	98	41	269
		%	17.8	30.5	36.4	15.2	100.0
	i17	Frequency	59	102	89	23	273
		%	21.6	37.4	32.6	8.4	100.0

As it is seen in Table 6, more than half of the participants marked 'completely agree' on the majority of the items. In line with this, the participants preferred computer-based application at a high rate in terms of marking the answers both faster and more easily for the items that compared the computer-based application they experienced within the project and pen-and-paper tests. In frame of the items which inquired the opinions towards the technical properties of CAT application, most of the participants demanded the possibility to go back and check their previous answers. Apart from this, they stated that they understood the instructions at beginning of the test and that they did not have any problems reading using computers.

The participants have positive opinions towards the visual properties of the computer-based application in the project. According to this, seeing only one question on the screen, the visual quality, the screen brightness and the resolution were evaluated positively in terms of its relation with the participants' performance on answering. Regarding the content of the items used in CAT application, almost half of them considered it to be easy as they could answer the questions showing up on the screen while almost half of them found the content difficult. Similarly, half of the participants thought the questions were easier than those of MoNE. In addition, the majority of the participants (almost 80%) had in the opinion that the questions were clear and understandable.

4. DISCUSSION and CONCLUSION

The purpose of this study was to develop a valid and reliable application for electronic examinations for driver's licenses. Accordingly, the applicability of CAT to the ECODL theoretical examination was tested and verified. In this context, the ideal conditions for live

CAT were defined through a simulation. This study can serve as basis for further research to improve and upgrade examinations for driver's license.

The results of simulations showed that when the number of questions in the MoNE ECODL was considered fixed with a termination rule, the CAT application yielded estimation proficiency with a smaller error. When the EAP method was used for proficiency estimation, reliable results were obtained. Some studies that have employed CATs have also concluded that the EAP method gives better results than other techniques (Eroğlu, 2013; Kezer, 2013).

According to the findings, the total time taken by respondents to answer the entire test was approximately 15 minutes on average. In this respect, although the ECODL_CAT had a fixed number of questions in this study, it was practical in terms of time taken to complete the test. Because ECODL_CAT considered the level of difficulty of questions in terms of individual proficiency, candidates did not have to deal with questions above or below their levels of proficiency. This is an indication of the practicability of live CAT, a major advantage of adaptive tests (Kimura, 2017; Thissen & Mislevy, 2000; Thompson & Weiss, 2011).

The mean and median values of the standard measured error for the ECODL_CAT application was within the limits (0.50–0.30) suggested and accepted in the literature, even though the CAT application with a fixed number of questions was used (Embretson & Reise, 2000; Linacre, 2006). The standard error in measurement in this case was expected to range from 0.50 to 0.40, provided that 15–30 questions had been posed, to obtain the probability of correct responses (achievement). An error of 0.50–0.70 was obtained. The results show that the ECODL_CAT theoretical test measures differences in drivers' proficiency in theoretical knowledge more reliably than the ECODL e-test application of MoNE.

The bank consisting of items with a good model–data fit according to IRT provided the questions used in the CAT application and the predicted proficiency parameters (obtaining fixed items and proficiency estimations) (Hambleton et al., 1991). Moreover, given that, it is compatible with CAT applications; IRT-based item bank development enhances the reliability and practicability of the ECODL_CAT (Thissen & Mislevy, 2000; Thompson & Weiss, 2011). However, written based only on expert views and test subjects, items in the MoNE e-examination application have been used without any pilot study. No objective proof of the difficulty of the questions or their ability to discriminate among candidates' proficiency has thus been provided.

The results here have shown that ECODL can reliably measure (by ensuring adequate test information) candidate drivers' proficiency even though the CAT application had a fixed number of questions. They also show that when ECODL was applied as a CAT, it reliably predicted individual ability levels. Most of the participants stated that they considered the computer-based application more practical and easier in terms of testing. These findings are consistent with the findings of the research done by Schmidt, Urry & Gugel (1978), in which the opinions of 163 participants who took part in the adaptive test pilot study in American Civil Service Commission were taken. On the other hand, the participants stated that they did not experience any technical difficulties. This case contradicts the finding in the study done by Schmidt et al. (1978). One of the topics that the participants complained about in their study was technical details such as screen light; however, this is thought to stem from the fact that the study was a very old one. The possibility to experience such a problem with today's technological opportunities is low. Georgiadou, Triantafillou & Economides (2006), in their study in which they determined the standards for CAT application, stated technical properties and visual design to be one of the dimensions they cared. They said that one of the important points regarding CAT application was designing the screen so clear that the test-takers could understand it on their own and they identified the situation as "Using CAT should be so simple, clear and easy that the test-taker can focus on the question's correct answer rather than how the system works" (p. 268).

The participants stated that the item contents were clear and understandable in terms of visual properties and content. Moreover, the numbers of ones who claimed that they encountered easier and more difficult questions than they could do are quite close to each other, which means there is no apparent agglomeration of a particular view. Although it is stated that CAT provides the individuals to encounter the questions which are consistent with their skills as a working principle in theory, Linacre (2000) states that this situation creates different situations in terms of the perception of the individuals with high performance. Highly skilled individuals are accustomed to a 90% correct response rate in paper-and-pencil applications. However, in CAT application, the more correct answers an individual gives, the more difficult questions he or she encounters and the success of the that person is determined jointly by the person's performance with the item selection algorithm. This may cause the successful individuals to be surprised during the test.

When the study group's experience with computers is examined, it is seen that the majority of the group is a computer user for more than 3-5 years. Computer familiarity is a variable that has been studied in various aspects of CAT applications in the literature. When the literature is examined, it is possible to come across findings that state that there is a relationship between computer familiarity and attitude towards computer (Powers & O'Neill, 1993; Wilder, Mackie, & Cooper, 1985) and the relationship between familiarity and attitude towards computer testing (Burke, Normand & Raju, 1987). Similarly, in this study, it was found that the majority of the study group (86.4%) had a high computer familiarity and likewise, the majority of the group (89.3%) preferred to take the test in the computer environment. Similarly, the majority of the study group (67%) stated that they did not experience any anxiety about computerized testing. In some studies (Kernan & Howard, 1990; Powers & O'Neill, 1993), while there was a relationship between familiarity and computer anxiety, Kim and McLean (1995) stated that individuals' participation in paper-and-pencil form or CAT application did not make a significant difference in terms of anxiety towards testing.

It is possible to say that the most basic point that individuals are not satisfied with is the lack of opportunity to review their answers. The majority of the group (95%) stated that they would prefer to go back and check their answers. This finding is consistent with the study conducted by Schmidt et al. (1978). On the other hand, Vispoel (1998, 2000) worked on two groups in which the students were allowed and were not allowed to return to their answers in the computer-based exam in their studies and examined the answer patterns. As a result of both studies, a very small percentage (3-5%) of the students changed their answers. In addition to this, it was stated that there was no significant difference between the groups who changed their answers and those who did not.

This research has certain limitations. Because the items were randomly assigned to respondents in the e-examination administered by MoNE, the 377 new items that were generated were not responded to by a sufficient number of candidates and thus were not included in the analysis, although the initial number of items in the pilot study was higher (876 items). In future studies on this subject, necessary precautions must be taken to ensure that every candidate driver is given new items at a certain rate in the e-exam. Another limitation is that the items in this study were multiple-choice questions and fitted the 3PL IRT model well. In future work, open-ended items with restricted responses can also be used in ECODL tests in addition to multiple-choice questions. In this case, psychometric results can be tested by sampling these items, and the precision of capability estimation between multiple-choice questions and mixed ECODL CAT test questions can be compared.

In this study, item exposure was not considered in the e-ECODL applied as live CAT. Item exposure represents the number of times a given item is used in all CAT applications (Georgiadou, Triantafillou & Economides, 2007; Revuelta & Ponsoda, 1998). It is associated with the validity of test scores and is intended to prevent changes in level of difficulty of questions in case test takers are familiar with a few. Future work can examine the effects of the incorporated use of item exposure with other termination strategies by considering different values of the reliability of proficiency estimation.

The MoNE carries out ECODL e-exam applications in 31 different halls. According to the findings obtained in the study, it is recommended that the MoNE should increase the number of halls and use other advantages provided by the computerized examination. The most important of these advantages is the inclusion of questions that determine the risk-taking disposition of individuals. It has been concluded that measuring this skill with internet or computer-based tests instead of traditional paper-and-pencil tests reduces both costs and increases the validity (Horswill & Coster, 2001; McKenna & Horswill, 1999). It is recommended to increase the size and quality of the ECODL item pool, to ensure the protection of confidentiality of items, and to carry out more studies on the transition to CAT application. Taking into consideration the opinions of individuals regarding CAT application, it is thought that CAT application can be started in ECODL theoretical exam with the arrangements made based on these research findings.

Acknowledgements

This research (Project # 215k018) was supported by the TUBITAK Project 1001 program in Turkey. The first draft of the paper presented at 6th International Congress on Measurement and Evaluation in Education and Psychology, Prizren, Kosova.

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).

ORCID

Nükhet ÇIKRIKÇI https://orcid.org/0000-0001-8853-4733
Seher YALÇIN https://orcid.org/0000-0003-0177-6727
İlker KALENDER https://orcid.org/0000-0003-1282-4149
Emrah GÜL https://orcid.org/0000-0001-8799-3356
Cansu AYAN https://orcid.org/0000-0002-0773-5486
Gizem UYUMAZ https://orcid.org/0000-0003-0792-2289
Merve Şahin KÜRŞAD https://orcid.org/0000-0002-6591-0705
Ömer KAMIŞ https://orcid.org/0000-0003-0605-087X

5. REFERENCES

AngloInfo, The Global Expat Network Berlin. (2015). *Taking a German driving test*. Retrieved May 24, 2017, from http://berlin.angloinfo.com/information/transport/driving-licences/the-driving-test

Balkız, C. (1999). *The effects of traffic education on the prevention of traffic accidents*. (Unpublished master thesis). Gazi University, Graduate School of Natural and Applied Sciences, Ankara.

Bozkurt, M (2003). *Evaluation of the first aid knowledge of driving nominees at the begining and at the end of driving course in Ankara*. (Unpublished master thesis). Gazi University, Graduate School of Natural and Applied Sciences, Ankara.

Bugbee, A. C. & Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982–1988. *Journal of Research on Computing in Education*, *23*(1), 87-100. https://doi.org/10.1080/08886504.1990.10781945

Burke, M. J., Normand, J., & Raju, N. S. (1987). Examinee attitudes toward computer-administered ability testing. *Computers in Human Behavior*, 3(2), 95-107. https://doi.org/10.1016/0747-5632(87)90015-X

Çakır, H. (2006). *Effectiveness of web and computer assisted instructions developed concerning dominant intelligence type on traffic education*. (Unpublished doctoral dissertation). Ankara University, Graduate School of Educational Science, Ankara.

Davey, T. (2011). *A guide to computer adaptive testing systems educational testing service for technical ıssues in large-scale assessment (TILSA).* State Collaborative on Assessment and Student Standards (SCASS). Retrieved May 11, 2018, from https://files.eric.ed.gov/fulltext/ED543317.pdf

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

Economides, A. A. & Roupas, C. (2007). Evaluation of computer adaptive testing systems. *International Journal of Web-Based Learning and Teaching Technologies*, *2*(1), 70–87. https://doi.org/10.4018/jwltt.2007010104

Elander, J., West, R., & French, D. (1993). Behavioral correlates of individual differences in road-traffic crash risk: An examination of methodology and findings. *Psychological Bulletin*, *113*(2), 279-294. Retrieved from https://www.researchgate.net/profile/Davina_French/publication/14748646_Behavioral_correlates_of_individual_differences_in_road-traffic_crash_risk_An_examination_of_methods_and_findings/links/00b49530310a4ec836000000/Behavioral-correlates-of-individual-differences-in-road-traffic-crash-risk-An-examination-of-methods-and-findings.pdf

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates.

Eroğlu, M. G. (2013). *Comparison of different test termination rules in terms of measurement precision and test length in computerized adaptive testing.* (Published doctoral dissertation). Hacettepe University, Graduate School of Educational Sciences, Ankara.

Georgiadou, E, Triantafillou, E., & Economides, A. A. (2006). Evaluation parameters for computer-adaptive testing. *British Journal of Educational Technology*, 37(2), 261-278. https://doi.org/10.1111/j.1467-8535.2005.00525.x

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8), 1-39. Retrieved from https://files.eric.ed.gov/fulltext/EJ838610.pdf

Gökçe, S. (2012). *Comparison of linear and adaptive versions of the Turkish pupil monitoring system (PMS) mathematics assessment.* (Unpublished doctoral dissertation). Middle East Technical University, The Graduate School of Natural and Applied Sciences, Ankara.

Gülecen, M. (1998). *The effect of driving courses in traffic education*. (Unpublished master thesis). Gazi University, Graduate School of Natural and Applied Sciences, Ankara.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Baston, MA: Kluwer Nijhoff Publishing.

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

Horswill, M. S., & Coster, M. E. (2001). User-controlled photographic animations, photograph-based questions, and questionnaires: Three internet-based instruments for measuring drivers' risk-taking behavior. *Behavior Research Methods, Instruments, & Computers*, 33(1), 46-58. https://doi.org/10.3758/BF03195346

İnal, K. (2001). *Traffic safety and assessment of driving education in Turkey*. (Unpublished master thesis). Gazi University, Graduate School of Natural and Applied Sciences, Ankara.

İşeri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures.* (Unpublished doctoral dissertation). Middle East Technical University, The Graduate School of Natural and Applied Sciences, Ankara.

Kaçmaz-Omak, S. (2012). *Awareness levels of primary school 5th-grade students about traffic information and traffic accidents*. (Unpublished master thesis). Istanbul University, Institute of Health Sciences, Istanbul.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability*. (Unpublished doctoral dissertation). Middle East Technical University, The Graduate School of Natural and Applied Sciences, Ankara.

Kaptan, F. (1993). *Comparison of adaptive (individualised) test application and traditional paper-pencil test application in estimation of ability*. (Unpublished doctoral dissertation). Hacettepe University, Graduate School of Social Sciences, Ankara.

Kernan, M. C., & Howard, G. S. (1990). Computer anxiety and computer attitudes: An investigation of construct and predictive validity issues. *Educational and psychological measurement*, 50(3), 681-690. https://doi.org/10.1177/0013164490503026

Kezer, F. (2013). *Comparison of computerized adaptive testing.* (Published doctoral dissertation). Ankara University, Graduate School of Educational Science, Ankara.

Kezer, F. & Koç, N. (2014). A comparison of computerized adaptive testing. *Journal of Educational Sciences Research*, *4*(1), 145–174.

Kim, J., & McLean, J. E. (1995, April). *The influence of examinee test taking motivation in computerized adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions, 14*(12), 1-5. https://doi.org/10.3352/jeehp.2017.14.12

Köklü, N. (1990). *A comparison of tailored test, which is developed according to classical test theory, and group test*. (Unpublished doctoral dissertation), Hacettepe University, Graduate School of Social Sciences, Ankara.

Kuyumcu, M. (2001) *The ideas of the personnel and instructors who participate in the in-seervices training activities related to the traffic arranged by Gendarme Schools Command*. (Unpublished master thesis). Hacettepe University, Graduate School of Social Sciences, Ankara.

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come* [MESA Memorandum No. 69]. Retrieved from https://www.cehd.umn.edu/EdPsych/C-Bas-R/Docs/Linacre2000_CAT.pdf

Linacre, J. M. (2006). Computer adaptive tests (CAT): Standard errors and stopping rules. *Rasch Measurement Transactions*, *20*(2), 1062. Retrieved from www.rasch.org/rmt/rmt202f.htm

Linden, W. & Glas, G. (2002). *Computerized adaptive testing: Theory and practice*. New York: Kluver Academic Pub.

McKenna, F. P., & Horswill, M. S. (1999). Hazard perception and its relevance for driver licensing. *Journal of the International Association of Traffic and Safety Sciences*, *23*(1), 36-41. Retrieved from https://trid.trb.org/Results?txtKeywords=McKenna&txtTitle=&txtSerial=%22IATSS%20Research%22&ddlSubject=&txtReportNum=&ddlTrisfile=&txtIndex=&specificTerms=&txtAgency=&txtAuthor=&ddlResultType=&chkFulltextOnly=&recordLanguage=&subjectLogic=or&dateStart=&dateEnd=&rangeType=emptyrange&sortBy=publisheddate&sortOrder=DESC&rpp=25#/View/492227

Ministry of Transport and Communication. (2015). *Driving in Finland: Licences, driving schools, rules & vehicles*. Retrieved January 21, 2017, from www.expat-finland.com/living_in_finland/driving.html

Miser, R. (1999). *Halk eğitimi ve toplum kalkınması (Public education and community development).* Ankara: MEB-Türk Tarih Kurumu Pub.

Mulkat-Ajibola, Y. (2015). Impact assessment of road traffic accidents on Nigerian economy. *Journal of Research in Humanities and Social Science*, *3*(12), 8-16. Retrieved from http://eprints.federalpolyilaro.edu.ng/615/1/YUSUFFM.A12.pdf

Nydick, S. & Weiss, D. (2009). A hybrid simulation *procedure for the development of CATs*. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved May 15, 2018, from www.psych.umn.edu/psylabs/CATCentral

Orcutt, V. L. (2002, February). *Computerized adaptive testing: Some issues in development*. Paper presented at the annual meeting of the Educational Research Exchange, University of North Texas, Denton, Texas.

Özbaşı, D. & Demirtaşlı, N. (2015). Development of computer literacy test as computerized adaptive testing. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 218–237. https://doi.org/10.21031/epod.79491

Öztuna D. (2012). *A computerized adaptive testing software (CAT): SmartCAT*. European Rasch Training Group (ERTG) Meeting, 17-19 April 2012, Leeds, UK.

Powers, D. E., & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment*, *1*(2), 153-173. https://doi.org/10.1207/s15326977ea0102_4

Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(4) 311–327. https://doi.org/10.1111/j.1745-3984.1998.tb00541.x

Rudner, L. M. (1998). *An online, interactive, computer adaptive testing tutorial*. 11/98. Retrieved April 24, 2017, from http://EdRes.org/scripts/cat

Schmidt, F. L., Urry, V. W., & Gugel, J. F. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. *Educational and Psychological Measurement*, *38*(2), 265–273. https://doi.org/10.1177/001316447803800208

Sutton, R. E. (1991). Equity and computers in the schools: A decade of research. *Review of Educational Research*, *61*(4), 475–503. https://doi.org/10.3102/00346543061004475

Tanrıkulu, S. (2002). Trafik kazalarının önlenmesi bağlamında trafik güvenliği eğitiminin rolü ve trafik kültürü. (The role of traffic safety education in the prevention of traffic accidents, and traffic culture). *Polis Bilimleri Dergisi*, *5*(1), 45-60.

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103-134). Mahwah (NJ): Lawrence Erlbaum.

Thompson, N., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*(1), 1–9. Retrieved from https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1242&context=pare

Turkish Statistical Institute [Türkiye İstatistik Kurumu (TÜİK)]. (2018). *Highway traffic accident statistics*. Retrieved November 27, 2018, from http://www.tuik.gov.tr/PreTablo.do?alt_id=1051

Türkoğlu, M. (2002). *The Effect of the training in the private driving schools on traffic accidents*. (Unpublished master thesis). Gazi University, Graduate School of Natural and Applied Sciences, Ankara.

Tütüncü, M. (2001). *The Importance of traffic education in formal education for Turkey*. (Unpublished master thesis). Gazi University, Graduate School of Natural and Applied Sciences, Ankara.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, *35*(4), 328-345. https://doi.org/10.1111/j.1745-3984.1998.tb00542.x

Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement*, *60*(3), 371-384. https://doi.org/10.1177/00131640021970600

Vursavaş, F. (2004). *Evaluation of the driving school curriculum*. (Unpublished master thesis). Ankara University, Graduate School of Educational Science, Ankara.

Wainer, H. & Mislevy, R. J. (1990). *Item response theory, item calibration and proficiency estimation*. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 1-21). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 70–84. https://doi.org/10.1080/07481756.2004.11909751

Weiss, D. J. & Guyer, R. (2012). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul MN: Assessment Systems Corporation.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wilder, G., Mackie, D., & Cooper, J. (1985). Gender and computers: Two surveys of computer-related attitudes. *Sex Roles*, 13(3), 215-228. Retrieved from https://link.springer.com/content/pdf/10.1007/BF00287912.pdf

Yaşar, M. (1999). *A study about individualized tests* [*Bireyselleştirilmiş testler üzerine bir çalışma*]. (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Young, K., Regan, M., & Hammer, M. (2007). Driver distraction: A review of the literature. In: I. J. Faulks, M. Regan, M. Stevenson, J. Brown, A. Porter & J. D. Irwin (Eds.), *Distracted driving (pp. 379-405)*. Sydney, NSW: Australasian College of Road Safety.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 588–606
https://doi.org/10.21449/ijate.682103
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

An Examination of Turkish Students' PISA 2015 Collaborative Problem-Solving Competencies

Emine Yavuz [1,*], Hakan Yavuz Atar [2]

[1] Erciyes University, Department of Educational Science, Kayseri, Turkey
[2] Gazi University, Department of Educational Science, Ankara, Turkey

ARTICLE HISTORY

Received: Jan. 30, 2020
Revised: July 21, 2020
Accepted: Sep. 25, 2020

KEYWORDS

PISA 2015,
Turkey,
Collaborative Problem-Solving,
Cognitive Diagnostic Model,
DINO

Abstract: Technological advancements initially made it possible for individuals with different cultures from different parts of the world to consider working together, and eventually collaboration turned into a necessity. Based on this necessity, OECD measured collaborative problem-solving (CPS) competencies of 15-year-old students from the countries participating in PISA assessment in 2015. The objective of this study is to examine the data on Turkey from PISA 2015 CPS survey through Deterministic-Input, Noisy-Or-Gate (DINO) cognitive diagnostic model and to determine the students' levels of CPS competencies as well as difficulty levels of these competencies. In this context, primarily, attributes of CPS competency domain were examined and the model with the most appropriate number of attributes, classification accuracy and consistency was determined for the analysis of PISA 2015 CPS data. The sample of this descriptive study consists of 435 students. As a result of the analysis, the Q-matrix of the model data fit was observed to have the best goodness-of-fit when it had three attributes whereas model fit and classification consistency decreased as the number of attributes increased. Furthermore, the easiest competency for Turkish students was determined to be 'taking appropriate action to solve the problem' while the most difficult was 'establishing and maintaining shared understanding'. Finally, latent class distribution exposed that 56% of the students did not have any of the competencies, and 35% had all. Results of this examination indicate that all CPS competencies of Turkish students need improvement.

1. INTRODUCTION

All individuals face several problems throughout their daily lives. It is of great importance that individuals have problem-solving competencies in overcoming these problems. To this end, it would be befitting to determine the strengths of individuals in problem-solving and to examine these aspects since they might provide an insight into the abilities of individuals to use basic thought processes and to face challenges in life (Lesh & Zawojewski, 2007:769).

Training individuals to have problem-solving competency is among the objectives of the education provided both by families and in schools as part of the government policies (Ministry of Education [MEB], 2009:6-16). (Organisation for Economic Co-operation and Development [OECD], 2013:13; Soylu & Soylu, 2006:97). Problem has a variety of definitions although none seems to have been agreed-upon. For instance, Turkish Language Association

CONTACT: Emive YAVUZ ✉ yavuzemine0@gmail.com Erciyes University, Department of Educational Science, Kayseri, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

(TDK, 2019) defines it as “a question, an issue or a matter that needs to be solved by theories and rules”. It was defined as situations that require mathematical operations by Mayer and Hegarty (1996:31) and as issues requiring qualitative or quantitative solutions by Krulik and Rudnick (1985:686). There are also researchers who define the problem as questions requiring higher-order thinking skills such as critical, creative and analytical thinking (Nancarrow, 2004:22; Schoenfeld, 1992:337). Thus, it may be asserted that a situation can be considered as a problem if it leads to an increase in mental activities of an individual (Baki, 2006:20). It is also possible to make an overall definition of problem as the difference between the condition at hand and the condition to be attained (Kneeland, 1999:5). The reason why the problems are described in different ways, as seen here, may be that they vary in their structure. Based on their structures, problems are generally divided into two as routine and non-routine ones (Dede & Yaman, 2006). In the literature, routine problems are expressed as single-solution problems, four-operation problems or well-structured problems. These problems can be rendered as analogous situations where individuals were able to find resolutions before or similar situations or issues in which they can make use of their previous experiences (gained knowledge) (Polya, 1973:171). Since these problems require the use of formulas, rules or methods that individuals have previously learned (Arslan & Altun, 2007:50), they are significant in terms of reinforcing what has already been learned. Non-routine problems, on the other hand, appear to be called ill-structured or multiple-solution problems in the literature. Because non-routine problems cannot be explained clearly and their solutions vary depending on the criteria (Frederiksen, 1984:367), they might have results varying from person to person. For this reason, the focus in problem-solving shifted from the answer of problem to the solution process (Altun, 2000). In addition, since these problems are not well-structured, it is not enough for people to have the necessary information in the fields related to the problem (Dede & Yaman, 2006:126). People are supposed to use their higher-order thinking skills as well (Altun, 2014:340-355; Hartman, 1998:1; Nancarrow, 2004:73).

Problem-solving may be defined as a cognitive process to transform a given condition into a target condition when there is no obvious method of solution (Mayer, 1990 as cited in OECD, 2017a:9). As problem solving has been an ongoing exercise for centuries and will continue to be so in the future, it is not easy to determine its framework and measure it from student performances (OECD, 2013:125). Also, in the literature, problem-solving process is approached in various ways. For example, Dewey (1933:107-116) studied the problem-solving process in five sub-processes: encountering a problem, understanding and identifying the problem, determining solutions, applying solutions, and reviewing results. Polya (1973:5-6) covered the problem-solving process as four sub-processes. These processes are understanding the problem, making plans for the solution, implementing the plan, and checking the result. As can be seen, the way Dewey (1933:107) and Polya (1973:5-6) handle problem solving processes are similar, and it is possible to recount more examples from the literature.

Upon recognition of its significance, problem-solving started to be an integral part of various educational programs to make students acquire this skill (National Council of Teachers of Mathematics [NCTM], 1989; MEB, 2009). Thus, the evaluation of the problem-solving competencies of the students by teachers, national exams and international exams has gathered speed. Programme for International Student Assessment (PISA) is one of the international exams in which students' problem-solving competencies are evaluated.

1.1. Evaluation of Problem-solving Skill in PISA

PISA is an international exam aiming to measure and evaluate the literacy of 15-year-old students in three basic domains (mathematics, science and reading) and competencies such as problem-solving, which is of common interest to all domains. PISA is concerned with the literacy capacities that students will need in adult life rather than curriculum and school knowledge (OECD, 1999:9). Therefore, PISA measures students' problem-solving skills in each key domain, as well as their performance of interdisciplinary problem-solving capacities (OECD, 2003:175). First implemented in 2000, this exam is repeated every three years and each time the aforementioned domains have a varying weight of examination. For instance, Reading in 2000, Mathematics in 2003 and Science in 2006 were focused on more closely than other fields, and this detailed examination continued in other surveys in the same order.

During the period between the first (2000) and the last (2018) survey of PISA, both the changes in the qualifications required for the business world and the development of measurement methods and technology have brought about some changes in the definition and measurement of this competency and problem types used in the process.

For example, in PISA 2003, 2006, and 2009 surveys, problem-solving is defined as “an individual’s capacity to use cognitive processes to confront and resolve real, cross-disciplinary situations where the solution path is not immediately obvious and where the literacy domains or curricular areas that might be applicable are not within a single domain of mathematics, science or reading (OECD, 2003:156; 2006:3; 2009:11).” In these surveys, routine and non-routine daily life problems were measured by a pen-and-paper test.

In 2012, with the development of technology and measurement methods, problem-solving assessment started to be applied computer-based in PISA 2012 and student interaction with the problem became one of the focal points of evaluation. As software development tools got better and network computers became widespread, students were able to interact with an increasing variety of problems (OECD, 2013:127). Thus, static and interactive problems started to be used in PISA 2012 survey. If the information about a problem solving given to the student prior to the procedure is complete, the problem is called static. However, if the information about the problem given to the student is incomplete, and if the student needs to make trial-error or discovery to reach additional information, it is an interactive problem (OECD, 2013:125). Some examples for the interactive problems can be remote control units, personal devices (such as cellphones), electronic appliances and vending machines, sports education, and animal husbandry. An interactive problem situation may also be dynamic. This means that situation may be altered by influences beyond the student’s control (OECD, 2013:125). For example, if no button is pressed for 10 seconds during a transaction, a ticket vending machine might reset. The evaluation of such interactive problems delivered on computers in an international survey took place for the first time in PISA 2012. (OECD, 2013:131). In addition, when interactive problem situations are included in this computer-based problem-solving assessment, it becomes possible to deliver more realistic, real-life scenarios (simulations) than pen-and-paper tests would allow.

PISA 2012 defines problem-solving competency as “an individual’s capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious. It includes the willingness to engage with such situations in order to achieve one’s potential as a constructive and reflective citizen.” The first sentence of this definition is almost the same as the first part of the definition for problem-solving assessment in PISA 2003. Even though the 2003 definition only has a cognitive dimension (OECD, 2003:156), the second part of the 2012 definition highlights the OECD’s definition of “competency” by stressing the cross-disciplinary structure (OECD, 2013:122). The main aspect that distinguishes the problem-solving assessment in 2012 from that of 2003 is not the definition of problem-solving competency, but the use of computers and the inclusion of interactive problems (OECD, 2013:122). Therefore, creative problem-solving competency was measured and evaluated in PISA 2012 survey (OECD, 2013:120).

Delivering problems via computer allows data collection and scoring of things such as the length, frequency, type, and order of the actions taken by the student while responding to items, or subsequent performance analysis of students in a way that would not be possible in pen-and-paper assessment (OECD, 2003:182). However, since the software required to measure collaboration skills had not been developed yet (OECD, 2013:120), only individual problem-solving competency was examined in PISA 2012 survey.

1.2. Evaluation of Collaborative Problem-solving in PISA

Technological advancements initially made it possible for individuals with different cultures from different parts of the world to consider working together, and eventually collaboration turned into a necessity (OECD, 2011:38). Hence, collaborative problem-solving (CPS) emerges as a vital and requisite competency in workforce and education (OECD, 2017a:3; b:3). In addition, this competency becomes even more significant as the workforce with such competency is a criterion for determining the position of countries in the future world economy (Yalçın, 2018:184).

In the literature, 21st century skills especially needed by the business world are identified in different ways by different circles. The generally accepted framework was presented by the Partnership for 21st Century Skills (P21). P21 framework lays out 21st century skills under three titles: life and career skills; information, media and technology skills; and finally learning and innovation skills. Learning and innovation skills are examined under four skills: communication; creativity and innovation; critical thinking and problem solving; and collaboration (P21, 2009). Considering the fact that individuals employ some critical thinking, (Soylu & Soylu, 2006:99; Şenşekerci & Bilgin, 2008:17) a little creativity and innovation skills (Bozkurt & Çakır, 2016:71; Öztürk, 2014:158; Schoenfeld, 1992:337; Soylu & Soylu, 2006:99) as well as communication skills while working in collaboration (Hmelo & Silver, 2004:244), collaborative problem-solving competency may be deemed as an umbrella term encompassing learning and innovation skills from 21st century skills. Therefore, it is significant for both individuals and countries to measure and evaluate this competency at an early age and to determine the aspects where individuals are lacking or advantageous in the components of this competency.

In the learning sciences, there was a big shift in the 1990s from "cooperative learning" to "collaborative learning" (Hesse, Care, Buder, Sassenberg & Griffin, 2015:38). This was because Dillenbourg, Baker, Blaye and O'Malley (1996:189-211) determined the difference between these two terms. Knowing the difference between these two terms in order to fully comprehend the measured CPS structure is thought to be essential. According to the work of Dillenbourg et al. (1996:205), a cooperative view is called an activity with the division of labor. In other words, in a cooperative study, while students can coordinate at certain parts of their activities, they tend to work in parallel. Many scientists have stated that cooperative learning cannot fully exploit the potential of a group and that people cannot demonstrate all social skills while working together (Hesse et al., 2015:38). This has resulted in a focus on collaborative learning. During collaborative learning, students organize their activities together to address a task or problem. Students' activities are entangled; their contributions are mutual, and a student's actions can be carried out or completed by others (Hesse et al., 2015:39). To measure this collaboration, the CPS was evaluated for the first time in PISA 2015 survey through virtual tasks that require participants to collaborate with virtual participants (agents/representatives) on the computer (Herborn, Stadler, Mustafić, & Greiff, 2018:2).

Collaboration has evident advantages over individual problem-solving. These advantages are: a productive labor division, multiple perspectives of information, sources and experiences of knowledge, and the quality of solutions and creativity promoted by other group members' ideas (OECD, 2017a:3). Nonetheless, collaboration can also bring about potential challenges to group members. These challenges are unequal or inefficient division of workload among group members, unfair contribution of some members, members prioritizing their goals rather than the group's, conflict among the group members that prevent developing creative solutions, and not being able to coordinate tasks effectively (OECD, 2017a:3). In order to avoid the negative challenges of collaboration, students were made to interact with simulated participants (agents/representatives) on the computer in PISA 2015 (Rosen & Tager, 2013:3). To take advantage of collaboration, there is a message section on one side of the screen where problem tasks are presented, so that students can communicate with virtual participants, and when students come to a wrong conclusion about the problem task, virtual participants provide various suggestions to make students continue their collaboration and problem-solving.

PISA 2015 defines collaborative problem-solving as follows: "Collaborative problem-solving competency is the capacity of an individual to effectively engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution and pooling their knowledge, skills and efforts to reach that solution." As it is seen, unlike other PISA surveys (2000, 2003, 2006, 2009, 2012), this definition emphasizes the active participation of the individual in the collaboration process.

When the evaluation frameworks of problem-solving processes in PISA surveys are examined, it is possible to observe how related they are to each other. For example, in PISA 2003, the problem-solving process was divided into six sub-processes: understanding the problem, characterizing the problem, representing/modeling the problem, solving, reflecting the solution, and communicating the problem (OECD, 2003:175). On the other hand, sub-processes in the evaluation of the problem-solving process in PISA 2012 were determined by considering the recent studies on complex and dynamic problem-solving, based on the problem-solving and reasoning studies of Polya's research and cognitive psychologists (1973) (OECD, 2013:126). These sub processes are composed of exploring and understanding, representing and formulating, planning and executing, monitoring and reflecting. Consisting of six sub-processes in PISA 2003 and reduced to four sub-processes by PISA 2012, the problem-solving structure was the source for the PISA 2015 CPS structure and also within the context of this study for the formation of the conditions, namely the number of attributes, to examine the performance of Deterministic-Input, Noisy-Or-Gate (DINO) model.

The framework for the evaluation of the collaborative problem-solving process in PISA 2015 survey includes definitions and theoretical structures based on research (such as computer-assisted collaborative work, individual problem solving, etc.) and practices in various fields where the skills related to CPS are evaluated. The framework also includes information collected from the PISA 2015 CPS assessment, including Assessment and Teaching of 21st-Century Skills (ATC21s), Programme for the International Assessment of Adult Competencies (PIAAC), Partnership for 21st-Century Skills (P21) and PISA 2012 individual problem-solving assessment (OECD, 2017a:5). In PISA 2015 survey, the CPS assessment framework consists of three core competencies: "Establishing and maintaining shared understanding" signifies group members' sharing their perspectives on problem states and determining a common vision. "Taking appropriate action to solve the problem" means determining types of collaborative problem-solving actions needed to solve the problem and executing them to reach the solution. Finally, "establishing and maintaining team organization" involves understanding one's role as an agent as well as other virtual participants', following the rules while activating one's role, observing group organization, and facilitating the changes needed to deal with communication issues and obstacles to optimize the problem and the performance.

Studies related to CPS in Turkey was found to concentrate on teaching cooperative problem-solving strategies in the learning environment (e.g., Gök & Sılay, 2008, 2009; Yazlık & Erdoğan, 2016) or problem-based teaching (e.g., Özgen & Pesen, 2010). With the progress of technology as well as assessment and evaluation methods since the initial studies, this competency has started to be examined in a way to obtain maximum information about the competency of the participant in computer-assisted practices (see OECD, 2017a:4; b:27). There are quite a few studies in Turkey, which focus on computer-assisted collaborative problem-solving competency (e.g., Arıcı, 2019; Uzunosmanoğlu, 2013). In one of these studies, Uzunosmanoğlu (2013) examined the participants' eye movements and the areas where their eyes were concentrated while solving problems in pairs, and as a result, it was found that the points that individuals in pairs with high CPS levels looked at overlapped more. Findings of this study also revealed that these pairs were more capable of predicting a point where the next action would take place, helping each other and forming a mutual understanding. In another study, Arıcı (2019) examined the variables related to the success of Turkish students in PISA 2015 CPS with mediation models in which direct and indirect relationships could be determined. Arıcı (2019) examined the relationship of variables with this competency by using students' CPS scores, and no research was conducted on the sub-competencies. Similarly, no study was encountered in which a detailed examination of the CPS competencies of Turkish students was undertaken. Considering the role of this competency domain in current and future life conditions, a detailed examination of Turkish students' PISA 2015 CPS competencies appears to be necessary. The objective of this study is to examine the data on Turkey from PISA 2015 CPS assessment through DINO cognitive diagnostic model and to determine the students' levels of CPS competencies as well as the difficulty levels of these competencies. In this context, primarily, attributes of CPS competency domain were examined and the model with the most appropriate number of attributes, classification accuracy and consistency was determined for the analysis of PISA 2015 CPS data.

In line with this purpose, following are the research questions of this study:

1. With how many Q-matrix attributes (3, 7, 11) does DINO model have the best model fit?
2. How does the number of attributes given in the Q-matrix affect the classification accuracy and consistency of the DINO model?
3. What are the mastery percentages of Turkish students' PISA 2015 CPS competencies?
4. What are the difficulty levels of competencies that make up PISA 2015 CPS domain?

In order to answer these research questions, the method section provides information about how the research was conducted; which forms of the PISA 2015 CPS survey were used; which model was selected by which criteria for data analysis; and also the validity and reliability. The findings obtained from data analysis are summarized in their relevant sections listed in the same order as the research questions. In the discussion, conclusion and suggestions section of the article, firstly, the findings of the study are discussed and then suggestions for the implementation and research are given in the light of the results of the study.

2. METHOD

2.1. Research Design

This article was designed as descriptive research, because objective of this study is to reveal the classification performance of DINO model when the Q-matrix has different number of attributes on PISA 2015 CPS data and the students' performance on the sub-competencies that constitute PISA 2015 CPS competency.

2.2. Population and Sample

PISA 2015 data includes information on 5895 Turkish students detailed in 66 forms (booklets). The population of the study consists of 5895 Turkish students participating in this survey. Not all the students participating in the survey in Turkey receive the same CPS form. For this reason, some forms had to be determined as research sample. An examination of items that make up the forms exposed that the number of CPS skills that some forms represent are higher than some others. Because Form 93 and Form 96 have more items that represent the structure, they were decided to be studied. Because these forms contain the same items, only the order of these items in the forms is different, so all the data in these forms were combined. Also, in order for content validity not to decrease, polytomously scored items were modified as dichotomously scored items without removing them. In order to encode the items in binary, option distributions were examined. The remaining options in the first 50% of the distribution were coded as "0", and the other options as "1". Following missing data extraction, data on 435 people consisting of 77 items were obtained. The sample of the study comprises these 435 people who answered Form 93 and Form 96 completely in PISA 2015 CPS survey.

2.3. Data Collection Tool

Six units were developed and used in PISA 2015 CPS evaluation (OECD, 2017a:18; b:66). A cluster was created by combining each two units. Units that take 30 minutes to answer are made up of many scenarios where students work on a problem. Eventually, data gathered from a total of 117 items that make up six units or three clusters are used to analyze and scale collaborative problem-solving performance. All students participating in the CPS assessment completed one or two clusters of CPS in their assessment.

There are no open-ended items in the CPS assessment. All items urged students to pick one of multiple choices in various ways to respond to group members or drag and drop icons into appropriate slots or click an option within the visual display area. Thus, the ability to collaborate was evaluated through student responses in their interactions with one or more virtual participants.

Within the framework of PISA 2015 survey, the structure of the CPS competency domain was represented by a cross matrix of individual problem-solving processes and collaboration processes, resulting in a skills matrix of 12 cells in total. This matrix is given in Table 1. As can be seen in Table 1, the CPS structure has been gathered under three basic competency areas (OECD, 2017a:12; b:50). Essential attributes to provide correct responses to the items on PISA 2015 CPS survey were determined and reported according to this table.

The Q-matrix used in DINO analysis of this study was created on the basis of this table and technical reports (see OECD, 2017a; b). There are no items measuring A3 cell in PISA 2015 CPS survey. Therefore, the Q-matrix was created from 11 attributes instead of 12, originally expected to be one of the study conditions. This situation constitutes a limitation for the study. In Table 2, the Q matrix with three different number of attributes is given for the first 20 items that constitute the study conditions. In matrices, rows indicate items, and columns indicate attributes. Names of attributes are given in the first row to show how attributes of matrices are determined. When the first rows of matrices are studied, it can be observed that attributes in the Q-matrixes were created with the symbols representing the cells in Table 1.

Table 1. *Matrix of collaborative problem-solving skills for PISA 2015 (OECD, 2017a, s.12; b, s.50)*

		Collaborative Problem-Solving Competencies		
		(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organisation
Problem-solving processes	(A) Exploring and understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve the problem
	(B) Representing and formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organisation (communication protocol/rules of engagement)
	(C) Planning and executing	(C1) Communicating with team members about the actions to be/being performed	(C2) Enacting plans	(C3) Following rules of engagement, (e.g. prompting other team members to perform their tasks)
	(D) Monitoring and reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organisation and roles

Table 2. *The Q matrix with three different number of attributes is given for the first 20 items that constitute the study conditions*

Q matrix with 3 attributes			Q matrix with 7 attributes							Q matrix with 11 attributes										
1	2	3	A	B	C	D	1	2	3	A1	A2	B1	B2	B3	C1	C2	C3	D1	D2	D3
0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1
0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0
0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0

2.4. Data Analysis

In the research, the data were arranged in SPSS 22.0 (IBM Corp., 2013) and analyzed by DINO model, one of the cognitive diagnostic models, using the R software CDM package (Robitzsch, Kiefer, George, & Uenlue, 2019).

2.4.1. *Cognitive Diagnostic Models (CDM)*

CDMs are restricted latent class models (Templin & Henson, 2006:290). In CDMs, students' responses to the items are associated with the attributes (skills) expected of the students through Q-matrices. Rows and columns of a Q matrix are populated with the information about items and attributes respectively. Table 3 demonstrates a Q-matrix with 3 items and 4 attributes. As can be seen from this matrix, students must master attribute 1 and 3 to provide correct responses for item 1, attribute 2 and 3 for item 2, and attribute 1 for item 3.

Table 3. *A Q-matrix with 3 items and 3 attributes*

Items	Attributes			
	Attribute 1	Attribute 2	Attribute 3	Attribute 4
Item 1	1	0	1	0
Item 2	0	1	1	0
Item 3	1	0	0	0

In this study, Q-matrices of PISA 2015 CPS data were created in accordance with PISA reports and presented under the heading "data collection tool" to make their creation process more comprehensible. Since individual contribution of attributes in the analysis of items in PISA reports were not specified, Q-matrixes were formed with attributes of two categories.

In the mathematical formulations of CDMs whose Q-matrix has two categories, the probability of individuals' being in these two categories is shown with υ_1 and υ_2. In general, the probability of individual r answering item i from latent class c correct is calculated as shown below (Rupp, Templin & Henson, 2010:113):

$$P(X_{ri}=1|c)=\pi_{ic} \tag{1}$$

In equation (1), X_{ri} indicates the score of individual r observed for item i. This equation may also be interpreted as the probability of all participants in class c answering item i correctly or the difficulty of item i for the participants in class c. In CDMs, class membership probability is generally formulated as below (Rupp et al., 2010:114):

$$P(X_r=x_r)=\sum_{c=1}^{C}\overbrace{\upsilon_c}^{\text{Structural}}\overbrace{\prod_{i=1}^{I}\pi_{ic}^{x_{ir}}(1-\pi_{ic})^{1-x_{ir}}}^{\text{Measurement}} \tag{2}$$

In equation (2), P(·) denotes probability, X_r and x_r are responses received from all participants; x_{ir} is response of individual r for item i, υ_c is membership probability of class c, π_{ic} is the probability of a correct response for item i by an individual from latent class c; and $\sum(\cdot)$ means sum of all latent classes from c=1 to C a. Π(·) denotes that probabilities of all items from i=1 to I will be multiplied. In the formula, the items represented by I can only take values of 0 and 1, and the formula refers to a set of Bernoulli random variables for the items of I. This provides joint probability or likelihood of a given response pattern for each latent class. Since the multiplication part of the formula is the component that connects the unobservable latent variables with observable data, this portion is called measurement component. The addition part is called the structural component of the model because it is used to describe the relationship between latent attribute variables.

Cognitive diagnostic models can be classified in different ways under different conditions. One of these classifications (compensatory and non-compensatory models) is the relationship between the attributes required for the correct response to items. In non-compensatory models, individuals must master all the attributes required by an item in order to produce a correct response for that item. On the other hand, in compensatory models, it is not necessary to have all the necessary attributes for the correct response of the item. In a study where preservice teachers' problem-solving process competencies were examined (Yavuz, 2014:55), it was found that they were able to solve problems even if they did not have all process competencies. For this reason, within the scope of this study, PISA 2015 CPS competency was examined via Deterministic-Input, Noisy-Or-Gate (DINO) model, which is a dichotomously scored compensatory model.

2.4.1.1. *Deterministic-Input, Noisy-Or-Gate (DINO)*

DINO, one of the compensatory models, increases the probability of a correct response of the item if only one of the attributes required for the correct response is mastered. In other words, although the students do not have all the necessary skills to answer the item correctly, they can still provide the correct answer. In the model, there are two parameters affecting the probability of a correct response for each given attribute: slipping (*s*) and guessing (*g*). The equation for the DINO model is as below (Rupp et al., 2010:132):

$$\omega_{ic} = 1 - \prod_{a=1}^{A} (1 - \alpha_{ca})^{q_{ia}} \qquad (3)$$

The first component of the equation (3) is latent response variable, ω_{ic}. It refers to the responses provided by individuals in class C for item i. Because this variable uses the disjunctive condensation rule that indicates whether the individual has at least one of the measured attributes, it constitutes the deterministic input part of the model. If an attribute a does not measure item i, then q_{ia}=0. If an item i can be measured by attribute a, then q_{ia}=1, regardless of whether α_{ca}=0 or α_{ca}=1. α_{ca}=1 means that individuals in latent class c possess at least one attribute. Thus, the presence of one attribute can compensate for the absence of others. In the DINO model, the probability of a correct response for item i by an individual in latent class c is calculated as follows:

$$\pi_{ic} = P(X_{ic}=1 \mid \omega_{ic}) = (1-s_i)^{\omega ic} g_i^{1-\omega ic} \qquad (4)$$

As can be seen in equation (4), the probability of a correct response for an item depends on the parameters of latent response variable (ω_{ic}), slipping (s), and guessing (g). Here, P is the probability, π_{ic} is the probability of a correct response for item i in latent class c by an individual, X_{ic} is the observed response for item i by students in class c, ω_{ic} is the latent response variable for item i in latent class c, ($1-s_i$) is the probability of not slipping for item i, and g is the probability of guessing for item i. Equations (5 and 6) for the slipping (*s*) and guessing (*g*) parameters of the DINO model are as follows:

$$g_i = P(X_{ic}=1 \mid \omega_{ic}=0) \qquad (5)$$

$$s_i = P(X_{ic}=0 \mid \omega_{ic}=1) \qquad (6)$$

The guessing parameter shows the probability of a score 1 when the individual has none of the measured attributes whereas slipping parameter stands for the probability of a score 0 when the individual has at least one of the measured attributes.

2.4.2. *Evaluation of model fit*

Evaluation of the model fit is the evaluation of the degree of fit between the predicted model and the observed data (DiBello, Roussos & Stout, 2007:988) and is a validity indicator. In CDMs, model fit can be determined through two different perspectives. One of these is to compare indexes of the model with certain criteria (absolute fit), and the other is to compare the model with other models (relative fit). Both model fit approaches were used to answer the research problem. To evaluate model fit within the scope of this study, MADcor and MADQ3 were examined as absolute fit index, AIC and BIC values as relative fit index. In addition, as an indicator of model fit, mean RMSEA values were examined to check whether item-model fit was achieved on item basis.

In R software, the CDM package produces absolute fit measurements by comparing the observed frequencies of the data with the estimated frequencies. MADcor examined within the scope of the study is calculated as the mean of the absolute difference between the correlations of observed and estimated pairs of items (DiBello, et al., 2007:989); MADQ3 is calculated as the mean of the absolute values of the Q3 statistics (Yen, 1984:125) indicating the correlations related to item errors. The criterion for evaluating MADQ3 (Ravand, 2016:8) and MADcor (DiBello et al., 2007:989) indexes is .05. Statistics smaller than .05 is evaluated as a good model fit.

AIC and BIC relative fit indexes are log-likelihood values, which is a maximum likelihood estimation. The smaller of these indexes belonging to different models, or the model closer to 0 is considered to fit the data better.

As an indicator of item-model fit, “item mean of RMSEA” is a chi-square value. It is calculated by comparing the observed and estimated values of the correct responses to the item by individuals in different latent classes (Kunina-Habenicht, Rupp & Wilhelm, 2009:67). An index smaller than .05 is considered as good fit, smaller than .1 is considered as medium fit, and greater than .1 is considered as poor fit.

After the evaluation of model fit in the study, the classification accuracy and consistency of the models were examined. Classification accuracy and consistency are the analyzed model’s indicators of reliability in assigning students to the correct attribute classes (Sinharay & Johnson, 2019). After detecting the model with good model fit, high classification accuracy and consistency, students' level of CPS competencies and difficulty levels of competencies were determined.

3. RESULTS / FINDINGS

3.1. With how many Q-matrix attributes (3, 7, 11) does DINO model have the best model fit?

The model fit indexes obtained in cases where the Q-matrix of DINO model has different number of attributes are given in Table 4. It can be seen that MADcor values are .047 under the condition of three and seven attributes and .05 with 11 attributes. Considering the evaluation criterion (.05) of this index, the model can be claimed to fit the data better under the condition of three and seven attributes. MADQ3 value seems to decrease as the number of attributes increases (.050, .049 and .049 respectively) and the model fits the data better. As the number of attributes increases in both AIC and BIC values, it is seen that the model-data fit is getting worse. In other words, the condition under which these indexes take the closest value to 0 is when the Q-matrix is expressed with 3 attributes. Examination of mean RMSEA reveals that item-model fit is achieved under the condition of 3 attributes. As a result, the best model-data fit is achieved when the Q-matrix is expressed with 3 attributes.

Table 4. *Item and model-fit index values under various conditions*

Attribute number	3 attributes	7 attributes	11 attributes	Assessment criteria
MADcor	0.047	0.047	0.050	0.050
MADQ3	0.050	0.049	0.049	0.050
AIC	38602.06	38756.24	42240.12	Almost 0
BIC	39258.19	39901.41	51209.96	Almost 0
Mean RMSEA	0.039	0.085	0.22	0.050

3.2. How does the number of attributes given in the Q-matrix affect the classification accuracy and consistency of the DINO model?

The values regarding the classification accuracy and consistency of the DINO model on the basis of the attributes given in the Q-matrix are listed in Table 5.

Table 5. *The classification accuracy and consistency values under various conditions*

	3 attributes	7 attributes	11 attributes	Assessment criteria
Classification accuracy	0.922	0.696	0.710	0.70
Classification consistency	0.905	0.720	0.608	0.70

Table 5 reveals that the classification accuracy values of the DINO model are between 0.696 and 0.922. The model has the lowest classification accuracy when the Q-matrix is formed with 7 attributes, and the highest classification accuracy when the Q-matrix has three attributes. It can also be seen in Table 5 that the classification consistency of the DINO model is between 0.608 and 0.905. Classification consistency of the DINO model decreases as the number of attributes given in the Q-matrix increases. In conclusion, the classification accuracy and consistency of the DINO model is found to be the best when the Q-matrix is composed of 3 attributes.

3.3. What are the mastery percentages of Turkish students' PISA 2015 CPS competencies?

As a result of the first and second research questions, it can be observed that the DINO model adapts better to the data and the classification performance is better when the Q-matrix has three attributes. Therefore, the answer to the third and fourth research questions are based on the situation where the Q-matrix is given with three attributes. These three attributes indicate PISA 2015 CPS competencies and are named as "establishing and maintaining shared understanding", "taking appropriate action to solve the problem", and "establishing and maintaining team organization". Students are divided into eight (2^3) latent classes in the Q-matrix with three attributes. The mastery percentages of Turkish students' PISA 2015 CPS competencies are given in Table 6.

Table 6. *The mastery percentages of Turkish students' PISA 2015 CPS competencies*

Skills	0	100	10	1	110	101	11	111
%	0.561	0.004	0.044	0.000	0.012	0.007	0.026	0.346
N	244	1	17	0	4	4	13	152

Latent class distribution in Table 6 discloses that most of the students in the study sample (P [0,0,0] = 0.56) did not master any attributes and students mastering all attributes constitute 35% of the sample (P [1,1,1] = 0.35). There are no students in P [0,0,1] latent class. In P [1,0,0], P [0,1,0], P [1,1,0], P [1,0,1] and P [0,1,1] latent classes, there are 0.4%, 4%, 1%, 1% and 3% of them respectively. It is evident in Table 6 that the number of students in these latent classes is almost nonexistent.

3.4. What are the difficulty levels of competencies that make up PISA 2015 CPS domain?

The difficulty levels of competencies that make up PISA 2015 CPS domain are given in Table 7.

Table 7. *The difficulty levels of competencies that make up PISA 2015 CPS*

Competencies	*p*
Establishing and maintaining shared understanding	0.369
Taking appropriate action to solve the problem	0.428
Establishing and maintaining team organization	0.379

When Table 7 is examined, it is seen that the competency of "taking appropriate action to solve the problem" is easier for students than the other competencies and its difficulty level is 0.43. The difficulty level for the competency of "establishing and maintaining team organization" is 0.38 and 0.37 for "establishing and maintaining shared understanding". In addition, it can be inferred from Table 7 that the most difficult field of competency for students is "establishing and maintaining shared understanding".

4. DISCUSSION, CONCLUSION and SUGGESTIONS

Today, individuals need to be competent problem solvers to be constructive and reflective citizens (OECD, 2013:121). Moreover, since it will contribute to the success in their lives, individuals should master problem-solving competency in all educational institutions and levels (Dede & Yaman, 2006:126). With the development of technology, more and more individuals started to work together, and this required them to develop not only individual problem-solving skills but also collaborative problem-solving competencies. Collaborative problem-solving competency may be defined as finding solutions for a problem after working together and exchanging ideas. The objective of this study is to examine the data on Turkey from PISA 2015 CPS survey through DINO cognitive diagnostic model and to determine the students' levels of CPS competencies as well as the difficulty levels of the competencies themselves. In this context, primarily, attributes of CPS competency domain were examined and the model with the most appropriate number of attributes, classification accuracy and consistency was determined for the analysis of PISA 2015 CPS data. Since all students participating in the PISA survey did not receive the same forms, the study was carried out with the data from Form 93 and Form 96, considering the content and construct validity. The results of the study were discussed with this limitation in mind.

Considering the CPS structure laid out in the PISA 2015 technical reports, it was decided to form the Q-matrix with 3, 7 and 11 attributes. Upon an analysis of DINO model, the Q-matrix with three attributes was found to provide the best model fit. These three attributes making up PISA 2015 CPS competencies are named as "establishing and maintaining shared understanding", "taking appropriate action to solve the problem", and "establishing and maintaining team organization".

It was demonstrated that the classification consistency of the DINO model decreased inversely with the number of attributes in the Q-matrix, similar to the findings of Cui, Gierl, and Chang (2012) and Cui, Gierl, and Guo (2015). This reveals the relationship of classification consistency with the number of attributes. In CDMs, if the total number of items remains the same and the number of items that measure an attribute increases, the increase in the complexity of the Q-matrix can cause a decrease in the classification efficiency of the model (Madison & Bradshaw, 2015:509). Of the Q-matrices created within the scope of the present study, Q-matrices with 3 and 11 attributes have a simple structure, that is, a single attribute is required for each item to be responded correctly. Seven-attribute Q-matrices are complex and each item is measured with more than one attribute. Therefore, the attributes in seven-attribute Q-matrixes were measured more than those attributes in the other Q-matrix, resulting in a more complex Q-matrix structure. In the study, it was determined that the classification accuracy of the model was lower in the seven-attribute Q-matrices than in those with eleven attributes. The reason for this is thought to be the complex Q-matrix structure as mentioned in the studies of Madison and Bradshaw (2015:509).

It was discovered that more than half of Turkish students (56%) did not have any of the three competencies that constitute the CPS structure. 35% of students was found to have all three CPS competencies. There were almost no students who had one or two of these competencies. Furthermore, there were no students who only had the third competency. These findings seem to be in line with the performance of Turkey, which was ranked last among OECD countries (OECD, 2017b:33), in PISA 2015 CPS domain. Özkan and Öner (2019) conducted an experimental study with Turkish students on the geometric thinking development of secondary school students in computer-based learning environments. The researchers, who classified students as successful and unsuccessful based on their geometric thinking development, observed all PISA 2015 CPS competencies in the behaviors of group members whereas they uncovered that members of unsuccessful groups displayed less behaviors within the competencies of “establishing and maintaining shared understanding” and “establishing and maintaining team organization”. Unlike this study, Li and Liu (2017) discussed and analyzed the CPS structure in a 12-skill matrix. They found that the Taiwanese students had high competency especially in the skills of “describing roles and team organization (B3)” and “communicating with team members about the actions to be performed (C1)”, but they had low competency in the skill of “building a shared representation and negotiating the meaning of the problem (B1)”.

Moreover, the easiest competency for Turkish students was determined to be “taking appropriate action to solve the problem” while the most difficult one was “establishing and maintaining shared understanding”. This finding seems to support the findings of Özkan and Öner (2019). Chang et al. (2017), in their study with Taiwanese students of 11th grade, detected that students displayed similar behaviors in "establishing and maintaining a shared understanding" and "establishing and maintaining a team organization" but more so than in “taking appropriate action to solve the problem”. In other words, the competency of “taking appropriate action to solve the problem” was challenging for Taiwanese students. Similar to Chang et al. (2017), Kuo et al. (2019) determined that the most difficult competency for Taiwanese students was “taking appropriate action to solve the problem”. Researchers also highlighted that the competency of “establishing and maintaining shared understanding” was easier for Taiwanese students than the other two. Although the CPS competencies might be found easy or difficult depending on the countries where students are from, it can be asserted that this competency needs to be improved, considering the fact that students can increase their learning in the environments where they can easily share and discuss their opinions (Kutluca, 2013:1517).

In the current study, it was discovered that Turkish students had almost no competency in “establishing and maintaining shared understanding”, “taking appropriate action to solve the problem”, and “establishing and maintaining team organization”. However, studies conducted in recent years offer some optimism for improving CPS skills. In their study with Malaysian students, Nordin and Osman (2018) found that the module they used improved all CPS skills. Lin, Yu, Hsiao, Chang, and Chien (2018) conducted a semi-experimental study with Taiwanese students. In their study, the researchers tried to improve the CPS competencies of three groups of students who received STEM education with teacher-guided web-based applications in the first group, with only web-based applications in the second group and only in-class activities in the third group. In the end, they detected that the CPS competencies of the first and second group members increased in general, and the CPS competency of "establishing and maintaining a shared understanding" was more developed than the other two. In this regard, it is possible to think that the CPS competencies of Turkish students can be improved with various applications. For this reason, there is a need for various projects or experimental research on how to develop CPS competencies.

Finally, although there are many CPS studies focusing on it as a 21st century skill and exploring its significance, measurement and evaluation, there seems to be no applications or courses aiming to teach it within the Ministry of Education (MEB) context. Also, thanks to Fatih project, it is possible to provide computer-assisted education more comfortably today. From this point of view, the collaborative problem-solving competency can be incorporated as a course into the curricula of schools by MEB, whether it is computer-assisted or not. Similarly, OECD (2013:121) states that problem-solving competency can be improved through a process of high-quality education.

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).

ORCID

Emine YAVUZ https://orcid.org/0000-0002-1991-1416
Hakan Yavuz ATAR https://orcid.org/0000-0001-5372-1926

5. REFERENCES

Altun, M. (2000). İlköğretimde problem çözme öğretimi [Teaching problem solving in primary education], *Journal of Education and Social Sciences*, 147. Retrieved July 10, 2019, from http://dhgm.meb.gov.tr/yayimlar/dergiler/Milli_Egitim_Dergisi/147/altun.htm

Altun, M. (2014). *Eğitim Fakülteleri ve İlköğretim Matematik Öğretmenleri için Matematik Öğretimi* [Teaching Mathematics for Faculty of Education and Primary School Mathematics Teachers], 18th ed.; Bursa: Alfa Aktüel Publication.

Arıcı, Ö. (2019). *Investigating the factors related to Turkish students' collaborative problem solving skills with mediating models according to PISA 2015 results*. Doctoral Dissertation, University of Ankara at Ankara.

Arslan, Ç., Altun, M. (2007). Learning to solve non-routine problems. *Elementary Education Online*, *6*(1), 50-61.

Baki, A. (2006). *Kuramdan Uygulamaya Matematik Eğitimi* [Mathematics Education from Theory to Practice]. İstanbul: Bilge Printing.

Bozkurt, Ş. B., Çakır, H. (2016). 21st Century learner skills: An investigation of middle school students based on grade level and gender. *Pamukkale University Journal of Education*, 39, 69-82.

Chang, C.-J., Chang, M.-H., Chiu, B.-C., Liu, C.-C., Chiang, S.-H. F., Wen, C.-T., et al. (2017) An analysis of student collaborative problem solving activities mediated by collaborative simulations. *Computers & Education*, 114, 222-235. https://doi.org/10.1016/j.compedu.2017.07.008

Cui, Y., Gierl, M. J., Chang, H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2011.00158.x

Cui, Y., Gierl, M., Guo, Q. (2015). Statistical classification for cognitive diagnostic assessment: An artificial neural network approach. *Educational Psychology: An International Journal of Experimental Educational Psychology*, 1-24. https://doi.org/10.1080/01443410.2015.1062078

Dede Y., Yaman S. (2006). Fen ve matematik eğitiminde problem çözme: Kuramsal bir çalışma [Problem solving in science and mathematics education: A theoretical study]. *Çukurova University Faculty of Education Journal, 2*, 116-128.

Dewey, J. (1933). *How We Think: A Restatement of the Relation of Reflective Thinking to The Educative Process*. Boston: DC Heath and Company.

DiBello, L. V., Roussos, L. A., Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of Statistics*, *26*, 979–1030. https://doi.org/10.1016/S0169-7161(06)26031-0

Dillenbourg, P., Baker, M., Blaye, A., O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada, P. Reiman (Eds.), *Learning in Humans and Machines: Towards an Interdisciplinary Learning Science (pp. 189–211)*. Oxford: Elsevier.

Frederiksen, N. (1984). Implications of cognitive theory for instruction in problem solving. *Review of Educational Research 54*(3), 363–407. https://doi.org/10.1002/j.2330-8516.1983.tb00019.x

Gök, T., Sılay, İ.(2008). The effects of problem-solving strategies on students' achievement, achievement motivation and attitude in the cooperative learning groups in physics teaching. *Hacettepe University Journal of Education*, 34, 116-126.

Gök, T., Sılay, İ. (2009). The effects of problem-solving strategies teaching on students' achievement motivation, in the cooperative learning groups. *Kastamonu Education Journal*, *17*(3), 821-834.

Hartman, H. J. (1998). Metacognition in teaching and learning: An introduction. *Instructional Science*, *26*, 1-3. https://doi.org/10.1023/A:1003023628307

Herborn, K., Stadler, M., Mustafić, M., Greiff, S. (2018). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, *87*, 1-31. https://doi.org/10.1016/j.chb.2018.07.035

Hesse, F., Care, E., Buder, J., Sassenberg, K., Griffin, P. (2015). A framework for teachable collaborative problem-solving skills. In P. Griffin, E. Care (Eds.), *Assessment and Teaching of 21st Century Skills: Methods and Approach (pp. 37-56)*. Dordrecht: Springer.

Hmelo C., E. Silver (2004). Problem based learning; what and how do students learn? *Educational Psychology Review, 16*(39), 235-263. https://doi.org/10.1023/B:EDPR.0000034022.16470.f3

IBM Corp. (2013). *IBM SPSS Statistics for Windows*. Version 22.0. Armonk, NY: IBM Corp.

Kneeland, S. (1999). *Effective Problem-Solving: How to Understand the Process and Practise it Successfully*. London: Little, Brown Book Group.

Krulik, S., Rudnick, J. A. (1985). Developing problem solving skills. *Mathematics Teacher*, *79*(9), 685-692.

Kunina-Habenicht, O., Rupp, A. A., Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, 35, 64–70. https://doi.org/10.1016/j.stueduc.2009.10.003

Kuo, B.-C., Liao, C.-H., Pai, K.-C., Shih, S.-C., Li, C.-H., Mok, M. M. C. (2019). Computer-based collaborative problem-solving assessment in Taiwan. *Educational Psychology*, *39*(1), 1-22. https://doi.org/10.1080/01443410.2018.1549317

Kutluca, T. (2013). The effect of geometry instruction with dynamic geometry software; GeoGebra on Van Hiele geometry understanding levels of students. *Educational Research and Reviews, 8*(17), 1509-1518. https://doi.org/10.5897/ERR2013.1554

Lesh, R., Zawojewski, J.S. (2007). Problem solving and modeling. In F. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning (pp. 763-804)*. USA: Information Age Publishing.

Li, C.-H., Liu, Z.-Y. (2017). Collaborative problem-solving behavior of 15-year-old taiwanese students in science education. *EURASIA Journal of Mathematics Science and Technology Education*, *13*(10), 6677-6695. https://doi.org/10.12973/ejmste/78189

Lin, K-Y., Yu, K-C., Hsiao, H-S., Chang, Y-S., & Chien, Y-H. (2018). Effects of web-based versus classroom-based STEM learning environments on the development of collaborative problem-solving skills in junior high school students. *International Journal of Technology and Design Education*, 1-14. https://doi.org/10.1007/s10798-018-9488-6

Madison, M. J., Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3) 491–511. https://doi.org/10.1177/0013164414539162

Mayer, R.E. (1990). Problem solving. In M. W. Eysenck (Ed.), *The Blackwell Dictionary of Cognitive Psychology (pp.284-288)*. Oxford: Basil Blackwell.

Mayer, R. E., Hegarty, M. (1996). The process of understanding mathematical problems. In R.J. Sternberg, T. Ben-Zeev (Eds.), *The Nature of Mathematical Thinking (pp.29-54)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ministry of Education (2009). *İlköğretim Matematik Dersi 6-8. Sınıflar Öğretim Programı ve Klavuzu* [Elementary Mathematics 6-8th Classes Curriculum and Guide]. Ankara. Retrieved Jun 10, 2019, from http://mimoza.marmara.edu.tr/~apusmaz/dosyalar/6-8_Program%C4%B1.pdf

Nancarrow, M. (2004). *Exploration of metacognition and non-routine problem based mathematics instruction on undergraduate student problem solving success*. Doctoral Dissertation, Florida State University, Florida.

National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and Evaluation Standards for School Mathematics*. Virginia, Reston: NCTM Inc. Retrieved Jun 10, 2019, from http://csmc.missouri.edu/PDFS/CCM/summaries/standards_summary.pdf

Nordin, N.M., Osman, K. (2018). Students' generated animation: An innovative approach to inculcate collaborative problem solving (CPS) skills in learning physics. *Journal of Education in Science, Environment and Health (JESEH)*, *4*(2), 206-226. https://doi.org/10.21891/jeseh.436758

OECD, (1999). *PISA 2000 Measuring student knowledge and skills: A new framework for assessment*, OECD Publishing, Paris. Retrieved May 1, 2019, from https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/33693997.pdf

OECD, (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*, OECD Publishing, Paris. Retrieved May 1, 2019, from https://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/33694881.pdf

OECD, (2006). *Assessing Scientific, Reading and Mathematical Literacy: A framework for PISA 2006*, OECD Publishing, Paris. Retrieved May 1, 2019, from https://www.sel-gipes.com/uploads/1/2/3/3/12332890/assessing_scientific_reading_and_mathematical_literacy.pdf

OECD, (2009). *PISA 2009 assessment framework – Key competencies in reading, mathematics and science*, OECD Publishing, Paris. Retrieved May 1, 2019, from https://www.oecd.org/pisa/pisaproducts/44455820.pdf

OECD (2011). *Skills for Innovation and Research*, OECD Publishing, Paris. https://doi.org/*10.1787/9789264097490-en*

OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing, Paris. https://doi.org/*10.1787/9789264190511-en*

OECD (2017a). *PISA 2015 collaborative problem-solving framework*, OECD Publishing, Paris. Retrieved May 1, 2019, from https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

OECD (2017b). *PISA 2015 Results (Volume V): Collaborative Problem Solving*, OECD Publishing, Paris. https://doi.org/10.1787/9789264285521-en

Özgen, K., Pesen, C. (2010). Probleme dayalı öğrenme (PDÖ) yaklaşımı ile işlenen matematik dersinde öğrencilerin problem çözme becerilerinin analizi [Analysis of students' problem solving skills in mathematics course processed by problem based learning (PBL) approach]. *Journal of Education and Social Sciences*, 186, 27-37.

Özkan, E., Öner D. (2019). Investigation of the development of van Hiele levels of geometric thinking in a computer supported collaborative learning (CSCL) environment. *Mersin University Journal of the Faculty of Education, 15*(2), 473-490. https://doi.org/10.17860/mersinefd.522491

Öztürk, E. (2001). Yaratıcılık ve eğitim [Creativity and education]. *Sakarya University Journal of Education Faculty,* 1, 158-164.

Partnership for 21st Century Skills (2009). *P21 Framework definition*. Retrieved July 7, 2019, from https://files.eric.ed.gov/fulltext/ED519462.pdf

Polya, G. (1973). *How to Solve It?* USA: Princeton University.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*(8), 1-18. https://doi.org/10.1177/0734282915623053

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2019, September). CDM [software package in R]. Available online at https://cran.r-project.org/web/packages/CDM/index.html

Rosen, Y., Tager, M. (2013). *Computer-based assessment of collaborative problem-solving skills: Human-to-agent versus human-to-human approach*. Boston: Pearson Education Retrieved July 7, 2019, from https://www.researchgate.net/publication/258629038_Computer-based_assessment_of_collaborative_problem-solving_skills_Human-to-agent_versus_human-to-human_approach_Boston_MA_Pearson_Education

Rupp, A. A., Templin, J., Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications.* New York: The Guilford.

Sinharay, S., Johnson, M. (2019). Measures of agreement: Reliability, classification accuracy, and classification consistency. In M. von Davier, Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models (pp.359-377).* New York: Springer.

Soylu, Y., Soylu, C. (2006). The role of problem solving in mathematics lessons for success. *İnönü University Journal of the Faculty of Education*, *7*(11), 97-111. Retrieved July 7, 2019, from https://www.pegem.net/dosyalar/dokuman/8298-2011062915226-soylusoylu.pdf

Şenşekerci, E., Bilgin, A. (2008). Critical Thinking and its Teaching. *Uludağ University Faculty of Arts and Sciences Journal of Social Sciences, 9*(14), 15-43.

Templin, J., Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods,* 11, 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Turkish Language Association (2019). Türk Dil Kurumu sözlükleri [Dictionary of Turkish Language Association]. Retrieved July 7, 2019, from http://sozluk.gov.tr/

Uzunosmanoğlu, S. D. (2013). *Examining computer supported collaborative problem solving processes using the dual-eye tracking paradigm.* Master Dissertation, Middle East Technical University at Ankara.

Yalçın, S. (2018). 21st century skills and tools and approaches that are used to measure these skills. *Ankara University Journal of Faculty of Educational Sciences, 51*(1), 183-201.

Yavuz, E. (2014). *Determining the problem solving process skills of the primary education preservice mathematics teachers as defined in PISA.* Master Dissertation, Gazi University at Ankara.

Yazlık, D., Ö., Erdoğan, A. (2016). The effects of problem-solving strategies used in combinations with cooperative learning on learner achievement. *Ahi Evran University Journal of Kırşehir Education Faculty (KEFAD), 17*(3), 1-16.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement,* 8, 125-145.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 607–630

https://doi.org/10.21449/ijate.738560

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

A Systematic Review of Research Articles on Measurement Invariance in Education and Psychology

Betul Alatli [1,*]

[1]Department of Educational Sciences, Faculty of Education, Balıkesir University, Balıkesir, Turkey

ARTICLE HISTORY

Received: May 17, 2020
Revised: Aug. 28, 2020
Accepted: Sep. 25, 2020

KEYWORDS

Measurement invariance,
Confirmatory Factor Analysis,
Content analysis,
Bibliometric analysis

Abstract: This study aims to reveal the trends in the related field by examining the researches evaluating the measurement invariance in education and psychology between 2008-2019. Accordingly, 99 articles published in three journals that were selected using the purposive sampling method among the journals indexed on Social Sciences Citation Index (SSCI)were analyzed within the scope of the study. As a result of the content analysis, in the studies investigating the measurement invariance, typical response tests were observed to be the most frequently employed tests, sample sizes often included 1501 or greater number of subjects, and data were mostly collected from students. The measurement invariance of the tests was mostly analyzed in terms of the gender variable. According to the results of the bibliometric analysis, on the other hand, only Multi-Group Confirmatory Factor Analysis was mostly conducted on the Mplus software package. In the studies, the most cited article was "Cheung and Rensvold (2002)", the author was "Cheung, G. W.", and the journal was "Structural Equation Modeling: A Multidisciplinary Journal". According to the results of the analysis, studies, references, and keywords including factor analysis were among the most commonly used group, which denotes that factor analysis has a crucial role in invariance measurement analyses.

1. INTRODUCTION

Science is defined as a whole consisting of systematic information; yet, the validity of this information should be accepted (Karasar, 2016). Accordingly, the validity of the information must be recognized so that it can be evaluated scientifically. In order for the validity of the information to be accepted, measurement results are needed. The branches of science have two dimensions: theoretical and experimental. In the theoretical dimension, facts and factual relationships are conceptually explained. Experimental/observational studies make it possible to observe the phenomenon in question or relationships under suitable conditions and to quantify or qualify the results of observation. The two dimensions have intertwined processes. In other words, when making theoretical explanations according to the results of an observation, we need observation results again to confirm these theoretical explanations. Precisely at this point, science introduces the importance of measurement. The branches of science can maintain their development in parallel with the development of measurement processes. Measurement is the process of observing a characteristic and displaying the results of observation using symbols

CONTACT: Betül ALATLI ✉ betulkarakocalatli@gmail.com Department of Educational Sciences, Faculty of Education, Balıkesir University, Balıkesir, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

or numbers (Turgut, 1977). Measurement is the process of determining whether an individual, an object, or a phenomenon has certain characteristics, and showing the degree of the feature using numbers or symbols if the property sought is available (Tekin, 2012). A measurement tool is as important as the measured feature and the thing measured. The validity and reliability of the measurement results are directly correlated with the validity and reliability of the measurement tool. For this reason, it is considered quite important for science that measurement tools can make valid and reliable measurements.

Measurement tools vary by the fields of science. Fields of science are classified into three basic fields: natural sciences, social sciences, and mathematics. Mathematics is absolute, but natural, and social sciences have a relative nature. While social sciences focus on social phenomena, natural sciences address natural events. The desire of human beings to have knowledge and skills has come out to understand and control primarily their environment and then themselves. For this reason, natural sciences date back to much older times than social sciences. Social sciences concentrate human and human behaviors and interactions (Karasar, 2016). Therefore, compared to natural sciences, the difficulty in making and controlling objective observations in social sciences have a significant impact on the measurements made in this field.

Measurements made in social sciences often involve human characteristics. Measurement tools used in these types of measurements are generally called a test. Some features can also be measured with non-test techniques. While it is more convenient to observe some features with non-test techniques, others can be observed with tests. Tests are measurement tools that consist of stimuli (items) for a certain characteristic to be measured. The status of the individual regarding this characteristic is determined by the response shown to theseitems. The extent to which the tests or the results obtained from these tests serve the purpose and the error-free nature of the tests are highly important in making decisions based on these results. Special studies are required to obtain evidence for such features called validity and reliability. One of these studies is the measurement invariance. Measurement invariance is defined as a condition where individuals in different groups who have had the same observed score in terms of a specific implicit structure get the same score at the subscale and item levels. According to the statements in the Test Adaptation Guidelines (International Test Commission-ITC, 2005) and the Standards of Measurement in Education and Psychology (American Educational Research Association-AERA, American Psychological Association-APA, and National Council on Measurement in Education-NCME, 1999), evidence of measurement invariance must be obtained for tests which aim to make intergroup comparisons. Accordingly, to make decisions about individuals and the groups that they belong to for many features and to make comparisons between individuals and groups in social sciences, measurement invariance analyses are considered to be highly important in making fair and appropriate decisions.

Although there are primarily invariance analyses at the test level under the name of measurement invariance, studies for determining the Differential Item Functioning also aim to determine the measurement invariance at item level (Holland &Wainer, 1993). As in many study areas, measurement invariance studies, too, can vary in different aspects. Many variables such as the test under investigation, measured feature, study group, and statistical technique and statistical software used increase the variety of measurement invariance studies. In this sense, determining the trends in the field by reviewing the measurement invariance studies is considered important and necessary.

New developments in a particular field of study can be followed by reviewing scientific studies such as projects, theses, and articles obtained as a result of a literature review. For this purpose, periodical literature reviews help determine trends in a given field and guide new studies (Chang, Chang & Tseng, 2010; Cohen, Manion & Morrison, 2007; Falkingham& Reeves, 1998; Keselman et al., 1998; Kilbourne & Beckmann, 1998; Lee, Wu, & Tsai; 2009). In Turkey, the number of educational researches is known to show a huge increase after the 2000s (Karadağ, 2010; Göktaş et al., 2012; Vega Arce et al., 2019). To reveal the quality of educational research, information about the quality and quantity of studies should be questioned (Bacanak, Değirmenci, Karamustafaoğlu & Karamustafaoğlu, 2011; Fazlıoğulları & Kurul, 2012).

When the literature is reviewed, it can be seen that studies conducted in Turkey for reviewing the literature in the field of education consist of many review studies on Science Education (Arıcı, Yıldırım , Çalıklar & Yılmaz, 2019, Chang, Chang & Tseng, 2010; de Jong 2007; Lin, Lin & Tsai, 2014; Ören & Sarı, 2019; Sırakaya & Alsancak Sırakaya, 2020; Sözbilir & Kutu, 2008; Lee, Wu & Tsai, 2009; Tsai & Wen, 2005; White, 1997), Mathematics Education (Aztekin& Taşpınar Şener, 2015, Baki, Karataş, Akkan & Çakıroğlu, 2011; Çiltaş, Güler & Sözbilir, 2012; Hart, Smith, Swars & Smith, 2009; Ulutaş &Ubuz, 2008), Social StudiesEducation (Tarman, Güven & Aktaşlı, 2011) Pre-School Education (Yılmaz & Altınkurt, 2012), chemistry education (Eybe & Schmidt, 2001; Ulutaş, Üner, Turan Oluk, Yalçın Çelik & Akkuş, 2015) specialeducation (Aslan &Özkubat, 2019), Classroom Teacher Education (Küçükoğlu & Ozan, 2013; Akaydın & Çeçen, 2015), EducationalSciences (Arık & Türkmen, 2009; Doğan & Tok, 2018; Erdem, 2011; Erdem Aydın, Kaya, İşkol&İşcan, 2019; Hsu, 2005; Karadağ, 2009; Selçuk, Kandemir, Palancı & Dündar, 2014; Tavşancıl et al., 2010).PsychologicalCounseling and Guidance (Seçer, Ay, Ozan & Yılmaz, 2006), Curriculum and Instruction (Saracaloğlu & Dursun, 2010; Hazır Bıkmaz, Aksoy, Tatar & Atak Altınyüzük, 2013; Ozan & Köse, 2014), Educational Administration (Aydın, Erdağ & Sarıer, 2010; Aypayet al., 2010; Turan, Karadağ, Bektaş & Yalçın, 2014; Murphy, Vriesenga & Storey, 2007), Educational Technology (Bozkutet al., 2015; Erdem Aydın, Bozkaya& Genç Kumtepe, 2019; Erdoğmuş & Çağıltay, 2009; Göktaş et al., 2012; Gülbahar & Alper, 2009; Hew, Kale & Kim, 2007; Özyurt & Özyurt, 2015; Zainuddin,et al., 2019).

Studies that investigate the relevant literature can be conducted to determine trends in the field while the review of studies on a particular topic in the field allows a detailed examination of that area. One example of this situation involves studies on leadership (Özkan, 2016) or special education in early childhood in Turkey (Öncül, 2014). The field of study of this research is measurement and evaluation in psychology and education. According to the content analysis studies in journals of education, the field of measurement and evaluation ranks fourth and sixth in journals overseas and Turkey, respectively (Hsu, 2005; Selçuk et al., 2014; Yalçın, Yavuz ve İlgün Dibek, 2015). Accordingly, it is necessary to increase studies on research trends by conducting content analysis studies in an important field of educational sciences such as measurement and evaluation.

The review of trend studies in the field of measurement and evaluation shows that many content analysis studies have been conducted on scale development and adaptation (Acar Güvendir & Özer Özkan, 2014; Bastos, Celeste, Faerstein & Barros, 2010; Boztunç Öztürk, Eroğlu & Kelecioğlu, 2015; Çüm & Koç, 2013; Hinkin, 1995; Kapuscinski & Masters, 2010; Ladhari, 2010; Morgado, Meireles, Neves, Amaral & Ferreira, 2017; Sveinbjornsdottir & Thorsteinsson, 2008; Şahin & Boztunç Öztürk, 2019; Tavşancıl, Güler & Ayan, 2014; Worthington & Whittaker, 2006). In a trend study, 40 papers and 49 researches related to the "Measurement-Evaluation" dimension of the 2004 primary education program were examined (Kazu & Aslan, 2013). In another study conducted by Şenyurt and Özer Özkan (2017) master's theses on measurement and evaluation in education were examined methodologically and thematically. Kazu and Deniz (2019) evaluated the studies investigating teachers in terms of using measurement and evaluation techniques. Gotch and French (2014) reviewed studies on teachers' evaluation literacy. Apart from these, trend researches are required in many other areas of measurement and evaluation. Kieffer, Reese and Thompson (2001) investigated statistical techniques used in education and psychology studies. Yalçın (2016) studied 584 articles for many aspects from the field of measurement and evaluation indexed in the Social Science

Citation Index (SSCI). According to this study, measurement invariance took place in 41 areas. In this sense, measurement invariance can also be said to have a crucial role in the field of measurement and evaluation. Vandenberg and Lance (2000) analyzed 14 studies published between 1971 and 1998 aiming to define and develop measurement invariance theoretically and within the framework of ConfirmatoryFactor Analysis (CFA) and 67 studies published between 1982 and 1999 which tested hypotheses intending to study measurement invariance. In studies in both groups, reviews focused on how the CFA procedure is carried out. Another trend research that examined studies published between 2000 and 2007 and analyzed measurement invariance of measurement tools in terms of different groups by using CFA was conducted by Schmitt and Kuljanin (2008). This study addressed the study areas (intelligence, depression, etc.), groups that were compared (gender, age, etc.), whether the scale was translated, and the analysis steps of measurement invariance (Schmitt &Kuljanin, 2008). Accordingly, when the related literature is analyzed, it can be seen that the measurement invariance trend studies focus on the framework of CFA. Although measurement invariance studies have been handled in terms of variables such as the subject area, compared groups, and translation of the scale, it can be said that these variables are limited. The measurement invariance studies usually deal with with the CFA steps. Nevertheless, reviews examining the studies conducted in 2008 and on have not been found. The analysis of measurement invariance studies published in recent years in several aspects and in terms of several variables is considered important in this respect.

For inter-group comparisons to be made fairly and appropriately, the tests must meet the measurement invariance assumption. Studies conducted to determine the measurement invariance are known to differ in many aspects. There is a need for literature reviews to determine trends in the related area. As with similar studies in many fields, trend studies can also be conducted on measurement invariance. For this reason, this study aims to analyze studies conducted on measurement invariance in the last 12 years and reveal the trends in terms of several variables such as what groups were involved in measurement invariance, the type of measurement tool in terms of the measured feature, the size of the study group, the statistical technique and software used, and the bibliometric analysis regarding most frequently used keywords, cited articles, authors, and journals.

2. METHOD

This study used the review model since it aimed to examine the measurement invariance articles that were published between 2008 and 2019 in three journals, which are reviewed on SSCI, according to some specific criteria, and to reveal trends in this area (Karasar, 2017).

2.1. Population and Sample

The universe of the study consisted of articles on measurement invariance published in journals reviewed on SSCI. Sampling was conducted in two stages in this study, which used the criterion sampling method, which is a purposive sampling method (Patton, 2002). Accordingly, first, the journals were selected. The criteria in the selection of journals were as follows: the journal should be in the field of educational sciences, it should include the words "measurement" or "evaluation", and it should have a high impact factor. Accordingly, the following journals were included in the sample when they were sorted according to impact factors.

- Educational and Psychological Measurement
- Journal of Psychoeducational Assessment
- Applied Measurement in Education

In the second stage of sampling, measurement invariance studies that were published between 2008 and 2019 and whose full-text version could be accessed were determined. However, simulation studies that were not appropriate for the variables to be examined concerning the purpose of the research were excluded from the sample. Accordingly, the distribution of the articles included in the present study by years and journals is given in Table 1.

Table 1. *Distribution of articles examined within the scope of the study by journals and years*

	Years												
Journal	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
Applied Measurement in Education	0	0	1	0	0	0	0	1	0	1	0	0	3
Educational and Psychological Measurement	5	7	4	1	1	2	1	1	0	0	3	0	25
Journal of Psychoeducational Assessment	2	0	2	4	2	5	8	9	12	4	17	6	71
Total	7	7	7	5	3	7	9	11	12	6	20	6	99

As seen in Table 1, 71 articles were included from the "Journal of Psychoeducational Assessment", 25 articles from the "Educational and Psychological Measurement" journal, and three articles from the "Applied Measurement in Education" journal. Regarding the distribution of the articles by years, the majority of the articles (20) on measurement invariance were published in 2018. A total of 99 articles published between 2008 and 2019 in these three journals were reviewed.

2.2. Data Collection and Analysis

To collect data an article review form specific to this study was developed by examining the article or thesis review forms previously used in similar studies in the literature (Tavşancıl et. al., 2010; Yalçın, 2016). The form consisted of a total of 10 sections including the code of the article (M1, M2..); the name of the journal in which the article was published; the year of the publication; the name of the article; the test whose invariance was analyzed; the type of the test, whose measurement invariance was analyzed, in terms of the measured feature; the groups between which invariance was analyzed; sample group; sample size; data analysis method; and the statistical software package used.

Content analysis and bibliometric analysis methods were used for the analysis of the data obtained. Content analysis can be done in the form of systematic coding of qualitative or quantitative data within the framework of certain themes and classifications (Cohen et al., 2007; Fraenkel et al., 2007). According to Falkinham and Reeves (1998), content analysis is specified as a method used in studies in which piles of publications are analyzed or evaluated. In this study, each article was analyzed according to each variable in the evaluation form, and the data obtained were classified according to these variables. Themes were created for the qualitative data obtained for each category. For example, themes such as *maximum performance or typical response test* for the measurement tool whose invariance was analyzedor *ethnicity, gender* for groups for which invariance was analyzed were created. The coding process in determining the validity and reliability categories in content analysis depends on the comprehensibility, clarity, and overlapping (Tavşancıl& Aslan, 2001). For this purpose, 20 randomly selected articles from the article group were re-analyzed by the researcher for intra-rater and inter-rater reliability. Three different raters also reviewed another 20 randomly selected articles for inter-rater reliability. No discrepancy was determined in both the intra-rater and inter-rater reliability study.

Another data analysis technique used in the study was bibliometric mapping analysis. Bibliometrics is defined as the statistical analysis of articles, books, and other publications (Oxford Dictionary, 2017). With bibliometric methods, the image of the related science field can be obtained from bibliographic data obtained from databases (Zupic, 2015). Bibliometric mapping method, on the other hand, is a widely used method in which visual and quantitative findings obtained from the relationship between the data of certain fields in terms of certain variables can be obtained (Small, 1999; Wang et al., 2016). In the present study, VOSViewer 1.6.14 software package was used for bibliometric analysis. The steps followed for the analysis were as follows. The articles examined within the scope of the study were accessed on the WoS and Ebscohost databases. The database widely used in bibliometric studies, especially for social sciences, is the Web of Science (WoS) database (Yang et al., 2015). WoS database was preferred in this study because it is a user-friendly database with easy-to-use rich content. A folder containing the articles investigated within the scope of the study was obtained on the Ebscohost and WoS databases. The data related to the source file obtained can be downloaded by selecting the "Full Record and Cited References" option in the "Tab-delimited" file format, which is a suitable file type for VOSViewer. Afterward, the related file is transferred onto the program, the criterion (for example, the most used keywords) is determined, and the analysis is completed (van Eck & Waltman, 2020). For this purpose, the following criteria were used for the review: five matches for determining the most frequently used keywords, at least 20 views for the most cited articles and authors, and at least 50 views for the most cited source. To determine the criteria the criteria, the structure of the data was taken into account so that the findings could be interpreted.

3. RESULT / FINDINGS

This section deals with findings and interpretations. For this purpose, firstly, content analysis findings, and then bibliometric analysis findings were discussed.

3.1. Content Analysis Findings related to Studies and Trends in the Field of Measurement Invariance

In the study, each article was examined under predetermined categories. Accordingly, while examining the measurement invariance articles, the following categories were taken into consideration: the type of the test, whose invariance was examined, in terms of the measured feature; groups between which the invariance was examined; the study group; sample size; data analysis method; and the software package used. Accordingly, Table 2 presents the distribution of studies in the field of measurement invariance by maximum performance or the typical response test variables regarding “the type of the test, whose invariance was analyzed, in terms of the measured feature”.

Table 2. *The distribution of tests analyzed in studies in terms of measured feature*

Test type	f	%
Typical Response	75	75.76
Maximum Performance	24	24.24
Total	99	100.00

As seen in Table 2, 75.76% of the articles analyzed in the study were typical response tests (f = 75) while 24.24% of the tests were determined to be the maximum performance tests. When the tests were examined, it is noteworthy that nine of the maximum performance tests were IQ tests, which outnumbered others. Table 3 presents the distribution of the measurement invariance studies in terms of study group characteristics.

Table 3. *The distribution of the measurement invariance studies in terms of study group characteristics*

Study Groups	f	%
Students	84	84.85
Teachers	9	9.09
Adults	8	8.08
All age groups	2	2.02
Total	103	100.00

As seen in Table 3, students were the most frequent study groups selected in measurement invariance studies with 84.85%. On the other hand, studies selecting teachers as the study group ranked second with 9.09%. Measurement invariance studies based on data obtained from adults constituted 8.08% of all studies. Also, measurement invariance studies using all age groups made up 2.02%. Under this classification, there were a total of 103 study groups. There were four studies using more than one group. For example, the study coded M3 was found to use data obtained from students and teachers. The distribution of measurement invariance studies by sample size is given in Table 4.

Table 4. *The distribution of measurement invariance studies by sample size*

Size of the Study Group	f	%
1- 500	24	24.24
501 -1000	19	19.19
1001 – 1500	20	20.20
1501 or larger	36	37.37
Total	99	100.00

As seen in Table 4, the size of the study group was 1501 or larger in 37.37% of the 99 articles, which examined the measurement invariance. Also, studies with a study group size in the range of 1-500, 501-1000, and 1001-1500 made up 24.24%, 19.19%, and 20.20% of all studies, respectively. The study coded M1 was conducted with a group of 144 people. On the other hand, the study coded M46 was carried out with 42.163 people. This difference was observed to arise from the number of groups used in the studies according to certain variables. For example, in the study coded M1, the time-dependent invariance of the test was examined based on data from a single group, while in the study coded M46, there were students from 10 different grade levels. Table 5 shows the distribution of invariance studies by groups between which invariance was examined.

Table 5. *The distribution of invariance studies by groups between which invariance was examined*

Variable	f	%	Variable	f	%
Gender	42	42.42	Place of residence	3	3.03
Ethnicity	15	15.15	Socio-economic level	3	3.03
Age	15	15.15	Experiment-Control	2	2.02
Country	9	9.09	Paper Pen Test- Computer Based Test	2	2.02
Culture	9	9.09	Status of Learning Difficulty	2	2.02
Education	8	8.08	Self-Peer Assessment	2	2.02
Types of school	6	6.06	Use of Tests in Low-High Risk Exams	1	1.01
Grade level	6	6.06	With–Without Diagnosis of ADHD*	1	1.01
Time	6	6.06	Learning difficulty due to ADHD*	1	1.01
Inter-rater	4	4.04	Student-teacher	1	1.01
Language	3	3.03	Face-to-Face / Online Course	1	1.01
		Total	f=142		%=100.00

*ADHD: Attention Deficit Hyperactivity Disorder

The examination of the variables in Table 5 indicated that the tests were mostly analyzed regarding whether they showed measurement invariance for the gender variable. Accordingly, in 42.42% of the articles examined, the invariance of measurement was examined in terms of gender variable. Gender was followed by some demographic characteristics such as age (15.15%), socio-economic level (3.03%), and place of residence (3.03%). Also, cultural variables are important in the measurement invariance studies analyzed. Accordingly, the variables such as ethnicity with 15.15%, country with 9.09%, culture with 9.09%, and language with 3.03% drew attention as cultural variables. On the other hand, education-related variables such as education level (8.08%), school type (6.06%), and grade level (6.06%) also appeared to be among the most used variables in testing invariance of tests. The learning difficulty was another variable employed in the analysis of the measurement invariance of tests. Accordingly, the measurement invariance was also analyzed in terms of learning difficulty with 2.02%, with–without a diagnosis of ADHD with 1.01%, and learning difficulty due to ADHD with 1.01%. Also, the rate of articles analyzing the time-dependent invariance of the tests was 6.06%. In terms of the application methods of the tests, the invariance of the Paper-Pen Tests-Computer-Based Tests (2.02%) was also analyzed in the articles examined. In additiont the invariance of tests was analyzed in terms of their use in Low-High Risk Exams (1.01%). Inter-rater invariance was another variable that was analyzed in the studies. For example, inter-rater invariance studies accounted for 4.04% of the total measurement invariance studies, and the self-peer assessment made up 2.02%. Also, the invariance of inter-group tests was studied in the experimental studies. Accordingly, the measurement invariance of the tests applied in terms of experiment-control groups (2.02%) and students taking face-to-face / online courses (1.01%) was examined, too. The examination of the statistical techniques used for measurement invariance analysis in the articles which examined the measurement invariance indicated that the "Multi-Group Confirmatory Factor Analysis" technique was used in all studies. Table 6 presents the distribution of studies according to the statistical software packages used to conduct this analysis.

Table 6. *The distribution of studies according to the statistical software packages used*

Statistical Software	f	%
Mplus	54	54.55
LISREL	14	14.14
Amos	11	11.11
R	7	7.07
Not specified	6	6.06
EQS	6	6.06
SAS/STAT® software	1	1.01
Total	99	100.00

The examination of Table 6 showing the distribution by statistical software packages indicated that the most frequently used statistical software package in measurement invariance studies was Mplus with 54.55%. Also, LISREL, Amos, and EQS, which are often used for structural equation modeling, were also used in studies which investigate measurement invariance with 14.14%, 11.11%, and 6.06%, respectively. R, which is the latest launched software package, was used in 7.07% of the studies. On the other hand, SAS/STAT® was observed to be used in just one of the studies. Besides, 6.06% of the measurement invariance studies were found to not specify the statistical software employed.

3.2. Bibliometric Analysis Findings Regarding Studies and Trends in the Field of Measurement Invariance

With bibliometric mapping analysis, the findings obtained for 99 articles which investigate the measurement invariance were included in this study are presented under this heading. The articles were examined in terms of the most frequently used keywords, the most cited publications, the most cited authors, and the most cited journals. Most frequently used keywords, with at least five matches, were examined in the measurement invariance articles. Accordingly, the map obtained as a result of the analysis is given in Figure 1 and the frequency values for each keyword are given in Table 7.

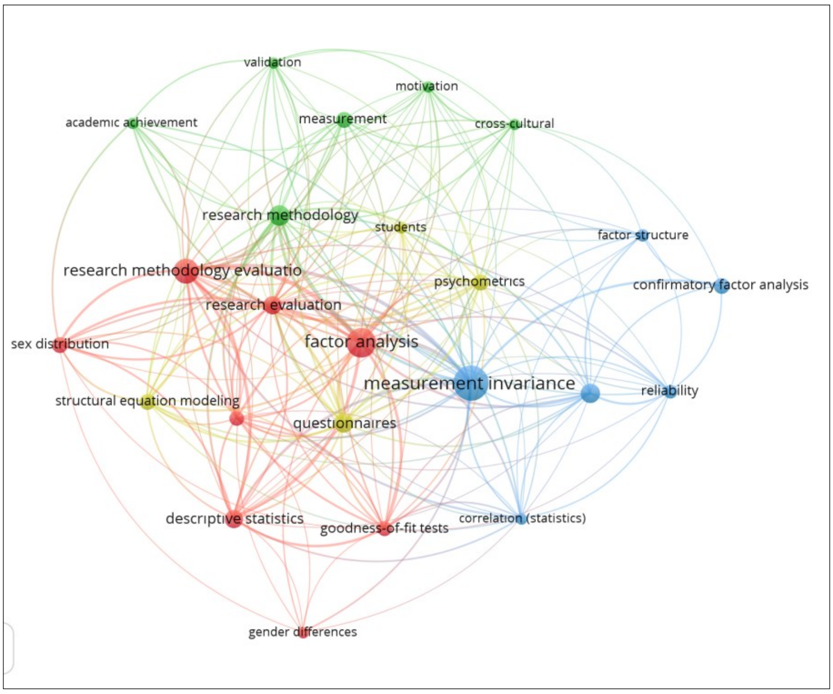


Figure 1. *The most frequently used keywords in studies in the field of measurement invariance*

Table 7. *The most frequently used keywords in studies in the field of measurement invariance*

Keywords	f	Keywords	f
Measurement invariance	46	Confirmatory factor analysis	9
Factor analysis	31	Goodness-of-fit tests	8
Research methodology evaluation	23	Self-evaluation	8
Research methodology	15	Reliability	7
Questionnaires	14	Factor structure	6
Validity	14	Correlation	5
Descriptive statistics	13	Students	5
Research evaluation	12	Academic achievement	5
Measurement	10	Cross-cultural	5
Psychometrics	9	Motivation	5
Structural equation modeling	9	Validation	5
Sex distribution	9	Gender differences	5

As seen in Figure 1 and Table 7 showing the most frequently used keywords in measurement invariance studies, as expected, the most frequently used keyword was "measurement invariance" (f=46). "Factor analysis" (f=31), "Structural equation modeling" (f=9), "Confirmatory factor analysis" (f=9), "Goodness of fit tests" (f=9), and "Factor structure" (f=6) were among the most used keywords, which revealed the importance of factor analysis for measurement invariance studies. Also, the frequent use of "Research methodology evaluation" (f=23), "Research methodology" (f=15), and "Research evaluation" (f=12) keywords showed the importance of research methodology in measurement invariance studies. Besides, the frequent use of keywords such as "Validity" (f=14), "Reliability" (f=7), and "Validation" (f=5) supported the view that measurement invariance was an important proof of validity and reliability. Of the most frequently used keywords in measurement invariance studies, "Questionnaires" (f=14), "Descriptive statistics" (f=13), "Measurement" (f=10), "Psychometrics" (f=9), "Self-evaluation" (f=8), and "Correlation" (f=5) were observed to be important concepts for the field of Measurement and Evaluation. Similar to the findings of content analysis, the use of keywords such as "Students" (f=5), "Gender distribution" (f=9), "Gender differences" (f=5), and "Intercultural" (f=5) indicated that measurement invariance studies often used data obtained from students, invariance was most frequently investigated over gender variable, and that intercultural invariance had a considerable place in measurement invariance studies. Moreover, the frequent use of keywords such as "Academic achievement" (f=5) and "Motivation" (f=5) showed that invariance studies regarding these features were often conducted. Figure 2 shows the map obtained according to the results of the analysis regarding the most cited articles with at least 20 views cited in measurement invariance articles. Also, Table 8 presents the citation frequency values of each publication.

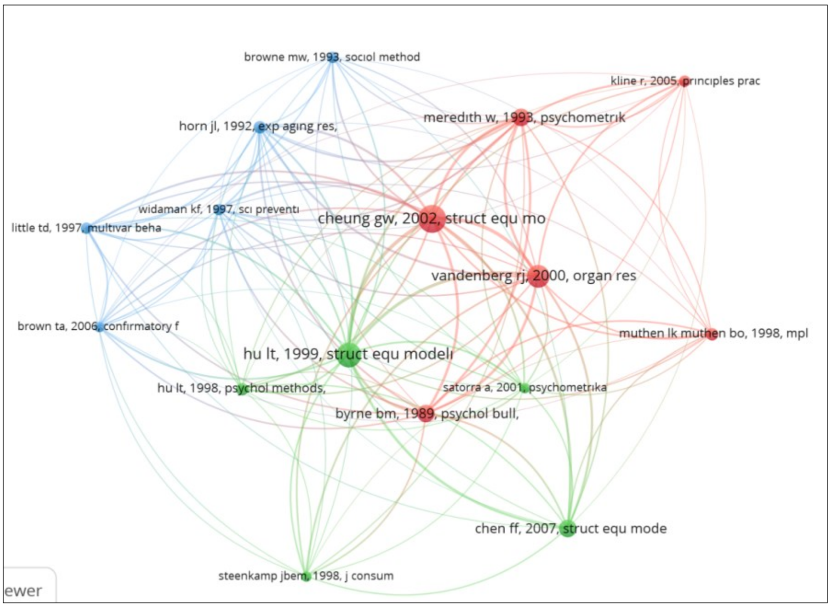

browne mw, 1993, sociol method
kline r, 2005, principles prac
meredith w, 1993, psychometrik
horn jl, 1992, exp aging res,
widaman kf, 1997, sci preventi
cheung gw, 2002, struct equ mo
little td, 1997, multivar beha
vandenberg rj, 2000, organ res
brown ta, 2006, confirmatory f
muthen lk muthen bo, 1998, mpl
hu lt, 1999, struct equ modeli
hu lt, 1998, psychol methods,
satorra a, 2001, psychometrika
byrne bm, 1989, psychol bull,
chen ff, 2007, struct equ mode
steenkamp jbem, 1998, j consum
ewer

Figure 2. *The most cited publications in studies in the field of measurement invariance*

The examination of the measurement invariance studies in Table 8 showing the most cited publications indicated that publications using factor analysis, goodness of fit, and structural equation modeling statistics were cited considerably (f_{Y1}=65, f_{Y2}=51, f_{Y4}=27, f_{Y5}=27, f_{Y6}=26, f_{Y8}=14, f_{Y11}=12, f_{Y12}=11, f_{Y13}=11, f_{Y14}=10, f_{Y16}=10). Accordingly, 11 of the 16 most cited publications were based on the statistical processes used in the analysis of measurement invariance. The publication titled "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research" by Vandenberg and Lance (2000) was the most cited (f_{Y3} = 43) publication. There were four studies dealing with the theoretical, application, and evaluation dimensions of measurement invariance (f_{Y3}=43, f_{Y7}=14, f_{Y10}=12, f_{Y15}=10). In parallel with the content analysis findings, the user manual (f_{Y9}=13) about the Mplus software package, which is widely used in structural equation modeling statistics, was among the most cited publications. Figure 3 shows the map obtained according to the results of the analysis done for determining the most cited authors with at least 20 citations, and Table 9 shows the frequency values of the citations.

Table 8. The most cited publications in studies in the field of measurement invariance

Study Code	Name of The Author	Publication Title	f
Y1	Cheung &Rensvold (2002)	Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance	65
Y2	Hu &Bentler (1999)	Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives	51
Y3	Vandenberg & Lance (2000)	A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research	43
Y4	Meredith (1993)	Measurement invariance, factor analysis and factorial invariance	27
Y5	Byrne, Shavelson &Muthén (1989)	Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance	27
Y6	Chen (2007)	Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance	26
Y7	Horn &Mcardle (1992)	A practical and theoretical guide to measurement invariance in aging research	14
Y8	Hu &Bentler (1998)	Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification	14
Y9	Muthén and Muthén (2007)	Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén&Muthén	13
Y10	Widaman&Reise (1997)	Exploring the measurement invariance of psychological instruments: Applications in the substance use domain	12
Y11	Kline (2005)	Methodology in the social sciences. Principles and practice of structural equation modeling (2nd ed.). Guilford Press.	12
Y12	Little (1997)	Mean and Covariance Structures (MACS) Analyses of Cross-Cultural Data: Practical and Theoretical Issues	11
Y13	Browne &Cudeck (1993)	Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), Testing structural equation models (pp. 136-162)	11
Y14	Satorra&Bentler (2001)	A scaled difference chi-square test statistic for moment structure analysis	10
Y15	Steenkamp & Baumgartner (1998)	Assessing Measurement Invariance in Cross-National Consumer Research	10
Y16	Brown (2015)	Confirmatory factor analysis for applied research. The Guilford Press.	10

As seen in Table 9 and Figure 3, the most cited author was Cheung, G. W. (f=80). Cheung, G. W was also one of the authors of the most cited publications (f_{Y1}=65). According to the findings, the most cited authors were sequenced similar to the most cited publications. Also, the references to some authors who did not have a publication among the most cited publications were quite high. This might be because the total number of citations to different publications of the related authors was high. These authors included Marsh, H.W. (f=47), Millsap, R.E. (f=29), Martin, A. (f=26), Reynolds, C.R. (f=22), Hancock, G. (f=20), and Bandura, A. (f=21). On the other hand, the book titled "Statistical Approaches to Measurement Invariance" which was written by Millsap, R.E. (f=29) was seen as an important source that addresses measurement invariance. Regarding the investigation of references to measurement invariance studies, Figure 4 shows the bibliographic map obtained considering the references to each journal according to the results of the analysis done for determining the most cited journals that had at least 50 views, and Table 10 shows the frequency values of the citations.

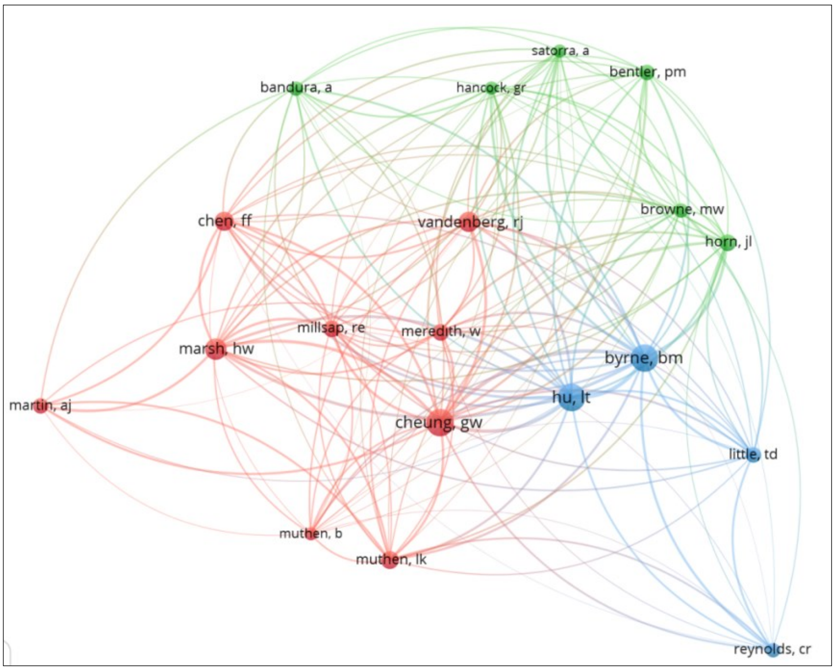


Figure 3. *The most cited authors in studies in the field of measurement invariance*

Table 9. *The most cited authors in studies in the field of measurement invariance*

Author	f	Author	f
Cheung, G. W.	80	Bentler, P.	27
Byrne, B. M	77	Martin, A.	26
Hu, L.	75	Little, T.	24
Marsh, H. W.	47	Browne, M. W.	22
Vandenberg, R. J	45	Reynolds, C. R.	22
Chen, F. F.	39	Bandura, A	21
Muthen, L.	36	Muthen, B.	20
Meredith, W.	30	Satorra, A.	20
Millsap, R. E.	29	Hancock, G.	20
Horn, J. L.	28		

As seen in Figure 4 and Table 10, the reference to the journal named "Structural Equation Modeling: A Multidisciplinary Journal" was quite high (f=235). This might be because structural equation modeling has an important place in the investigation of measurement invariance. On the other hand, journals that use the word "psychology" or its derivations in their name were also observed to be cited Journals such as "Psychometrıka" (f=103), "Psychological Bulletin" (f=96), "Educational and Psychological Measurement" (f=84), "Psychological Methods" (f=79), "Journal of Psychoeducational Assessment" (f=75), "Journal of Educational Psychology" (f=73), and "Psychological Assessment" (f=51) can be given as examples. This reveals that measurement invariance is important especially in psychological measurements. Also, parallel to the fact that the most used keywords were related to the research methodology, these journals were found to contain the word "research" e.g. "Multivariate Behavioral Research" (f=57), "Organizational Research Methods" (f=54). Considering the importance of measurement invariance for measurements conducted in the field of education, journals such as "Educational and Psychological Measurement" (f=84), "Journal of Psychoeducational Assessment" (f=75), and "Journal of Educational Psychology" (f=73) were observed to have educational content. On the other hand, journals such as "Journal of Psychoeducational Assessment" and "Journal of Educational Psychology", which contained the articles analyzed in the study, were also among the most cited journals.

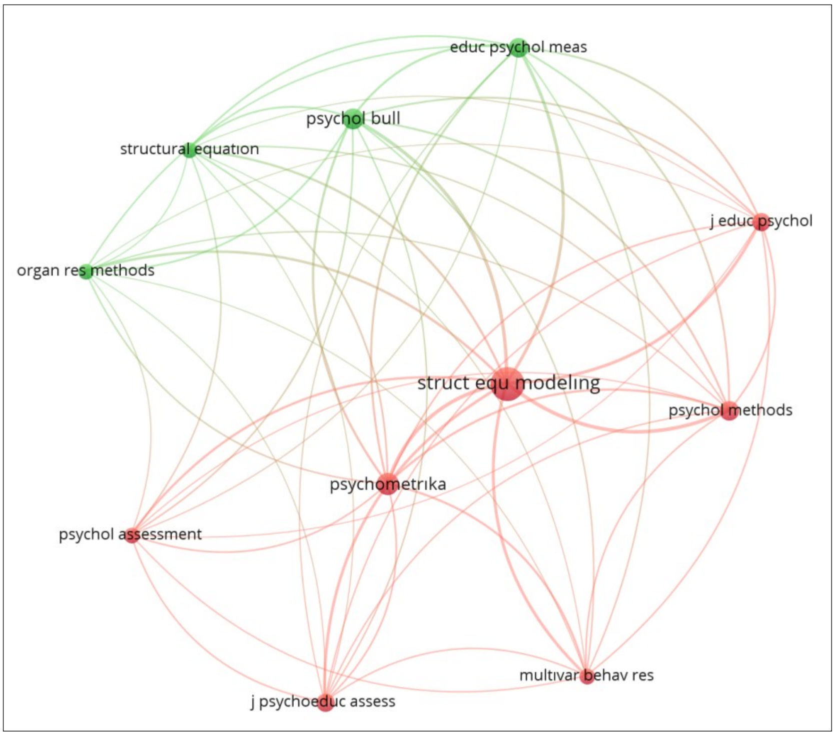


Figure 4. *The most cited sources in studies in the field of measurement invariance*

Table 10. *The most cited sources in studies in the field of measurement invariance*

Source	f
Structural Equation Modeling: A Multidisciplinary Journal	235
Psychometrıka	103
Psychological Bulletin	96
Educational and Psychological Measurement	84
Psychological Methods	79
Journal of Psychoeducational Assessment	75
Journal of Educational Psychology	73
Multivariate Behavioral Research	57
Organizational Research Methods	54
Structural Equation	54
Psychological Assessment	51

4. DISCUSSION and CONCLUSION

This study examined 99 articles in the field of measurement invariance published between 2008 and 2019 using content and bibliometric analyses. Accordingly, the majority of the articles were found to address the measurement invariance of typical reaction tests. Also, maximum performance tests were among the tests whose measurement invariance was examined. Studies handling inter-group measurement invariance regarding characteristics such as interest, perception, attitude, and personality measured by typical reaction tests were found to outnumber studies investigating measurement invariance of features such as intelligence or success which were measured by maximum performance tests. Similarly, typical reaction tests were found to be widely used in trend surveys related to the field of science education in Turkey (Erdem, 2011; Göktaş et al., 2012; Selçuk et al., 2014). Also, according to the results of a content analysis study on educational journals with the highest impact factor, the most used tests were achievement tests (Yalçın, et al., 2015). However, in another study in which journals in the field of measurement and evaluation with high impact factors were examined, achievement tests were found to be used in more than half of the studies (Yalçın, 2016). Accordingly, unlike studies in educational sciences and especially in the field of measurement and evaluation, the invariance of typical reaction tests can be said to be analyzed more in measurement invariance studies.

In studies investigating the measurement invariance, the study group or sample was found to consist mostly of students. On the other hand, the study group or samples were also observed to include teachers, adults, and all age groups in measurement invariance studies. This is because students are the focus group in education, and measurement tools developed for students are more than other elements of education. In trend surveys conducted in Turkey in the field of educational sciences, the research group or sample was determined to often consist of students (Arık & Türkmen, 2009; Göktaş, et. al., 2012; Selçuk et al., 2014; Şenyurt & Özer-Özkan, 2017; Yalçın, et al., 2015).

Since structural equation modeling is used in measurement invariance analysis, comparisons between models are known to be based on model fit indexes. For this reason, since many studies have revealed that the group size affects the model fit indexes (Fan & Sivo, 2007; Fan, Thompson & Wang, 1999; Hu & Bentler, 1998; Lei & Lomax, 2005; Mahler, 2011), the size of the group has an important place in measurement invariance studies. Particular attention should be paid to the size of the compared groups. In articles investigating the measurement invariance, the size of study groups or samples was mostly 1501 and above. The sample size increases according to the number of groups for which the invariance is examined. Yet, the average study group or sample size was determined as 3436.39. In the studies found in educational journals with a big impact factor, sample sizes were much higher (Yalçın, 2016; Yalçın, et al., 2015). For example, the average sample size was 81.008 according to the content analysis related to journals that had the highest impact factor and which were reviewed on SSCI in the field of educational sciences (Yalçın, et al., 2015). This is associated with simulation studies and the availability of data from large-scale applications. However, since this study did not include simulation studies, the sample sizes of studies reported here were lower. Sample sizes were even smaller in studies investigating research in the field of educational sciences held in Turkey (Arık & Türkmen, 2009; Göktaş et al., 2012; Selçuk et al., 2014).

In this study, which examined measurement invariance studies, the invariance of measurement tools was found to be mostly analyzed in terms of gender variable. Also, demographic variables such as age, socio-economic level, and place of residence were among the variables that were analyzed for the invariance of measurement tools. Besides, according to the articles examined within the scope of the study, the invariance of the tests was also analyzed in terms of cultural variables such as ethnicity, country, culture, and language. However, some variables related to education were also considered important for the invariance of the tests. These variables can be listed as education level, school type, and grade level. Although the rate was not high, the measurement invariance of measurement tools was also analyzed according to the diagnostic variables of the learning difficulties of individuals. There were also time-dependent invariance analyses that provided important evidence of validity and reliability. The invariance of measurement tools was also analyzed in terms of exam types including paper-pen, computer-based, or low-high risk tests. Finally, measurement invariance analyses were carried out based on the rater (such as self-peer) or groups created for experimental studies (such as experiment-control). This can be explained by the fact that measurement invariance analyses are performed in terms of many variables since human behaviors measured by measurement tools are complex and abstract. While developing measurement tools, many variables are taken into account from writing items to the selection of groups in experimental studies. However, evidence of whether the measurement tool shows invariance in terms of some variables cannot be obtained during the development stage of the measurement tool. For this reason, conducting measurement invariance studies regarding the important variables is considered to be very important primarily for the validity and reliability of the measurement tool. In their trend analyses on measurement invariance articles, Schmitt and Kuljanin (2008) conducted measurement invariance analyses considering the sub-groups created according to gender, ethnic, cultural, linguistic, and other demographic variables. Indeed, cultural variables have great importance. It is especially necessary to analyze invariance for cultural variables in tests adapted to different cultures. This may explain the abundance of invariance studies conducted in terms of cultural variables in this study. In their trend study investigating measurement invariance studies, which was conducted by Schmitt and Kuljanin (2008), 20 out of 75 articles were tests translated from another language. Similarly, it can be concluded that in invariance studies, invariance analyses are conducted for demographic variables and adapted tests.

In this study, which discusses the measurement invariance of some measurement tools in terms of some variables, the measurement tools and the variables for which the invariance is analyzed show variety, as stated earlier. All of these studies have one aspect in common; that is, they use "Multi-Group Confirmatory Factor Analysis" as the statistical analysis technique. Accordingly, 75 out of 88 articles which were published between 2000 and 2007 and which used the term "measurement invariance" were found to employ confirmatory factor analysis (Schmitt & Kuljanin, 2008). Studies in which measurement invariance studies are analyzed were also determined to use measurement invariance analyses based on the analysis of the difference in variance and covariance matrices, but they were not preferred as much as CFA (Vandenberg & Lance, 2000; Schmitt & Kuljanin, 2008).

The examination of the measurement invariance studies according to the statistical programs used indicated that more than half of the studies had used Mplus statistical software package, and this was followed by LISREL, Amos, R, EQS, and SAS/STAT® according to the frequency of use. Also, some studies had not specified the statistics program employed. In a study investigating articles from the field of measurement and evaluation which had a high impact factor and were reviewed on SSCI, the most frequently used statistical software packages in the descending order were R, Mplus, and SAS (Yalçın, 2015). In another study in the field of educational sciences analyzing articles from journals that had a high impact factor and which were reviewed on SSCI, the frequently used statistical software packages were Mplus, SPSS, and SAS, respectively. On the other hand, the statistical software packages mostly used in trend research in Turkey were SPSS and LISREL, respectively (Arık & Türkmen, 2009; Doğan & Uluman, 2015). The findings obtained in this study were almost similar to other studies conducted in the field of educational sciences. In some articles, the name of the statistical software was not given. This can be due to concerns about avoiding the advertisement in the case of paid software.

As a result of the bibliometric analyses of the measurement invariance studies, the analysis of most frequently used keywords, most cited publications, authors, and sources was conducted. The most frequently used keyword was found to be "measurement invariance". In addition to this, the words specific to factor analysis were used also widely. Among them were the keywords such as "factor analysis", "confirmatory factor analysis", "factor structure", and "goodness of fit tests". Another group of frequently used keywords was related to research methodology such as "evaluation of research methodology", "research methodology", and "research evaluation", which indicate that measurement invariance studies have an important place in research methodology. Similarly, "validity", "reliability", and "validation" keywords were also among the frequently used keywords. Accordingly, it has been re-established that the measurement invariance is directly related to validity and reliability. Validity and reliability cannot be achieved unless evidence regarding measurement invariance is obtained (Vanderberg & Lance, 2000). Some basic concepts of measurement and evaluation such as "surveys", "descriptive statistics", "measurement", "psychometrics", "self-evaluation" and "correlation" were also included in the measurement invariance studies. Another set of frequently used keywords was observed to include "gender", "gender distribution", “gender differences", and "intercultural". In line with the results of the content analysis, the keyword "students" was found to be among the frequently used keywords. Moreover, "academic achievement" and "motivation" keywords, which are frequently measured concepts in education and psychology, were among the most used keywords in measurement invariance studies.

The article named "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance" written by Cheung and Rensvold (2009) was determined to be the most frequently cited publication among measurement invariance studies published in three journals in the field of measurement and evaluation, which had a high impact and which were reviewed on SSCI. This publication was followed by "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives", which was written by Hu and Bentler (2009). Apart from these, nine other publications, which were among the most cited were related to statistical processes for analyzing measurement invariance. Vandenberg and Lance's (2000) publication named "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research" was another most frequently cited study that addressed both statistical processes and trend research regarding measurement invariance comprehensively. Schmitt and Kuljanin (2008) referred to the study of Vandenberg and Lance (2000) in their trend research on measurement invariance and stated that they aimed to examine studies published after this study. Similar to the study of Vandenberg and Lance (2000), the studies of Horn and Mcardle (1992), Widaman and Reise (1997), and Steenkamp and Baumgartner (1998), which dealt with the theoretical, application and evaluation aspects of measurement invariance, were also among the most cited publications. The publications of Horn and Mcardle (1992) and Steenkamp and Baumgartner (1998) were included in the analysis group during reviewing the study of Vandenberg and Lance (2000), which aimed to define and develop measurement invariance theoretically. According to the content analysis findings, the most used statistical software package was "Mplus". The user manual (f_{Y9}=13) of the Mplus statistical software package, which is widely used in structural equation modeling statistics, was also among the most cited publications. This user manual was written by Muthén and Muthén (2007) and named "Mplus User’s Guide".

According to the results of bibliometric analysis of the measurement invariance studies, the most cited author was identified as Cheung, G. W. In parallel to the most cited publications, the authors of the publications were also in the most cited authors list. This list also included the following authors: Marsh H. W., Millsap R. E., Martin A, Reynolds C. R, Hancock G., and Bandura A. Moreover, the author of the publication that addressed statistical methods regarding measurement invariance, Millsap, R. E. (2011) was among the most cited authors.

When the measurement invariance studies were examined in terms of the most cited sources (journals, publishing houses, etc.), the journal with the highest reference was the "Structural Equation Modeling: A Multidisciplinary Journal". In addition to this, journals from the field of psychology such as "Psychometrics", "Psychological Bulletin", "Educational and Psychological Measurement", "Psychological Methods", "Journal of Psychoeducational Assessment", "Journal of Educational Psychology", and "Psychological Assessment" were also found to be among the most cited journals. This may be associated with the fact that the measurement invariance of typical response tests was analyzed more. On the other hand, journals such as "Multivariate Behavioral Research" and "Organizational Research Methods" were also frequently cited, which were considered to show up associated with research methodology related keywords. Concerning the place and importance of measurement invariance in the field of measurement and evaluation in education, the following journals from the field of educational research were among the most cited journals: "Educational and Psychological Measurement", "Journal of Psychoeducational Assessment", and "Journal of Educational Psychology".

4.1. Recommendations

According to the findings, the measurement invariance of the maximum performance tests was analyzed less frequently compared to the typical response tests. However, the results of maximum performance tests can provide insights into making important decisions about individuals such as placement in an educational institution or recruitment for a job. For this reason, we recommended that the measurement invariance analysis of maximum performance tests should be increased in terms of many variables to make fairer and more appropriate decisions about individuals.

According to the results obtained from the study, the majority of measurement invariance analyses were found to be done considering the gender variable. Also, many other variables were handled in measurement invariance studies. Accordingly, more reviews and studies should be carried out to guide the developers of measurement tools in terms of showing which variables should be analyzed for a given measurement tool. Considering the importance of measurement invariance in terms of validity and reliability, measurement invariance analyses should be carried out within the scope of validity and reliability studies to emphasize this significance for measurement tool developers, especially for the item writing process.

Given the finding that measurement invariance studies are mostly based on data obtained from students, the measurement invariance studies of measurement tools used in the field of education and psychology targeting groups such as teachers, administrators, and parents can also be carried out. According to the results obtained from the study, many statistical software packages were used in measurement invariance studies. In terms of measurement invariance analysis, the advantages and disadvantages of the related software packages over each other can be investigated.

In this study, measurement invariance studies were examined in terms of various variables. However, we could not analyze the findings obtained as a result of the studies conducted. With the investigation of studies which focus on the measurement invariance of certain measurement tools through certain variables, significant interpretations can be put forward regarding the validity and reliability of the related measurement tool. In this study, measurement invariance studies published in three high impact factor measurement and evaluation journals that were reviewed on SSCI were analyzed. Measurement invariance studies from different databases, journals, years and countries can also be analyzed.

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).

ORCID

Betul ALATLI https://orcid.org/0000-0003-2424-5937

5. REFERENCES

Acar, G. M. & Özkan, Ö. Y. (2015). Türkiye'deki eğitim alanında yayımlanan bilimsel dergilerde ölçek geliştirme ve uyarlama konulu makalelerin incelenmesi. *Elektronik Sosyal Bilimler Dergisi*, *14*(52), 23-33. http://dx.doi.org/10.17755/esosder.54872

Akaydın, Ş. & Çeçen M. A. (2015). Okuma becerisiyle ilgili makaleler üzerine bir içerik analizi. *Eğitim ve Bilim*, *40*(178), 183-198. http://dx.doi.org/10.15390/EB.2015.4139

American Educational Research Association, American Psychological Association, NationalCouncil on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing*. Washington: American Psychological Association.

Arıcı, F., Yıldırım, P., Çalıklar, Ş., & Yılmaz, R. M. (2019). Research trends in the use of augmentedreality in science education: Content and bibliometric mapping analysis. *Computers & Education*, *142*(December), 103647. http://dx.doi.org/10.1016/j.compedu.2019.103647

Arık, R. S., & Türkmen, M. (2009). *Eğitim bilimleri alanında yayınlanan bilimsel dergilerde yer alan makalelerin incelenmesi.* Retrieved November 11, 2019, from http://www.eab.org.tr/eab/2009/pdf/488.pdf

Aslan, C., & Özkubat, U. (2019). Ulusal özel eğitim kongresi bildirilerindeki araştırma eğilimleri: Bir İçerik analizi. *Türkiye Sosyal Araştırmalar Dergisi*, *23*(2), 535-554.

Aydın, A., Erdağ, C., & Sarıer, Y. (2010). Eğitim yönetimi alanında yayınlanan makalelerin konu, yöntem ve sonuçlar açısından karşılaştırılması. *Eurasian Journal of Educational Research*, 39, 37-58.

Aypay, A., Coruk, A., Yazgan, D., Kartal, O., Çağatay, M., Tuncer, B., & Emran, B. (2010). The status of research in educational administration: An analysis of educationaladministrationjournals, 1999-2007. *Eurasian Journal of Educational Research*, *39*, 59-77.

Aztekin, S., & Taşpınar Şener, Z. (2015). Türkiye'de matematik eğitimi alanındaki matematiksel modelleme araştırmalarının içerik analizi: Bir meta-sentez çalışması, *Eğitim ve Bilim*, *40*(178), 139-161. http://dx.doi.org/10.15390/EB.2015.4125

Bacanak, A., Karamustafaoğlu, S., Değirmenci, S., & Karamustafaoğlu, O. (2011). E-dergilerde yayınlanan fen eğitimi makaleleri: Yöntem analizi. *Türk Fen Eğitimi Dergisi*, 8(1), 119-132.

Baki, A., Güven, B., Karataş, İ., Akkan, Y., & Çakıroğlu, Ü. (2011). Türkiye'deki matematik eğitimi araştırmalarındaki eğitimler:1998 ile 2007 yılları arası. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *40*, 57-68.

Bastos, J. L., Celeste, R. K., Faerstein, E., & Barros, A. J. D. (2010). Racial discrimination and health: a systematic review of scales with a focus on their psychometric properties. *Social Science and Medicine*, *70*(7), 1091-1099. https://doi.org/10.1016/j.socscimed.2009.12.20

Bozkurt, A., Akgün-Özbek, E., Yılmazel, S., Erdoğdu, E., Uçar, H., Güler, E., Sezgin, S., & Dincer, G.D. (2015). Trends in distance education research: Acontent analysis of journals 2009-2013. *The International Review of Research in Open and Distributed Learning*, *16*(1), 330-363. https://doi.org/10.19173/irrodl.v16i1.1953

Boztunç Öztürk, N., Eroğlu, M. G., & Kelecioğlu, H. (2015). Eğitim bilimleri alanında yapılan ölçek uyarlama makalelerinin incelenmesi, *Eğitim ve Bilim*, *40*(178) 123-137. http://dx.doi.org/10.15390/EB.2015.4091

Brown, T. A. (2015). *Methodology in the social sciences. Confirmatory factor analysis for applied research (2nd ed.).* New York: Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chang, Y., Chang, C., & Tseng, Y. (2010). Trends of science education research: An automatic content analysis. *Journal of Science Education and Technology, 19*(4), 315-331.

Chen, F., F. (2007). Sensitivity of goodness of fit ındexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung G., W., & Rensvold R., B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. London and New York, NY: Routledge Falmer.

Çiltaş, A., Güler, G., & Sözbilir, M. (2012). Türkiye'de matematik eğitimi araştırmaları: Bir içerik analizi çalışması. *Kuram ve Uygulamada Eğitim Bilimleri, 12*(1), 565-580.

Çüm, S., & Koç, N. (2013). Türkiye'de psikoloji ve eğitim bilimleri dergilerinde yayımlanan ölçek geliştirme ve uyarlama çalışmalarının incelenmesi. *Eğitim Bilimleri ve Uygulama, 12*(24), 115-135.

de Jong O (2007) Trends in western science curricula and science education research: A bird's eyeview. *Journal of Baltic Science Education*, *6*(1). 15–22.

Doğan C., D., & Uluman M. (2016). İstatistiksel Veri Analizinde R Yazılımı ve Kullanımı. *İlköğretim Online*, *15*(2), 615-634., https://doi.org/10.17051/io.2016.24991

Doğan, H., & Tok, T. N. (2018). Türkiye'de eğitim bilimleri alanında yayınlanan makalelerin incelenmesi: Eğitim ve Bilim Dergisi örneği. *Current Research in Education, 4*(2), 94-109.

Erdem Aydın İ., Kaya S., İşkol S., & İşcan A., (2019). Anadolu Üniversitesi uzaktan eğitim bölümünde yayınlanmış yüksek lisans ve doktora tezlerinin içerik analizi. *Journal of Higher Education and Science, 9*(3), 430-441. https://doi.org/10.5961/jhes.2019.343

Erdem Aydın, İ., Bozkaya, M., & Genç Kumtepe, E. (2019). Research trends and issues in educational technology: Content analysis of TOJET (2012–2018). *The Turkish Online Journal of Educational Technology, 18*(4), 46-61.

Erdem, D. (2011). Türkiye'de 2005–2006 yılları arasında yayımlanan eğitim bilimleri dergilerindeki makalelerin bazı özellikler açısından incelenmesi: Betimsel bir analiz. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *2*(1), 140-147.

Erdoğmuş, F., U., & Çağıltay, K. (2009, Şubat). *Türkiye'de eğitim teknolojileri alanında yapılan master ve doktora tezlerinde genel eğilimler*. Paper presented at the XI. Akademik Bilişim Konferansı, Harran Üniversitesi, Şanlıurfa. Retrieved May 15, 2019, from https://ab.org.tr/ab09/kitap/erdogmus_cagiltay_AB09.pdf

Eybe, J., & Schmidt, H.-J. (2001). Quality criteria and exemplary papers in chemistry education research. *International Journal of Science Education, 23*, 209-225. https://doi.org/10.1080/09500690118920

Falkingham, L. T. & Reeves, R. (1998). Context analysis a technique for analysing research in a field, applied to literature on the management of R and D at the section level. *Scientometrics, 42*(2), 97-120. https://doi.org/10.1007/bf02458351

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indicesto model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509-529. https://doi.org/10.1080/00273170701382864

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*(1), 56–83. https://doi.org/10.1080/10705519909540119

Fazlıoğulları, O., & Kurul, N. (2012). Türkiye'deki eğitim bilimleri doktora tezlerinin özellikleri. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 12*(24), 43-75.

Fraenkel, J.R. & Wallen, N. (2005). *How todesign and evaluate research in education*. New York, NY: McGrawHill.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice, 33*, 14-18. http://dx.doi.org/10.1111/emip.12030

Göktaş, Y., Küçük, S., Aydemir, M., Telli, E., Arpacık, Ö., Yıldırım, G., & Reisoğlu, I. (2012). Türkiye'de eğitim teknolojileri araştırmalarındaki eğilimler: 2000-2009 dönemi makalelerinin içerik analizi. *Kuram ve Uygulamada Eğitim Bilimleri*, *12*(1), 177-199.

Gülbahar, Y., & Alper, A. (2009). Öğretim teknolojileri alanında yapılan araştırmalar. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *42*(2), 93-111. https://doi.org/10.1501/Egifak_0000001178

Hart, L. C., Smith, S. Z., Swars, S. L., & Smith, M. E. (2009). An examination of research methods in mathematics education: 1995–2005. *Journal of Mixed Methods Research*, *3*(1) 26–41. https://doi.org/10.1177/1558689808325771

Hazır Bıkmaz, F., Aksoy, E., Tatar, Ö., & Atak Altınyüzük, C. (2013). Eğitim programları ve öğretim alanında yapılan doktora tezlerine ait içerik çözümlemesi (1974-2009). *Eğitim ve Bilim, 38*(168), 288-303.

Hew, K. F., Kale, U., & Kim, N. (2007). Past research in instructional technology: Results of a content analysis of empirical studies published in three prominent instructional technology journals from the year 2000 through 2004. *Journal of Educational Computing Research*, *36*(3), 269-300. https://doi.org/10.2190/K3P8-8164-L56J-33W4

Hinkin, T. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*(5), 967-988. https://doi.org/10.1016/0149-2063(95)90050-0

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillside, NJ: Lawrence Erlbaum.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3-4), 117-144. https://doi.org/10.1080/03610739208253916

Hsu, T. (2005). Research methods and data analysis procedures used by educational researchers. *International Journal of Research & Method in Education*, *28*(2), 109–133. http://dx.doi.org/10.1080/01406720500256194

Hu L-T. & Bentler, M., P. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hu, L.T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to under parameterized model misspecification. *Psychological Methods, 3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

ITC (2005). *International test commission guidelines for test adaptation*. London: Author.

Ivanović L., & Ho Y. S. (2019) Highly cited articles in the education and educational research category in the Social Science Citation Index: A bibliometric analysis, *Educational Review*, 71(3), 277-286. https://doi.org/10.1080/00131911.2017.1415297

Kapuscinski, A. N., & Masters, K. S. (2010). The current status of measures of spirituality: A critical review of scale development. *Psychology of Religion and Spirituality, 2*(4), 191–205. https://doi.org/10.1037/a0020498

Karadağ, E. (2009). Eğitim bilimleri alanında yapılmış doktora tezlerinin tematik açıdan incelemesi. *Ahi Evran Üniversitesi Eğitim Fakültesi Dergisi*, *10*(3), 75-87.

Karasar N. (2017). *Bilimsel araştırma yöntemi: Kavramlar ilkeler teknikler*. Ankara: Nobel Yayıncılık.

Kazu, H., & Aslan, S. (2013). 2004 ilköğretim programının ölçme-değerlendirme boyutu ile ilgili yapılan araştırmaların değerlendirilmesi. *İlköğretim-Online*, *12*, 87-108.

Kazu, İ., Y., & Deniz, E. (2019). Öğretmenlerin ölçme ve değerlendirme tekniklerini kullanma durumlarını inceleyen araştırmaların değerlendirilmesi. *Journal of History School, 12*(XXXVIII) 174-195. https://doi.org/10.14225/Joh1527

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., & Levin, J. R. (1998). Statistical practices of educational researchers: Ananalysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of educational research*, *68*(3), 350-386. https://doi.org/10.3102/00346543068003350

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. *Journal of Experimental Education*, *69*(3), 280-309. https://doi.org/10.1080/00220970109599489

Kilbourne, W. E., & Beckmann, S.C., (1998). Review and critical assessment of research on marketing and the environment. *Journal of Marketing Management 14*(6), 513-32. https://doi.org/10.1362/026725798784867716

Kline, R. B. (2005). *Methodology in the social sciences. Principles and practice of structural equation modeling (2nd ed.).* New York: Guilford Press.

Küçükoğlu, A., & Ozan, C. (2013). Sınıf öğretmenliği alanındaki lisansüstü tezlere yönelik bir içerik analizi. *Uluslararası Avrasya Sosyal Bilimler Dergisi*, *4*(12), 27-47.

Ladhari, R. (2010). Developing e-service qualityscales: A literature review. *Journal of Retailing and Consumer Services, 17*, 464-477. https://doi.org/10.1016/j.jretconser.2010.06.003.

Lee, M.H., Wu, Y.T. & Tsai, C.C. (2009). Research trends in science education from 2003 to 2007: A content analysis of publications in selected journals. *International Journal of Science Education, 31*(15), 1999-2020. https://doi.org/10.1080/09500690802314876

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling, 12*(1), 1-27. https://doi.org/10.1207/s15328007sem1201_1

Lin, T. C., Lin, T. J., &Tsai, C. C. (2014). Research trends in science education from 2008 to 2012: A systematic content analysis of publications in selected journals. *International Journal of Science Education, 36*(8), 1346-1372. http://dx.doi.org/10.1080/09500693.2013.864428

Little, T., D. (1997). Mean and Covariance Structures (MACS) analyses of cross-cultural data: Practical and theoretical issues, *Multivariate Behavioral Research, 32*(1), 53-76. https://doi.org/10.1207/s15327906mbr3201_3

Mahler, C. (2011). *The effects of misspecifcation type and nuisance variables on the behaviors of population fit indicesused in structural equation modeling.* (Unpublished master dissertation The Faculty of Graduate Studies, University of British Columbia, Vancouver, Canada). Retrieved January 17, 2020, from https://open.library.ubc.ca/cIRcle/collections/ubctheses/24/items/1.0105120

Meredith W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge: Taylor & Francis Group.

Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2017). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica, 30(*3). https://doi.org/10.1186/s41155-016-0057-1

Murphy, J, Vriesenga, M., & Storey, V. (2007). Educational administration quarterly, 1979-2003: An analysis of types of work, methods of investigation, and influences. *Educational Administration Quarterly*, *43*(5), 612-628. https://doi.org/10.1177/0013161X07307796

Muthén, L. K., & Muthén, B. O. (2008). Mplus (Version 5.1) [Computer software]. Los Angeles: Muthén & Muthén.

Oxford Dictionary (2017). Retrieved January 15, 2020, from https://en.oxforddictionaries.com/definition/bibliometrics

Öncül, N. (2014). Türkiye'de erken çocuklukta özel eğitim ile ilgili yapılmış makalelerin gözden geçirilmesi. *International Journal of Early Childhood Special Education, 6*(2), 247-284.

Ören F. Ş. & Sarı, K. (2019). Web of Science veri tabanında fen eğitimi üzerine yapılan araştırmaya dayalı öğrenme stratejisi konulu çalışmaların değerlendirilmesi. *İlköğretim Online, 18*(4), 1875-1901. https://doi.org/10.17051/ilkonline.2019.639353

Özkan, M. (2016). Liderlik hangi sıfatları, nasıl alıyor? Liderlik konulu makalelerin incelenmesi. *Gaziantep Üniversitesi Sosyal Bilimler Dergisi, 15*(2), 615-639. https://doi.org/10.21547/jss.256732

Özyurt, Ö., & Özyurt, H. (2015) Learning style based individualized adaptive e-learning environments: Content analysis of the articles published from 2005 to 2014. *Computers in Human Behavior, 52*, 349-358. https://doi.org/10.1016/j.chb.2015.06.020

Patton, M. Q. (2002). *Qualitative research and evaluation methods.* Thousand Oaks, CA: Sage.

Saracaloğlu, A. S., & Dursun, F. (2010, Mayıs). Türkiye'de eğitim programları ve öğretim alanındaki lisansüstü tezlerinin incelenmesi. Paperpresented at the 1. Ulusal Eğitim Programları ve Öğretim Kongresi, Balıkesir Üniversitesi, Ayvalık. Retrieved October, 9, 2019, from https://www.pegem.net/akademi/kongrebildiri_detay.aspx?id=117909

Satorra, A., & Bentler, P., M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514. https://doi.org/10.1007/BF02296192

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210-222. https://doi.org/10.1016/j.hrmr.2008.03.003

Seçer, İ., Ay, İ., Ozan, C., & Yılmaz, B. Y. (2014). Rehberlik ve psikolojik danışma alanındaki araştırma eğilimleri: Bir içerik analizi. *Turkish Psychological Counseling and Guidance Journal, 5*(41), 49-60.

Selçuk, Z., Palancı, M., Kandemir, M. & Dündar, H. (2014). Eğitim ve bilim dergisinde yayınlanan araştırmaların eğilimleri: İçerik analizi. *Eğitim ve Bilim, 39*(173), 430-453. https://doi.org/10.15390/eb.v39i173.3278

Sırakaya M., & Alsancak Sırakaya, D. (2020). Augmented reality in STEM education: A systematic review. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2020.1722713

Small, H. (1999). Visualizing science by citation mapping for Information Science. *Journal of the American Society, 50*, 799-813. https://doi.org/10.1002/(SICI)1097-4571(1999)50:93.3.CO;2-7

Sözbilir, M., & Kutu, H. (2008). Development and currentstatus of scienceeducationresearch in Turkey. *Essays in Education, Special Issue*, 1-22.

Steenkamp, J., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*(1), 78-107. https://doi.org/10.1086/209528

Sveinbjornsdottir, S., & Thorsteinsson, E. B. (2008). Adolescent coping scales: A critical psychometric review. *Scandinavian Journal of Psychology, 49*(6), 533-548. https://doi.org/10.1111/j.1467-9450.2008.00669.x

Şahin, M. G., & Boztunç Öztürk, N. (2018). Eğitim alanında ölçek geliştirme süreci: Bir içerik analizi çalışması. *Kastamonu Eğitim Dergisi, 26*(1), 191-199. https://doi.org/10.24106/kefdergi.375863

Şenyurt, S., & Özer Özkan, Y. (2017). Eğitimde ölçme ve değerlendirme alanında yapılan yüksek lisans tezlerinin tematik ve metodolojik açıdan incelenmesi. *Elementary Education Online, 16*(2), 628-653. https://doi.org/10.17051/ilkonline.2017.304724

Tarman, B., Güven, C., & Aktaşlı, İ. (2011). Türkiye'de sosyal bilgiler eğitimi alanında yapılan doktora tezlerinin değerlendirilmesi ve alana katkıları. *Selçuk Üniversitesi Ahmet Keleşoğlu Eğitim Fakültesi Dergisi, 32*, 391-410.

Tavşancıl, E., & Aslan, E. (2001). *Sözel, yazılı ve diğer materyaller için içerik analizi ve uygulama örnekleri.* İstanbul: Epsilon Yayınları.

Tavşancıl, E., Çokluk, Ö., Çıtak, G., Kezer, F., Yıldırım, Ö., Bilican, S., Büyükturan, E., Şekercioğlu, G., Yalçın, N., Erdem, D., & Özmen, T. (2010). *Eğitim Bilimleri Enstitülerinde Tamamlanmış Lisansüstü Tezlerin İncelenmesi (2000-2008).* Ankara Üniversitesi Bilimsel Araştırma Projeleri. Retrieved June 9, 2019, from https://dspace.ankara.edu.tr/xmlui/handle/20.500.12575/68960

Tavşancıl, E., Güler, G., & Ayan, C. (2014, June). 2002-2012 yılları arasında Türkiye'de geliştirilen bazı tutum ölçeği geliştirme çalışmalarının ölçek geliştirme süreci açısından incelenmesi. Paperpresented at the IV. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi (Uluslararası Katılımlı) 9-13 Haziran, Hacettepe Üniversitesi, Ankara.

Tekin, H. (2008). *Eğitimde ölçme ve değerlendirme*. Ankara: Yargı Yayınevi.

Tsai, C.-C. & Wen, L.M.C. (2005). Research and trends in science education from 1998 to 2002: A content analysis of publication in selected journals. *International Journal of Science Education*, *27*, 3–14. https://doi.org/10.1080/0950069042000243727

Turan, S., Karadağ, E., Bektaş, F., & Yalçın, M. (2014). Türkiye'de eğitim yönetiminde bilgi üretimi: Kuram ve Uygulamada Eğitim Yönetimi Dergisi 2003-2013 yayınlarının incelenmesi. *Kuram ve Uygulamada Eğitim Yönetimi Dergisi, 20*(1), 93-119. https://doi.org/10.14527/kuey.2014.005

Turgut, M., F., (2006). *Eğitimde ölçme ve değerlendirme.* Ankara: Saydam Yayıncılık.

Ulutaş, B., Üner, S., Turan Oluk, N., Yalçın Çelik, A., & Akkuş, H. (2015). Türkiye'deki kimya eğitimi makalelerinin incelenmesi: 2000-2013. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 16*(2), 141-160.

Ulutaş, F. &Ubuz, B. (2008). Matematik eğitiminde araştırmalar ve eğilimler: 2000 ile 2006 yılları arası. *İlköğretim Online, 7*(3), 614-626.

vanEck, N. J., & Waltman, L. (2020). *VOSviewer manual. Manual for VOSviewer version 1.6.14, software documentation*. Leiden: Univeristeit Leiden.

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*, 4−69. https://doi.org/10.1177/109442810031002

Vega-Arce, M., Salas, G., Núñez-Ulloa, G., Pinto-Cortez, C., Fernandez, I. T., & Ho, Y.S. (2019). Research performance and trends in child sexual abuse research: A Science Citation Index Expanded-based analysis. *Scientometrics*, *121*(3), 1505-1525. https://doi.org/10.1007/s11192-019-03267-w

Wang, Y., Lai, N., Zuo, J., Chen, G., & Du, H. (2016) Characteristics and trends of research on waste-to-energy ıncineration: A bibliometric analysis, 1999–2015. *Renewable and Sustainable Energy Reviews, 66*, 95–104. http://doi.org/10.1016/j.rser.2016.07.006

White, R. (1997). Trends in research in science education. *Research in Science Education, 27*(2), 215-221. https://doi.org/10.1007/BF02461317

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substanceuse 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

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806–838. https://doi.org/10.1177/0011000006288127

Yalçın, S. (2016). Content analysis of research articles in measurement and evaluation journals. *Ankara University Journal of Faculty of Educational Science*, *49*(1), 65-84. https://doi.org/10.1501/Egifak_0000001375

Yalçın, S., Yavuz, H. Ç. ve İlgün Dibek, M. (2015). An examination of articles published in educational journals having highest impact factors: Content analysis. *Eğitim ve Bilim, 40*(182), 1-28.

Yang, Y. T., Iqbal, U., Ching, J. H. Y., Ting, J. B. S., Chiu, H. T., Tamashiro, H., & Hsu, Y. H. E. (2015). Trends in the growth of literature of telemedicine: A bibliometric analysis. *Computer Methods and Programs in Biomedicine, 122*(3), 471-479. https://doi.org/10.1016/j.cmpb.2015.09.008

Yılmaz, K., & Altınkurt, Y. (2012). An examination of articles published on preschool education in Turkey. *Kuram ve Uygulamada Eğitim Bilimleri, 12*(4), 3227-3241.

Zainuddin, Z., WahChu, S. K., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research, 30.* https://doi.org/10.1016/j.edurev.2020.100326

Zupic, I. (2015). Bibliometric methods in management and organization. *Organizational Research Methods, 18*(3), 429-472. https://doi.org/10.1177/1094428114562629

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 631–656
https://doi.org/10.21449/ijate.730454
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Design and Development of a Web Based Dynamic Assessment System to Increase Students' Learning Effectiveness

Arif Tuluk [1,*], **Halil Yurdugul** [2]

[1] Ministry of National Defense, 06550, Ankara, Turkey

[2] Hacettepe University, Faculty of Education, Computer Education & Instructional Technology Department, 06800, Ankara, Turkey

ARTICLE HISTORY

Received: May 01, 2020
Revised: Sep. 07, 2020
Accepted: Oct. 15, 2020

KEYWORDS

Web-based assessment,
Formative assessment,
Dynamic assessment,
Web-based dynamic assessment

Abstract: As Bloom (1984) stated that the way to increase students' achievements with one-to-one tutorial support can be the subject of instructional technologies today. In this context, the aim of the study was to design, develop and improve a web-based dynamic assessment system aimed at contributing to mathematics learning of secondary school students. For this purpose, a teaching environment has been designed in which 5th grade students can test themselves independently of time and place through internet technologies to learn mathematics lesson topics. A web-based environment combining the principles of dynamic assessment and gamification has been developed with the dynamic assessment algorithm developed within the scope of the research. The research is structured in accordance with the "Type 1" study features, one of the developmental research types. During the development of the prototype, pilot implementation was carried out with 47 5th grade students. In addition, focus group discussions were held with 12 students using the system. According to the findings obtained, it was concluded that the secondary school students' success scores constantly increase in each achievement, the instructional guidance they receive in each test decreases continuously and they can perform the next assessment tasks without help. The students stated that they were directed to the correct solution by the system while solving the questions they did not know, the instructional guidances for the solution were very helpful, the visual elements (monkey, robot and banana) were interesting, fun and beautiful and they wanted to use this system in other lessons.

1. INTRODUCTION

One of the main objectives of instructional technologies is to develop technological applications to facilitate and / or increase learning. As a result of his extensive research, Bloom (1984) emphasized that the basic conditions of increasing the learning on a subject by two standard deviations from the average are a) mastery learning model and together with b) one-to-one tutoring. Students need support in their learning processes, in other words, intervention. Especially considering that there is no teacher in new generation learning environments, digital systems should provide this support. The concept of support mentioned here has two different dimensions, respectively, the support in the learning process and the other is the support during problem solving. Supporting students in the problem solving phase is modeled with intelligent

CONTACT: Arif TULUK ✉ ariftuluk@gmail.com ✉ Ministry of National Defense, 06550, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

tutoring systems (ITS). However, the point to be emphasized here is that ITS and structured learning systems (e-learning systems) are different concepts. To explain this difference, the students' learning systems at school (learning systems) and at institutions such as classrooms and courses (tutoring systems) can be given as examples. While learning systems (such as learning management systems) offer enriched learning environments for the purposes of the course and based on the instructional design, the main purpose in ITS is to support students in the problem solving process. In this context, the support offered to students in problem solving situations can be provided through a dynamic assessment system.

As Bloom (1984) stated that the way to increase students' achievements with one-to-one tutorial support can be the subject of education technologies today. In this context, web-based dynamic assessment systems that can provide instructional support to learners can be made thanks to instructional technologies. It is seen in the literature that many technology supported environment designs have been developed in order to support students (Ashton, et al. 2006; Costa, Kothe, Mullan & Butow, 2010; Critchley, Ware, Kumta & Wong, 2009; Marinagi, 2011; Wang, 2007, 2010, 2011, 2014; Wang, Wang, Wang & Huang, 2006; Zou & Zhang, 2013). When the studies conducted to increase student success are examined, there are differences in the designs of these studies; some studies have included gamified dynamic assessment items (eg, a false option is reduced, I want to ask my friend option, leader board, etc.), while others (for example, algorithms based on the assessment task and the scaffolding relationship) have been differences in the algorithms used. It can be argued that it is important to examine the algorithms in gamification elements and assessment task-guiding relations. In the studies analyzed in this context, no study with web-based dynamic assessment has been encountered in our country. It can be suggested that an e- assessment system, which supports the out-of-school learning experiences in addition to the in-school learning experiences of the students at the target audience level (secondary school level), includes different instructional guidances according to different types of knowledge. In addition to the rapid increase in the use of web-based technologies for learner-oriented assessments in higher education, the use of such assessment systems, especially in classroom environments at the K-12 level, is new. Within the scope of the research, it is thought that the literature will contribute to secondary school students in terms of providing them with the opportunity to monitor their own learning and development instantly by using the web based dynamic assessment system. In addition, it is thought that it will contribute to the literature in terms of the design model proposed by the research according to its educational purpose. Other issues that the research is thought to contribute to the literature are as follows:

- The research is a study carried out in terms of making a significant contribution to the achievement of the educational objectives of the students at the secondary school level by the web-based dynamic assessment system.
- In web-based assessment systems, there are serious problems in keeping the learners in the system and ensuring their continuity, and these results in the learners leaving the system. In this research, there are gamification elements to keep the learners in the system and ensure their continuity in each acquisition test process. With this aspect, it is a study carried out in order to ensure the permanence and continuity of the learners by embedding gamification elements into the system.
- The research is a study in which 5th grade students at K-12 level at K-12 at the level of web-based dynamic assessment can take instructional guides as instructors in the process and improve themselves continuously.

The research is a study that shows whether 5th grade students at secondary school level have an effective learning experience, whether there are assessment items that are incorrectly answered as a result of carelessness, if there are lesson topics that are deficient, and that they can eliminate the missing subjects through instructive measures and show their educational development levels.

Research Questions

Within the scope of the research, assessment environments and processes have been designed with the dynamic assessment system based on web-based teaching developed. With this environment, it has been tried to contribute to the learning and teaching process with the processes developed to identify and eliminate the deficiencies in learning mathematics. In this research, it is aimed to examine how to design a web-based dynamic assessment system for students at secondary school level to learn mathematics. For this purpose, the following questions were sought in this research;

1. For the design process of a web-based dynamic assessment system,
    a. What is needed to develop a web based dynamic assessment system according to the literature?
    b. What should be the features and functions of a web-based dynamic assessment system?
    c. What are the issues that need to be developed in the web-based dynamic assessment system?
    d. What are the design changes based on the assessment of the developed system?
2. What are the students' views on the web-based dynamic assessment system?

2. REVIEW OF THE LITERATURE

This topic covers the concept of dynamic assessment and the design of the web-based dynamic assessment system.

2.1. Dynamic Assessment

The focus of formative assessment is very different from summative assessment. According to Haywood, Brown and Wingenfeld (1990), summative assessment is used to assess performance at a given moment and there is no attempt to change this performance. Formative assessment aims at revealing students' learning deficits and difficulties leading to these deficiencies. Dynamic assessment is a type of formative assessment, a general and inclusive concept that is used to explain different types of approaches, involves teaching, is presented in the assessment process of feedback and differs on the basis of individual performances (varying according to the individual's performance). One of the main objectives of dynamic assessment is to improve the performance of learners by providing instructional aids with assessment tasks.

The development of dynamic assessment has been heavily influenced by L.S. Vygotsky (Allal & Ducrey, 2000; Haywood, et al., 1990). Vygotsky (1978) stated that, with the help of an adult, the performance level of children can be improved. Vygotsky proposed the theory of "The zone of proximal development (ZPD)" to describe the difference between children's performance without the help of adults and more competent peers. ZPD represents children's learning potential. By interacting with adults or more competent peers, children's learning potential can be revealed and learning activities can also be improved. By interacting with the dynamic assessment system, the learners try to improve their learning performance with the help of tutorials in the assessment tasks presented to them. In this context, it is thought that learners will contribute to the ZPD by developing their learning potential with the dynamic assessment system.

Dynamic assessment is an interactive assessment commonly given as “test-teach-retest” (Haywood & Lidz, 2007; Moore-Brown et al., 2006). It has been revealed that the researchers who examined the formats of dynamic assessment and related research findings had various opinions on how to effectively manage dynamic assessment (Elliott, 2003; Sternberg & Grigorenko, 2001). Two basic instructional features were shared on this subject:

- Individuals are given the opportunity to learn (Bransford, et al., 1987).
- Teaching and feedback are included in the testing process (Elliott, 2003).

In this context, it is possible to divide the dynamic assessment into two as interactionist and interventionist dynamic assessment. In interactionist dynamic assessment, measurement is avoided and qualitative assessment of a person's learning potential is concerned. It aims to provide the failed individual with a necessary understanding of the assessment items and possible ways of resolving these assessment items. Then, after the individual has developed the ability to solve this task (through mediation work with the teacher or trainer), their ability to overcome similar tasks is evaluated. Interventionist dynamic assessment includes measurable pre-programmed help and is directed to measurable psychometric measurement. In other words, aid takes the form of standardized interventions developed to measure the capacity of individuals or groups to use predetermined guidance, feedback and support to "increase predictive validity of the assessment process".

According to Sternberg and Grigorenko (2001), dynamic assessment has two forms: sandwich form and cake form. Both formats are implemented as "test-teach-retest". Sandwich formatted dynamic assessment means that teaching is held between pre-test and post-test, thus creating a sandwich-like process. Teaching in cake-type dynamic assessment is a response to the researchers' answers to each question. The main difference between sandwich form dynamic assessment and cake form dynamic assessment is that education and assessment are separate from dynamic assessment in sandwich format but are combined with dynamic assessment in cake form.

2.2. Web Based Dynamic Assessment System

The web-based dynamic assessment system adopted in this research was designed as a cake form interventionist dynamic assessment. The web-based dynamic assessment system is a learner-oriented system, and it is expected that students can determine whether they have performed effective learning, whether there are questions that they answer incorrectly as a result of carelessness, whether there are issues missing, to contribute to achieving their educational goals, and to improve learning strategies and to organize learning processes.

The main feature of dynamic assessment in the form of cakes is the design of consecutive clues with a series of progressive clues. With the designed system, students are provided with gradual instructional guidance (IG) regarding their problems regarding assessment tasks. In this design, reference is made to the "graduated prompt approach" proposed by Campione and Brown (1985, 1987) and the mathematical problem solving theory of Mayer (1992). According to Campione and Brown, hints in the phased graduated approach are presented in a predetermined order, arranged according to the level of disclosure (Bransford, et al., 1987). They begin with "general tips" and gradually become "specific tips". General hints offer relatively little specific information about the solution, while a specific hint provides a detailed instruction in which students can produce the correct answer (Campione & Brown, 1985, 1987). Mayer argued that solving mathematical problems involves two sub-processes, namely "problem representation" and "problem solving", and "translation", "integration", "planning and monitoring" and "execution". Three IGs (IG1, IG2, IG3) are designed to compensate students for missing information at each stage and sub-process. With these IGs, the necessary mathematical problem solving knowledge is expected to compensate for the lack of learners, but also improve their learning. In this research, the “graduated prompt approach” and the tips provided by the dynamic assessment elements were used as a reference to develop the web-based dynamic assessment system. These clues are called instructional guidance (IG), as they are used to guide and teach students. When students answer an item incorrectly, they gradually receive IGs and learn to find the right answer step by step.

3. METHOD

This part covers the research model, the participants of the research, the design and development of the web-based dynamic assessment system, data collection process, data collection tools and data analysis.

3.1. Research Model

In this research, the design, development and improvement processes of the web based dynamic assessment system are planned as "developmental research". Developmental research are considered as two types; Design, development and assessment of a particular product (Type 1), focusing on specific design, development and assessment processes, tools or models (Type 2) (Richey, Klein & Nelson, 2004). While the researcher has the role of both researcher and designer in a special development context in Type 1, the researcher has the role of implementing only one tool / model in Type 2 (Van den Akker, 1999). This research is structured in accordance with the characteristics of "Type 1" study, one of the developmental research types. Working process in the design, development and improvement of the web-based dynamic assessment system; starting with the analysis of the problem, the development of systemic solutions (prototype creation-design), testing of solutions (use of prototype-research) and improvement processes were carried out. Summary information about the stages followed in this research, the operations performed at these stages, data sources and output are presented in Table 1.

Table 1. *Stages of the Research, Processes, Data Sources and Outputs*

Stages of the Research	Operations	Data Sources	Output
Requirement Analysis	Examining the related literature (systematic content analysis)	Literature (98 articles)	Requirement list
Design (Creating the Prototype)	- Creating a prototype in accordance with the elements and components planned to be included in the system design. - Prototype testing and revision	Expert opinions	- System components - System elements and features - System prototype
Development (Using the Prototype)	- Pilot implementation of the prepared prototype - Focus group meeting	Target group	Learning environment (final product)
Improvement	- Focus group meeting - Making design changes based on assessment	Target group	Learning environment (improved product)

As seen in Table 1, the research was carried out in a formative manner in four basic stages. The research process started with literature review, referring to what was done at each stage of this research and the syllabus given in Table 1, and finalized the process of designing, developing and improving the web-based dynamic assessment system. The research was completed within eight months.

3.2. Participants

This research was carried out together with field experts and 5th grade students in accordance with the design principles. During the design, development and improvement of the web-based dynamic assessment system prototype, semi-structured interviews were made with working groups continuously. In this research, the information about the period when the data was collected, the people with which the data were collected, their frequencies, data sources, data collection tools and research problem numbers are presented in Table 2.

Table 2. *Information on the Data Collection Process*

Process Name	Time	Participants	Frequency	Data Sources	Data Collection Tools	Research Problem No
Requirement Analysis	2017-2019	Related Studies	98	Literature	Systematic descriptive scanning	Sub-Problem No:1a
Design (Creating the Prototype)	2017-2018 academic year fall and spring semester	Field experts	7	Expert opinions	Getting expert opinions	Sub-Problem No.:1b
Development (Using the Prototype)	2017-2018 academic year spring semester	Secondary school 5th grade students	47	Secondary school students	Use of the system by students and focus group discussions	Sub-Problem No.:1c
Improvement	2018-2019 academic year fall semester	Secondary school 5th grade students	12	Secondary school students	Focus group discussions	Sub-Problem No.:2

Within the scope of the requirement analysis for the design and development of the web-based dynamic assessment system, the studies conducted on the dynamic assessment and static assessment systems developed for e-learning and personalized learning environments in the literature were analyzed by systematic review. In the process of establishing the web-based dynamic assessment system prototype, the opinions of 3 experts from the field of computer and instructional technologies, 2 as subject matter experts for the course contents in mathematics lesson and 2 from the field of measurement and assessment were consulted. In the process of using the prototype, the web-based dynamic assessment system has been presented as a pilot application to 47 5th grade students in a public school in a province of İzmir under the Ministry of National Education. In the process of improving the web-based dynamic assessment system, the system was introduced to 12 5th grade students in a public secondary school and focus group discussions were held with these students. For the students who participated in the research, the students were allowed to participate on a voluntary basis by obtaining permission from the ethics committee commission.

3.3. Design and Development of Web-Based Dynamic Assessment System

In this section, the steps taken during the design and development of the web-based dynamic assessment system are described. The purpose of this process is to develop effective software by using resources efficiently. For this purpose, the Rapid Prototyping Model of Tripp and Bichelmeyer (1990) was taken as reference within the scope of its fitness for purpose in the design of the education system for the design and development of the system. In this model, first a prototype was produced, then this prototype was tested, necessary corrections were made in accordance with the available data, and the system was completed when it reached to the point without error. There are four main processes in the Reference Rapid Prototyping Model. These processes,

- requirement analysis,
- creating the prototype (design),
- using the prototype (development),
- improving of the system.

3.3.1. *Stage 1: Requirement analysis*

Within the scope of the requirement analysis for the design and development of the web-based dynamic assessment system, studies on the dynamic assessment and static assessment systems developed for e-learning and personalized learning environments have been examined. In order to access related studies, first of all, keywords that are suitable for the purpose of the research and the problem sentence have been determined. Keywords were determined as web-based assessment, formative assessment, dynamic assessment. The determined keywords were scanned by creating different search queries in academic databases (ERIC, JSTOR, Science Direct, Scopus, Springer, Taylor & Francis, Web of Science). By using search queries, related studies have been reached by searching through relevant international academic databases, printed and electronic journals and books, Google Scholar and various academic social networks (Academia, ResearchGate). As a result of the examination of the studies in the literature, the need to design and develop a web-based dynamic assessment system has emerged. In this context, it was decided to support this research with algorithms based on the interventionist dynamic assessment process by examining the interventionist and interactive dynamic assessment approaches and considering the research problem. Accordingly, the learner-instructional guidance interaction algorithm has been developed (Figure 1).

As indicated in Figure 2, in cake-form interventionist dynamic assessment, the assessment is always carried out in an individualized way. The learners who are subject to the test receive instructions by answering the assessment items one after another. When learners answer an assessment item incorrectly, they see a series of consecutive instructional guidances. These consecutive instructional guidances are designed to help progressively clarify the answer. The instructional guidances are teaching and teaching activities continue until the assessment item is answered correctly. As a result of the literature reviews, it has been determined in studies conducted abroad that similar systems, albeit few, are used in limited numbers. In this research, students were given the right to repeat the same item (by offering progressive instructional guidance each time). Thus, a system design has been designed according to the characteristics of the target audience, with a choice of four options, with only one option left, in other words, three times according to the wrong answer situation. One of the aspects of this research that is different from the studies in the literature is this feature.

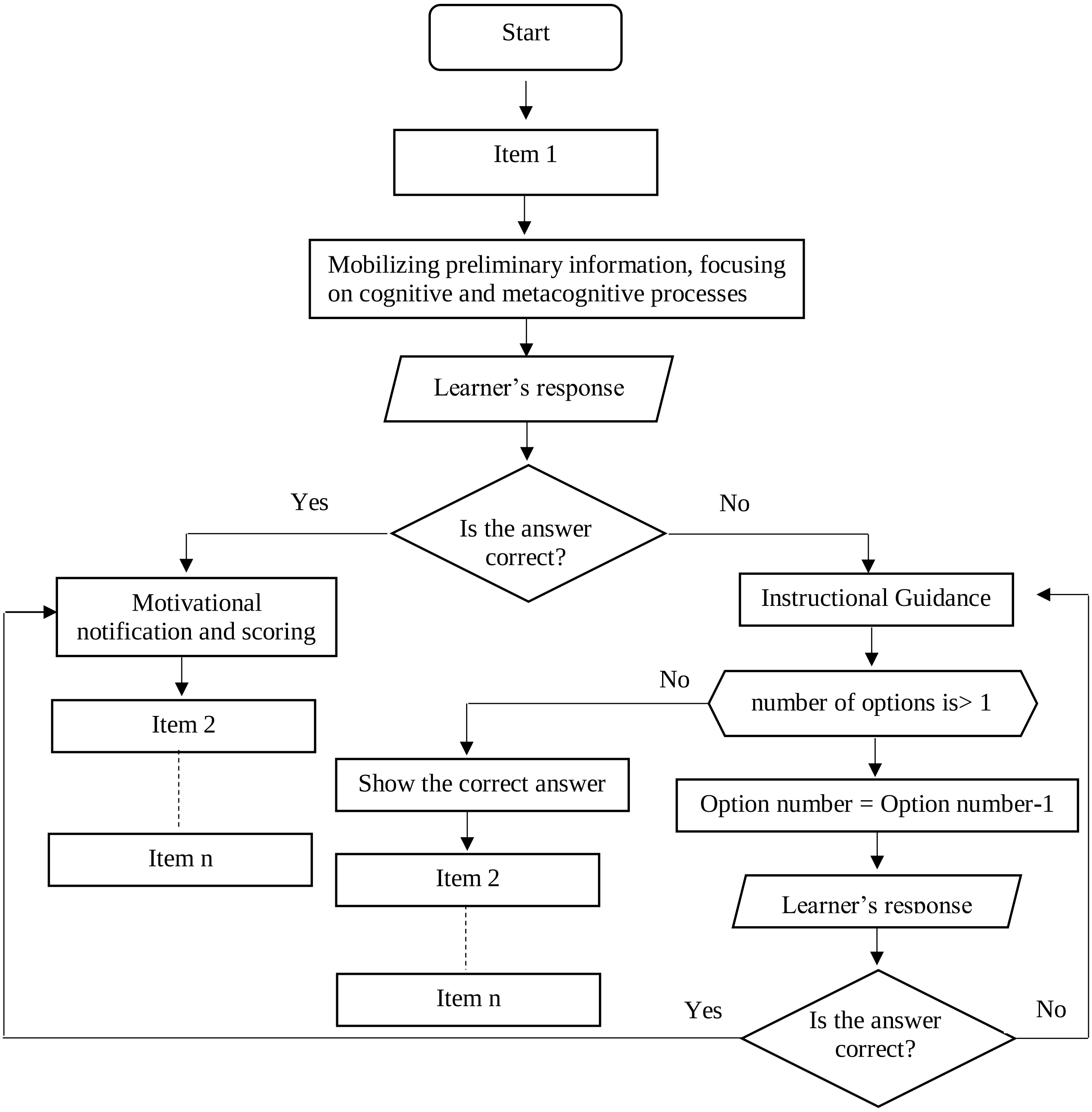


Figure 1. *Learner-instructional guidance interaction algorithm*

3.3.2. *Stage 2: Creating the prototype (design)*

As a result of the systematic descriptive review of the literature reviews regarding the design of the web-based dynamic assessment system, the studies of the design of the system and the creation of a prototype have been started in order to achieve the objectives of the system to be developed. In this direction, studies on the features of the system to be developed, the creation of contents, the creation of user interfaces and the drafting of assessment tools have been carried out. Angular 5, material design and bootstrap technologies were used on the front of the system, and nodejs, expressjs and mongodb technologies were used as the database on the back of the system.

3.3.3. *Stage 3: Using the prototype (development)*

After the design of the web-based dynamic assessment system, expert opinions were consulted. In this context, the opinions of 3 experts from the field of computer and instructional technologies for the development of the environment, 2 experts from the field of mathematics for the course contents and 2 experts from the field of measurement and assessment were consulted. Interviews were conducted with experts to obtain their opinions using a semi-structured interview form. System design has been developed in line with the opinions of the field experts regarding the suitability of the design to the dynamic assessment approach, the suitability of the feedbacks developed, the suitability of the interface and visual elements to the target audience, and the suitability of the number of tests and questions. In this context, the appropriateness of the design has been provided by examining the content validity (the content of validity ratio: 1.0). The web-based dynamic assessment system, the prototype of which has been developed and developed in line with expert opinions, has been presented as a pilot application to 47 5th grade students in a public school in İzmir province under the Ministry of National Education. After the pilot implementation of the web-based dynamic assessment system, a three-dimensional focus group interview was made with 18 5th grade students' perception of satisfaction and benefit, motivation and engagement, and interface design. During the focus group interview, a semi-structured interview form developed by the researcher was used. The interviews were analyzed by content analysis and the system was evaluated. Themes and frequencies obtained as a result of content analysis are given in Table 5. Following the pilot implementation, studies on the development of the web-based dynamic assessment system continued in line with the feedback received from the students.

3.3.4. *Stage 4: System improvement*

Following the development studies, focus group discussions were held with 12 students prior to the experimental study. In line with the feedback received from the students, efforts to improve the web-based dynamic assessment system continued. As the final product, the system was configured on a server by taking a domain name. System backups were taken periodically.

3.4. Data Collection Process

The data collection process of the research started with literature review. As a result of the scanning, 98 articles were examined. Systematic descriptive scanning of the determined articles was made. During the design phase, expert opinion was consulted in structuring the subject contents, revealing the user interfaces and preparing the first drafts of the assessment tools. During the design phase, development focuses were determined in step cycles and these issues were studied. The design focus areas include the interface design of educational web based dynamic assessment system software, user roles and their control design, the design of educational visual elements and the principles of using web based technologies. In design-based research, the improvement process was carried out within the framework of step cycles. The validity and reliability of the subject acquisition items prepared by the experts in the scope of the design of assessment tasks and the creation of the item pool were checked. Cross-checks were provided by three different groups of experts: language expression experts, measurement and assessment experts, and subject matter experts.

As e-learning materials, acquisition-based tests for teaching mathematics subjects of 5th grade students were prepared on the basis of 5th grade mathematics curriculum published by the Ministry of National Education Board of Education and Discipline. Within the scope of the research, “Numbers and Operations” learning area, “Reading and Writing Numbers”, “Digit Numbers and Values”, “Number and Figure Patterns”, “Addition-Subtraction Estimation” and “Addition-Subtraction Process” acquisitions were presented. A total of 15 tests were directed towards five determined acquisitions (there are three multiple choice tests with 10 assessment items in one acquisition). Gamification elements are included to keep students in the system during each acquisition test process and to ensure their continuity in the assessment tasks.

Students can exit the assessment task at any time. Learners can continue their assessment tasks from where they left off.

3.5. Data Collection Tools

Data collection tools used in the research are semi-structured interview form, focus group interview form, achievement test, log recordings and audio and video recordings.

3.6. Data Analysis

The data obtained in the research were analyzed using quantitative and qualitative data analysis techniques. Content analysis was used in the analysis of qualitative data. The main purpose in content analysis is to reach the concepts that can clearly describe the data collected and the relationships between these concepts. The most important process in content analysis is to gather similar data together under certain themes and concepts and to illuminate them by organizing them in a way that the reader can understand (Yıldırım & Şimşek, 2011). The data obtained as a result of the interviews were first transferred to a written environment on a computer and no spelling and punctuation correction was made on any sentence and word. The data obtained was coded and tables specific to each question were created separately and similar or identical themes in the statements of the participants were brought together on a common denominator. Regarding the validity of the content analysis, the data obtained from the interviews were evaluated only by the researcher, and interpreted and cross-checked with another expert from the field. The frequencies of the data are presented through the tables.

The quantitative data obtained in this research include summative assessment pretest scores, summative assessment posttest scores, and student response histories in the web-based dynamic assessment system. In the analysis of quantitative data, descriptive statistics such as percentage, frequency, arithmetic mean were used.

In order to ensure the internal validity of the research, the interview forms and assessment items used within the scope of the research were applied to all learners in the same environment and at the same time and were applied by the researcher. The data obtained from interview forms and assessment items were assessed only by the researcher and interpreted after being examined with another expert from the field. Considering that the research is carried out with the working group, a limited generalization can be made with the results obtained. In this case, the results can only be generalized to groups of students whose educational level corresponds to that of the study group. During the research process, there was no loss of participants that would threaten external validity.

4. FINDINGS

In this section, the findings of the research obtained in the order of sub-problems related to the analysis of the data and the assessments related to these findings are presented.

4.1. Sub-Problem 1: Web-Based Dynamic Assessment System Design Process

4.1.1. *Findings at the requirement analysis stage*

In order to determine the requirement situation for the development of a web-based assessment system, studies on this subject in the literature have been examined. The studies on the dynamic assessment and static assessment systems developed for e-learning and personalized learning environments in the literature were examined between 2000 and 2019 (Table 3).

Table 3. *Information on Articles Published in the Field Between 2000-2019*

Source Name	Dynamic Assessment	Formative Assessment	Web-Based Assessment
Science Direct	257	283	39
Springer	211	296	157
Taylor & Francis	52	322	42
Jstor	16	111	17
Eric	42	309	15
DergiPark	14	61	3
Ulakbim TR Directory Dizin	5	31	2
Total	**597**	**1413**	**275**

As it can be seen in Table 3, when the articles published in the literature between 2000 and 2019 are scanned according to keywords, 597 articles with dynamic assessment, 1413 articles with formative assessment, 275 articles with web-based assessment are published. It was found that most of the research were carried out using formative assessment. Within the scope of the research, 98 articles appropriate for the context of the research, the purpose of the research and the problem sentence were selected from among the resources accessed in the literature. It has been determined that 32 (32.65%) of the assessment studies used in the articles assessed were worked with dynamic assessment. 28 of the 32 articles studied with dynamic assessment were published in English and 4 were published in Turkish. 1 of 4 articles published in Turkish is scale development, 2 are literature review and 1 is qualitative descriptive analysis. The sample level distribution used in the articles examined within the scope of the research is given in Figure 2.

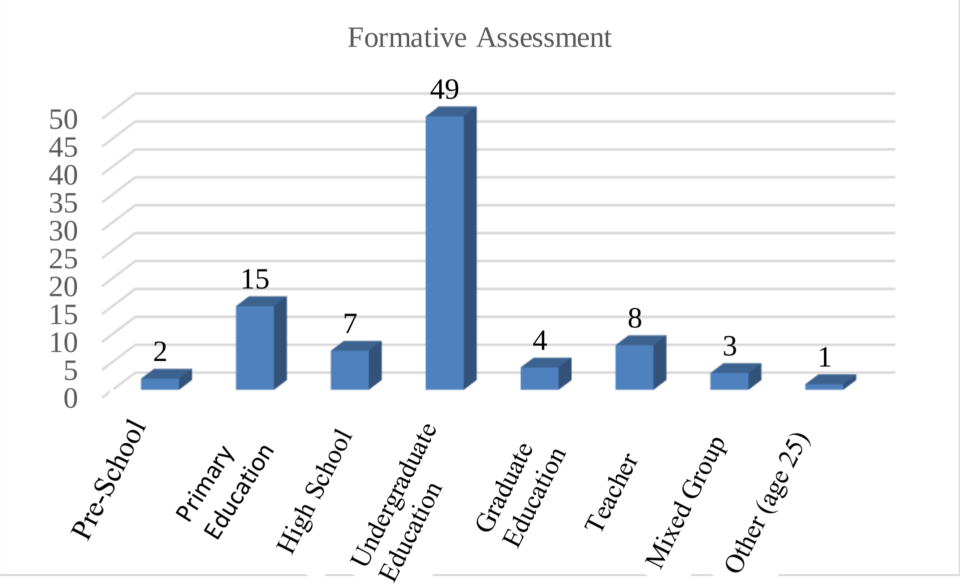


Figure 2. *Sampling level distribution used in articles (formative assessment)*

As seen in Figure 2, when the sample level distribution used in the articles examined within the scope of the study is examined, it was found that formative assessment was studied mostly at the undergraduate (49 articles, 52%), at least other (25-age group course, 1 article) and pre-school (2 articles) level. The sampling level distribution of the articles studied in dynamic assessment is presented in Figure 3.

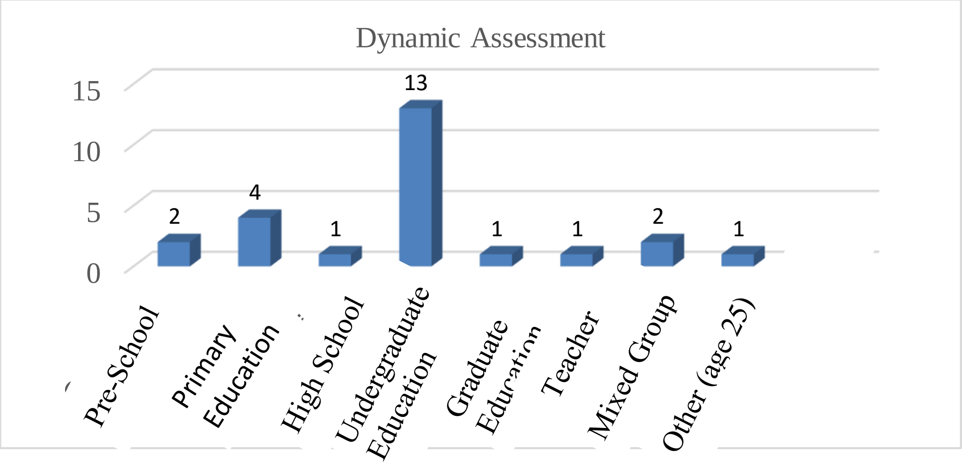


Figure 3. *Sampling level distribution used in articles published in the literature (Dynamic assessment)*

When the sample level distribution is analyzed in the articles in which dynamic assessment is preferred, it is seen that the highest level is studied with the undergraduate (13 articles) level, at least is high school (1 article), graduate (1 article), teacher (1 article) and other (25 years old) (1 article) level. It was stated in 25 (78.13%) of 32 articles, in which dynamic assessment study was carried out, that dynamic assessment increased success and was effective. In 7 articles (21.87%), the variable of success was not examined (review of the literature, scale adaptation study, etc.). It was found that 19 of 25 dynamic assessment articles (76%) were made with interactionist dynamic assessment, and 6 (24%) were made with interventionist dynamic assessment. In addition, 20 of these articles (80%) are intended to support in-school learning experiences, and 5 of them (20%) are intended to support out-of-school learning experiences. Conceptual information was presented in 19 (76%), conceptual-procedural-strategic information in 5 (20%) and conceptual and procedural information in 1 (4%) for learners.

When studies conducted with dynamic assessment in Turkey are examined, only 4 studies were encountered. On the other hand, it was determined that scale development in 1 article, literature review in 2 articles and qualitative descriptive analysis in 1 article were conducted. However, when the studies conducted in the literature are examined, no study conducted in Turkey based on web-based dynamic assessment has been encountered. However, Kılıç (2004) and Deniz Kan (2007) pointed out the use of portfolios as a dynamic assessment tool in some of their studies. It is determined that many studies have been carried out abroad with the dynamic assessment approach.

When the studies in this direction are examined, there are differences in the designs of these studies. In some studies, there are dynamic assessment elements (such as the option I want to ask my friend, leader board, etc.) while others (for example; assessment task and path. It has been observed that there are differences in algorithms based on scaffolding relationship) are differences in the algorithms used. In this research, it was decided to differentiate the algorithms in assessment task-guiding relations by retaining gamification elements. As a result of the literature review, it has been determined that similar systems, albeit few, are used in a limited number of instructional guidances. In this research, students were given the right to repeat the same item (by offering gradual instructional guidance each time). In this way, a system design has been made according to the characteristics of the target audience, with a choice of four options, with only one option left. In other words, three times according to the wrong answer situation. One of the aspects of this research that is different from the studies in the literature is this feature. In the e-assessment system, which was designed by using similar dynamic assessment systems, points system and gamification elements were included in order to increase learning motivation. The unique aspect of this system on a national basis is that, in addition to the limited number of studies abroad, such a system has not yet been configured in our country. The unique aspect of the design in this research, which is international in scale, is that it has a unique quality compared to its peers, with the ability to offer students repeated instruction and instructional guidance in each retry. Another point to be emphasized here is inspired by the trial approach until we find the correct answer, which is a type of feedback. In this context, it was decided to support this research with algorithms based on interventionist dynamic assessment process. Accordingly, the learner-instructional guidance interaction algorithm has been developed (Figure 1).

In the light of the information obtained above, as a result of the study of literature studies, the lack of a study in the context of Vygotsky's sociocultural learning theory (ZPD) and the system design, development and improvement using dynamic assessment in our country, the dynamic assessment can also be used at the target audience level (secondary school level) in our country. In order to design, develop and improve the web baed dynamic assessment system, a unique system should be designed considering the following:

- Supporting students' learning out-of-school in addition to their classroom learning,
- Establishing an environment where students can work outside the class hours regardless of location and time,
- Providing students with an environment where they can work in accordance with their own pace,
- Monitoring and analyzing students’ learning process

Based on the learner-instructional guidance interaction algorithm given in Figure 1, considering the characteristics of the target audience through a unique system to be developed as a result of literature reviews and systematic descriptive analysis for the design, development and improvement of the web-based dynamic assessment system, it was decided to go through the stages of determining the system features and creating the prototype.

4.1.2. *Findings at the design (creation of the prototype) stage*

In this section, the design, development and features of the web based dynamic assessment system are explained.

4.1.2.1. *Web based dynamic assessment system software design process*

Within the scope of the research, it is possible for secondary school students to test themselves with the flexibility of time and space through the internet technologies for learning mathematics lesson topics, to be able to interact with a system that is an instructional approach while testing themselves, to obtain more objective measurements, a quick and instant way towards student performance. A web-based dynamic assessment system has been designed to improve learning and teaching by giving demonstrators, to monitor and analyze assessment results. The architecture of the designed system is given in Figure 4.

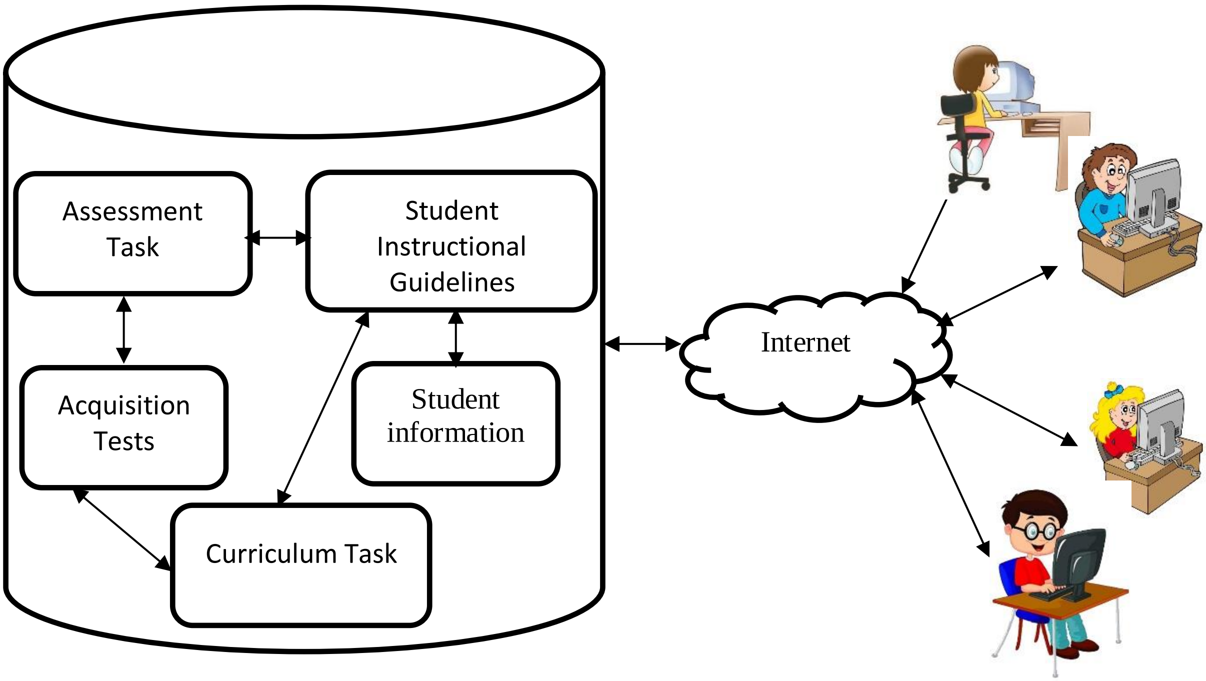


Figure 4. *Web-based assessment system architecture*

Assessment tasks for the dynamic assessment system were prepared according to the 5th grade mathematics curriculum and added to the acquisition tests in the system. Within the scope of the acquisition tests, 5 acquisition tests were prepared, 3 dynamically designed tests for each acquisition. Students can access the system developed to perform these assessment tasks via the internet and log in with the user information provided to them or by registering the system. Students who log in to the system can perform the assessment tasks that are managed online and receive instructional guidance on the basis of test, acquisition and item for assessment results. Assessment results are instantly calculated with the functions of the programming languages used in the development of the system and presented to students. Acquisition-based test results are presented to students in detail as instructional guidelines at the end of each test. In addition, instructional guidelines regarding the acquisition and test results are available on each student's profile page.

4.1.2.2. *Development of web based dynamic assessment system*

When the individual differences are considered, it is important that the feedbacks are individual. However, it is very difficult for educators to determine the deficiencies of the learner and give feedback to the individual (Boud, 2000). Technology-supported environments in terms of feedback should be provided in order for the learners to be able to see his / her current status by testing himself / herself. However, it is noteworthy that there is no common design in creating road maps for assessment systems and feedbacks for web-based formatting in this process where developments continue. In addition, when the distance education studies conducted in Turkey are examined, it is seen that there is not enough focus on e- assessment, and in the studies conducted, the feedbacks are provided to the teacher and the institution rather than the learner. However, there is a need for e-assessment environments where the learner can self-test to observe his own development (Boud, 2000) and make judgments. When the studies carried out using the dynamic assessment method in Turkey are examined, it is found that these studies are quite limited and there is no study conducted with students at secondary school level. Based on this point, a web-based dynamic assessment system has been developed.

Angular 5, material design and bootstrap technologies were used in the development of the web based dynamic assessment system, nodejs, expressjs and mongodb technologies were used as database. The developed systems are configured on a Linux based server that is rented through a hosting company. Considering the fact that many students have instant access to the system at the same time, the server is configured to allow 1 terabyte bandwidth for the system to operate uninterruptedly and quickly. After the installation process on the server has been completed, technical tests of the system have been carried out. Following the completion of the tests, the system was made available to students. Within the scope of the research, a simple and easy-to-use design was preferred considering the characteristics of the target audience and the system was developed to work on personal computers and tablets.

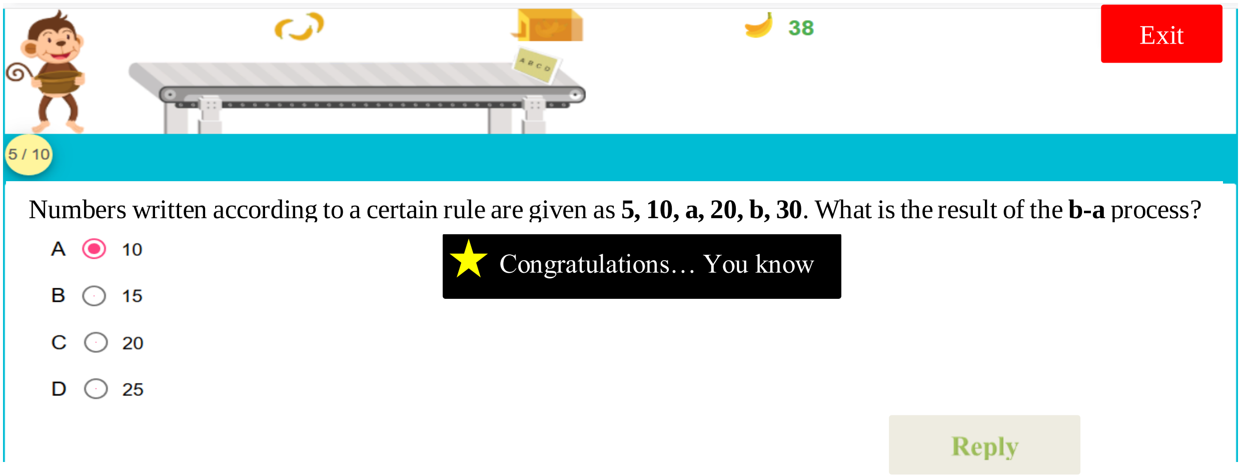
4.1.2.3. *Features of the web-based dynamic assessment system*

The web-based dynamic assessment system is a student-centered system, and students can instantly see their achievements for the tests they have taken, examine the assessment tasks, receive instructional guidance that have already been entered into the system, and thus organize their learning instantly. In this system, there are acquisition-based tests prepared for the learning of mathematics lesson subjects of 5th grade students based on primary mathematics curriculum.

Web-based dynamic assessment system can be entered in three different roles: student, teacher and administrator. The administrator can add students (users) to the system, update, delete, add gains suitable for the learning area and sub-learning areas in the curriculum, create, update, delete, create item pool for each acquisition, update, delete. The researcher has assumed the role of manager in the system. The user in the role of teacher can add, update, delete, create item pool for each acquisition, update, delete, and receive detailed reports of each assessment item created by the item pool in accordance with the learning area and sub-learning areas in the curriculum. On the other hand, students can make their own assessments by following the tests they have taken from the list of acquisitions according to the learning area from the profile page, and take dynamic assessment tasks for a new acquisition.

In web-based assessment systems, there are serious problems in keeping the learners in the system and ensuring their continuity, and this results in the learners leaving the system in a short time. In this context, gamification elements are included to keep students in the system in each acquisition test process and to ensure their continuity in the assessment tasks. If the students answer correctly without any guidance (instructional guidance) to the assessment item directed to them, 3 bananas are sent from the box on the tape to the monkey to represent that the item was answered with full score (10 points). Bananas collected in the monkey's basket are instantly reflected in the success score received. If the students respond correctly by taking one instructional guidance to the assessment item, 2 bananas are sent from the box on the tape to the monkey, representing that the item knows correctly (8 points) by receiving one instructional guidance. If students respond to the assessment item by taking two instructional guidances, one banana is sent from the box on the tape to the monkey, representing that the item knows correctly (6 points) by receiving two instructional guidances. If students respond to the assessment item with three instructional guidances, no bananas are sent from the box on the tape to the monkey to represent that the item responded by receiving three instructional guidances (0 points). With this aspect, it is aimed to ensure the permanence and continuity of the students in the system by embedding gamification elements into the system. A sample assessment item page view is given in Figure 5.

Within the scope of the tests they have taken for the assessment tasks directed to them, three instructional guidances are offered gradually in the system when they encounter difficulties in their learning lives or experience problems in their performance. Instructional guidances are given gradually through an equivalent example without an accurate answer to the assessment item accompanied by a moving robot. If the students respond correctly to the assessment item directed to them, they are supported with expressions to increase their motivation as "Congratulations ... You know.", "Bravo", "Excellent", "Great", etc. before proceeding to the next assessment item. After the tests they have taken for the assessment tasks directed to them, the students who took the test through a table stating whether their answers were true or false at the end of the test directed towards them at the end of the test for that achievement and the score that reflects the average score of all students who took that test at that moment it is shown.



Numbers written according to a certain rule are given as **5, 10, a, 20, b, 30**. What is the result of the **b-a** process?

A ● 10

B ○ 15

C ○ 20

D ○ 25

★ Congratulations… You know

Figure 5. *Assessment item page view*

4.1.3. *Findings during the development (Using the prototype) stage*

The web-based dynamic assessment system, whose prototype was developed, acquisition-based tests were offered to 47 5th grade students in a public school in İzmir province under the Ministry of National Education as a pilot application. Students were presented with five assessment tasks for acquisition. Within the scope of the research, in order to increase the learning efficiency of the students in accordance with the Rapid Development Model, the achievement developments in the acquisitions determined were examined. The instructional guidance information of students regarding pilot implementation data is presented in Figure 6.

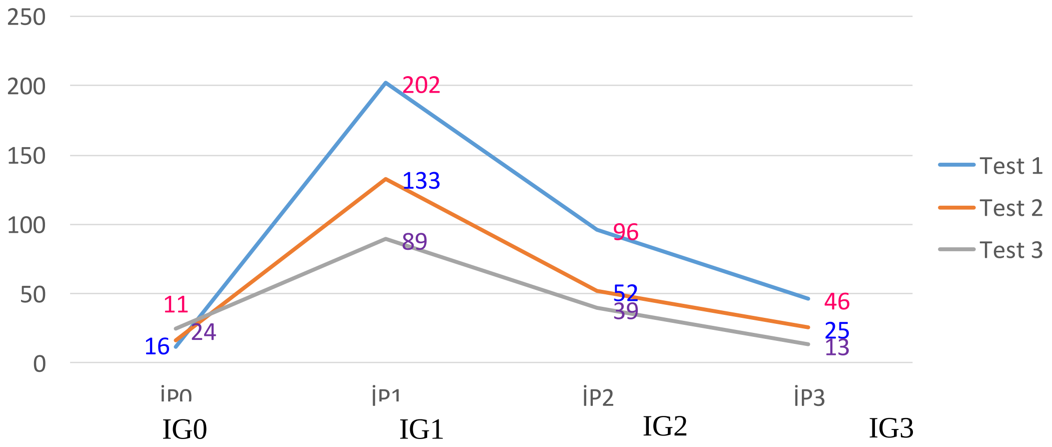


Figure 6. *Graph showing the instructional guidance information of students regarding pilot application data*

As can be seen in Figure 6, students received 202 instructional guidance 1, 96 instructional guidance 2, 46 instructional guidance 3 within the scope of Test 1, whereas 11 assessment items were answered without receiving instructional guidance. Within the scope of Test 2, 133 instructional guidance 1, 52 instructional guidance 2, 25 instructional guidance 3 received, whereas 16 assessment items were answered without receiving instructional guidance. Within the scope of Test 3, 89 instructional guidance 1, 39 instructional guidance 2, 13 instructional guidance 3 received, whereas 24 assessment items were answered without instructional guidance. In each test, it was determined that the students needed more instructional guidance in the first items of the test (they needed more for instructional help), and these needs (instructional help needs) gradually decreased in the following items of the test. Achievement points taken on the basis of gain in pilot implementation are presented in Table 4.

Table 4. *Pilot Application Achievement Based Success Points Chart*

Acquisition Name / Test No	Number Digits and Values	Reading and Writing Numbers	Number and Shape Patterns	Addition and Subtraction Estimation	Addition and Subtraction Process
Test1	84.32	90.73	80.22	74.00	79.33
Test2	91.87	92.12	85.60	86.00	86.50
Test3	94.40	93.09	89.60	92.00	92.00

It was found that the students' success scores constantly increase in each acquisition, the instructional guidance they receive in each test decreases constantly and they can perform the next assessment tasks without help. Based on the data here, students can see their success levels for the tests they have taken through the dynamic assessment system, they can instantly examine their assessment tasks and performances, they can arrange their learning instantly by taking the instructional guidances entered into the system, and thus, as Vygotsky expresses that the students can do without help. It can be stated that it is possible to fill the gap between what they can do with help, with dynamic assessments and support Vygotsky's sociocultural learning theory.

After the pilot implementation of the web-based dynamic assessment system, a three-dimensional focus group interview was made with the students' perception of satisfaction and benefit, motivation and engagement, and interface design. During the focus group interview, a semi-structured interview form developed by the researcher was used. The interviews were analyzed by content analysis and the system was evaluated. Themes and frequencies obtained as a result of content analysis are given in Table 5.

Table 5. *Content Analysis Themes and Frequencies*

Theme	Frequency (f)
Feeding the monkey	18
Competition score	16
Contribution of learning to mathematics	14
Ease of learning the system	12
The system is guiding	11
Visual elements	11
Legibility of the articles	11
Access to acquisitions and tests	10
Using hints	7

As can be seen in Table 5, as a result of the content analysis conducted as a result of focus group interviews, 9 themes were obtained. As a result of focus group meetings;

"How did the system guide you in unfamiliar questions?" Students' answers to the question posed as;

K1: *“It leads us to the correct solution of the questions and gives clues. I think we would not be able to solve the next questions more easily without clues.”*

K6: *"When we get the question wrong, we can see that we did it wrong. We can continue to answer the question correctly."*

While solving the questions in the questions that the students did not know, it was found that they were directed to the correct solution of the questions thanks to the instructional instructions given by the system, they found the correct answers more easily, and they used these instructional guidances in similar questions.

"Was the visual robot cute?" Students' answers to the question posed as;

K4: *"Yes, I think it is cute and it is fine."*

K6: *“I think it was good, the robot was swinging left and right, moving where it was.”*

It has been found that the visual elements (monkey, robot) in the system are interesting, cute and very beautiful, they help students, can be bigger, entertain themselves when they are tired or bored.

"Whether to feed the monkey or not?" Students' answers to the question posed as;

K1: *"I think it is fun."*

K7: *"I think the monkey is beautiful, looking at it feels fun."*

It was found that there was a monkey feeding status associated with student responses given to the assessment items in the system, that they motivated them positively and that they tried to feed the monkey by answering the assessment items more correctly.

“In your opinion, having a competition score; Whether or not? How did it affect you?” Students' answers to the question posed as;

K4: *“I think so. We can understand how we can improve ourselves better.”*

K7: *“I think so too. But it may be that he does greed, not to determine our level.”*

It was found that they should have a competition score in the system, they can understand how they can improve themselves better, they can follow their levels, they will solve less tests in the absence of a success score, and that they motivate them positively, they try to get more points.

"Would the learning here contribute to the Mathematics course?" Students' answers to the question posed as;

K3: *"Yes, it will help."*

K8: *“Yes it does. Because he explains it as our teacher explains and looks like the questions he solves.”*

It was found that learning in the system could contribute to mathematics lesson.

K2 *“The robot is big, the sample picture in the tips is small. It would be better if the robot got a little smaller and the picture was enlarged.”* made a statement.

K13 *“When we solve the question, pressing the answer button repeatedly, the situation of giving points constantly needs to be corrected.”* made a statement.

When marking the answer of the question in the system, a software error has been detected stating that it can exceed 100 points by clicking the answer button without waiting for the other problem to appear.

K7 *"The bravo statements, whose questions come to consciousness, were much more effective."* and

K9 *"The feedbacks were super, as I know, expressions like bravo and well done motivate me even more, I like it very much."* made a statement.

"Bravo, Perfect, Congratulations ... You know, etc." it was found that motivational expressions given in the system are much more effective.

K12 *"At the end of the test, you got the following score and you passed so many people, it is good for us to see our own development, but it is not realistic to always give it as "you have passed 25 people who have taken this test", it should be corrected."* declared.

At the end of the acquisition-based test, it was determined that giving the following score and giving the motivational message that you have passed so many people is beneficial in terms of seeing student developments.

K3 *"I like everything. The system was beautiful, you can easily move forward. Get in other lessons. It is nice to be in the computer environment, easier."* and

K6 *"... I was able to learn easily, it did not have a very difficult interface."* made a statement.

It has been found that when students use the system, they do not have to get help from someone else, but they have questions to take action, they may be good if they have a paper pencil, they can learn the system easily, they do not have a very difficult interface, they are not a difficult system, they can easily access, they can find the acquisitions related to the lesson and assessment they can easily reach their duties. It has been found that there is no difficulty in reading the articles in the system, but where the instruction is given, the articles are very intertwined and the numbers in the tables with sample solutions are not read.

According to the findings obtained from interviews with students using the system, it is possible to conclude that the web-based dynamic assessment system is useful and that information about their interactions in such systems is presented to them with graphic and visual elements positively affecting their learning processes.

4.1.4. *Findings at the Stage of Design Changes Based on Assessment*

Following the pilot implementation, studies on the development of the web-based dynamic assessment system continued in line with the feedback received from the students. In this context, as a result of focus group discussions with students;

- The problem that the system constantly scores (assessment of a software bug in the system) has been fixed as a result of clicking on the answer of an assessment item repeatedly before another assessment item is displayed in the assessment tasks that students take for the acquisition.
- It was stated that the visual robot could be bigger and the subject contents could also be included. In this regard, the robot has been developed a little more. Instructional guidance on the subject content of the achievements was decided to be given in more detail and on the example in the context of the subject scope and the system was developed in this context.
- The students stated that the sample pictures in the hints presented in the context of instructional guidance in the assessment item are small. In this regard, improvements have been made in the system.
- Expressions presented as motivation element have been further developed.
- Some students stated that they did not have difficulty in reading the articles on the pages, but the interfaces where instructional guidances were given were very intertwined and the numbers in the tables were not read. Improvements were made in these interfaces and made more readable.

In this context, the necessary improvements in the web-based dynamic assessment system were made with expert guidance. The final product for the system is configured on a server by purchasing a domain name. System's backups were taken periodically.

4.2. Sub-Problem 2: Findings Related to the Second Research Problem (Improvement Stage)

In this section, the opinions of the students using the web-based dynamic assessment system are given. In the 2018-2019 academic year, focus group interviews were conducted with 12 students using this system in two public secondary schools in Ankara. During the interviews, a semi-structured interview form developed by the researcher was used. Interviews were analyzed with content analysis and the system was evaluated. Themes and frequencies resulting from content analysis are given in Table 6.

Table 6. *Content Analysis Themes and Frequencies*

Theme	Frequency (f)
Instructional guidances	12
The effectiveness of the system	12
Visual elements	12
Competition score	12
Ease of learning the system	12
Access to acquisitions and tests	12
Monkey feeding anxiety	11
Individual and group graphics	10
Legibility of the articles	10
System sending notification	9

As a result of focus group discussions with students;

“How did the hints presented in the system guide you in questions you do not know?” The question posed as,

K9: *“I really liked how it helped with questions we couldn't do. It shows our score. I really liked such a test.*”

K12: *“Once I got the clues from the questions, I was able to do other questions in similar questions.”*

It has been found that the students are directed to the correct solution by solving the questions in questions they do not know, they may not be able to solve the following questions more easily without instructional guidance, and instructional guidance for the solution is very helpful.

"When you evaluate your acquisition-based learning, what are your thoughts on whether the system is effective in your success development?" The question posed as, K4: *“I really liked how he helped with questions we couldn't do. It shows our score. I liked such a test very much. My success score has increased continuously.”*

K7: *“I like everything. The system was nice, I enter from home. Get in other lessons. It is better to have a computer environment, we do not have difficulty. It helps us be more successful in lessons.”*

It has been found that students can observe their progress in learning acquisition-based mathematics, the system is very successful and they want to use this system in other courses.

“Were the visuals (monkey, hint robot, banana) in the system interesting and cute?” The question posed as,

K5: *“While it is difficult on the one hand, something is burning like a game, like a monkey. The robot helps us understand where we made mistakes. These were the things I liked the most. I think this system is very nice instead of paper and pencil.”*

K3: *"I like monkeys. It was nice to say how many people I passed."*

It has been determined that the visual elements of monkey, robot and banana in the system are interesting, fun and beautiful, and can entertain students when they are tired or bored.

"What are your thoughts about having a competition score in the system? Whether or not points?" The question posed as,

K1: *"I can see how many people I have passed as points. There must be points."*

K8: *"I think it is better to have a competition score, we try to pass, we feel an ambition."*

It was found that there should be a competition score in the system, students can follow their development level in their learning lives, that there is a competition score motivates their students positively and they make ambition.

"What are your thoughts on the individual and group graphics in the system?" The question posed as,

K1: *"Thanks to the graphics, I can see my past. I can also see our situation within the group. It was very understandable and beautiful."*

K8: *"With the graphics found at the end of the test, I can see both my own situation and the averages of others who took this test, I can see in which questions I got hints, how many hints I got, I can assess myself. It was pretty good."*

It was found that individual and group graphics in the system can see student developments by comparing them with group developments, and that the instructional guiding usage status charts given at the end of each test are understandable and useful and that they can follow their own levels.

"Did feeding the monkey cause anxiety?" The question posed as,

K4: *"Feeding the monkey did not cause any anxiety, I have to answer the questions more carefully and feed the monkey. It was more fun."*

K6: *"I don't have much trouble about this. Feeding the monkey is not a problem, it's fun."*

It was determined that feeding monkeys did not cause any anxiety in students and this situation led students to solve questions more carefully.

"How helpful was the system in notifying you?" The question posed as,

K2: *"It is very helpful for the system to provide feedback when we do it wrong."*

K9: *"The feedback has been super, as I know, phrases like bravo and well done motivate me even more, I like it very much."*

It has been found that it is motivating that the system gives feedback to the students while answering the questions and the notifications are useful.

"Did you have in difficulty reading the articles?" The question posed as,

K3: *"Normally I did not have any difficulties."*

K5: *"Normally legible, questions, tips, graphics can be read properly."*

It was found that students did not have any difficulties while reading the writings in the system and that the articles were legible.

According to the findings obtained from the interviews with students using the system, it is possible to conclude that the web-based dynamic assessment system is effective and useful, and that information about student interactions in the developed system with graphic and visual elements positively affects the learning processes.

5. DISCUSSION and CONCLUSION

Bloom (1984) stated that one-to-one tutorial support increases students' achievements by two standard deviations. In this context, the field where many technology supported environment designs are developed in order to support students from past to present and to ensure their development is seen in the literature (Ashton, et al. 2006; Costa, Kothe, Mullan & Butow, 2010; Critchley, Ware, Kumta & Wong, 2009; Marinagi, 2011; Wang, 2007, 2010, 2011,2014; Wang, Wang, Wang & Huang, 2006; Zou & Zhang, 2013). At the same time, this point of view is closely related to Vygotsky's sociocultural learning theory and "The Zone of Proximal Development". The convergent development area is expressed as the gap between what the student can do without help and what he can do with help (Bodrova & Leong, 1996; Vygotsky, 1978). It can be said that one of the alternatives to fill this gap is dynamic assessment. Because one of the main objectives of dynamic assessment is to improve the performance of learners by providing instructional aids with assessment tasks. In this context, the feature that distinguishes dynamic assessment from static assessment can be explained by feedback concept. In dynamic assessment, feedbacks are based on performance in an assessment task; rather than cognitive, affective and motivational information about students' performance, it is done with scaffolding to guide students to the correct answer. While dynamic assessments have been used extensively in foreign language teaching in classroom practices until today, they have started to be transferred to other fields thanks to the developing teaching technologies. One of the benefits of instructional technologies is that with the help of web-based systems, learners can monitor their development by testing themselves repeatedly, they can see their strengths and deficiencies with meaningful guides, and they can improve their learning performance by improving their learning strategies.

As Bloom (1984) stated that, the way to increase students' achievements with one-to-one tutorial support can be the subject of education technologies today. When the studies in this direction are examined in the literature, there are differences in the designs of these studies. In some studies, there are dynamic assessment elements (such as the option I want to ask my friend, leader board, etc.) while others (for example; assessment task and path. It has been observed that there are differences in algorithms based on scaffolding relationship) are differences in the algorithms used. In this research, it was decided to differentiate the algorithms in assessment task-guiding relations by retaining gamification elements. Accordingly, the learner-instructional guidance interaction algorithm has been developed (Figure 1). Instructional support is a concept related to the help given in case of a problem. As the working principle of the designed system; it is possible to give students (without moving to a new assessment task) consecutively the possible guiding (scaffolding) of the current assessment task in the context of a certain strategy. The design in this research has been associated with a special type of feedback, “continue to answer the assessment item until you find the correct answer” approach, and the design of the existing research is based on this.

In this research, a web-based environment that combines the principles of dynamic assessment and gamification with a dynamic assessment algorithm developed within the scope of the research has been developed. In the research, firstly, the requirement analysis for the web-based dynamic assessment system was done by examining the studies conducted in this context in the literature. Within the scope of the research, systematic descriptive scanning was carried out by selecting 98 articles suitable for the context, purpose and problem sentence among the resources accessed in the literature. It has been determined that 32 (32.65%) of the studies used in the articles evaluated were worked with dynamic assessment. When the sample level distribution was examined in the articles in which dynamic assessment was preferred, it was seen that the highest level of study was at the level of undergraduate (13 articles), at least high school (1 article) and graduate (1 article). In 25 (78.13%) of these articles, it was found that dynamic assessment further increased success and was effective. It was found that in 19 (76%) of the articles in which the success variable was examined, interactionist dynamic assessment was performed in 6 (24%) and interventionist dynamic assessment. When the studies conducted with dynamic assessment in our country are examined, the findings are quite limited (only 4 studies are available; 1 study is scale development, 2 studies are literature review and 1 study is a qualitative descriptive analysis study). In this research, no studies conducted in Turkey based on web-based dynamic assessment were encountered. The necessity of an e-assessment system, which has the feature of supporting the out-of-school learning experiences in addition to the in-school learning experiences of the students at the target audience level (secondary school level), which includes different instructional guidances according to different types of information, has been revealed. In this context, it has been found that the design, development and improvement of the web-based dynamic assessment system is needed.

As a result of the literature review, it has been determined that similar systems, albeit few, are used in a limited number of instructional guidances. In this research, students were given the right to repeat the same item (by offering gradual instructional guidance each time). In this way, a system design has been made according to the characteristics of the target audience, with a choice of four options, with only one option left, in other words, three times according to the wrong answer situation. One of the aspects of this research that is different from the studies in the literature is this feature. In the e-assessment system, which was designed by using similar dynamic assessment systems, points system and gamification elements were included in order to increase learning motivation. The unique aspect of this system on a national basis is that, in addition to the limited number of studies abroad, such a system has not yet been configured in our country. The unique aspect of the design in this research, which is international in scale, is that it has a unique quality compared to its peers, with the ability to offer students repeated instruction and instructional guidance in each retry. Another point to be emphasized here is inspired by the trial approach until we find the correct answer, which is a type of feedback. In this context, it was decided to support this research with algorithms based on interventionist dynamic assessment process. Accordingly, the learner-instructional guidance interaction algorithm has been developed (Figure 1).

The web-based dynamic assessment system adopted in this research was designed as a cake-form interventionist dynamic assessment. As a result of the literature reviews, a design was realized in line with the design principles determined by making use of the findings obtained from the dynamic assessment studies. The first prototype of the system was created with reference to the rapid prototyping model. System design has been developed in accordance with the opinions of the field experts regarding the suitability of the design to the dynamic assessment approach, the suitability of the feedbacks developed, the suitability of the interface and visual elements to the target audience, and the suitability of the number of tests and questions. In this context, the appropriateness of the design was provided by examining in the context of content validity. During the development of the system, a pilot application was made to 47 5th grade students in İzmir. Focus group meetings were held with 18 5th grade students determined after the pilot implementation. As a result of the pilot implementation and focus group discussions with students, efforts to improve the system continued. During the improvement phase, a web-based dynamic assessment system was opened for 12 5th grade students in Ankara. After the use of the system, semi-structured focus group interviews were held with the students. In the focus group interviews, the students are directed to the right solution while solving the questions that they do not know, the instructional guidance for the solution is very helpful, the students can observe the development of their achievements, the system is very successful, the other elements that they want to use this system in other courses, the visual elements (monkey, robot and banana) is interesting, fun and beautiful, it can entertain students when they are tired or bored, it is motivating that the system gives scaffolding to the students while solving the questions, and the notifications are useful, the students can learn the system easily and the system is guiding. Thus, as stated by Vygotsky, it is possible to fill the gap between what students can do without help and what they can do with help, and that supports Vygotsky's sociocultural learning theory.

Through the web-based dynamic assessment system developed in this research, students at the secondary school level can instantly see their achievements for the tests they have taken, instantly review the assessment questions and their answers, and can take pre-entered guides into the system and thus organize their learning instantly. In addition, it was tried to contribute to the students' educational goals by determining whether the student had an effective learning experience, questions that he / she answered incorrectly as a result of carelessness, and whether the course contents were deficient. In addition, with this study, it can be contributed to its extensive use in education and training in the context of the development of assessment environments in the context of advanced applications of computer-aided web-based education and training through internet technologies.

As part of the limitations of the research, there have been situations where the use of voice recorders for the recording of voice recordings by the parents of students in focus group interviews with students has occurred. In these cases, interview notes were kept by the researcher in order to determine the results of the research correctly and to realize their analysis realistically. For the effectiveness of the developed system, interaction situations, achievement developments, instructional orientation status of the students were studied within the scope of the acquisitions for 5th grade mathematics course. The interaction situations of students with regard to the effectiveness of the developed system, their success development, instructional guidance status and the probability of transition between situations can be revealed with an experimental study. The instructional guidance offered in this system is not limited to 5th grade mathematics lessons, but it is more holistic and covers videos, animations, lectures, etc. content designs can also be developed and presented to students. In addition, students' interactions in the system can be presented to students through learning panels using “Lag sequential analysis”, “markov chains”, etc. methods. In this context, it is assessed that students' sequential navigation will provide an important finding in estimating past, present and future patterns.

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).

ORCID

Arif TULUK https://orcid.org/0000-0003-3130-6005
Halil YURDUGÜL https://orcid.org/0000-0001-7856-4664

6. REFERENCES

Allal, L., & Ducrey, G. P. (2000). Assessment of- or in- the zone of proximal development. *Learning and Instruction*, 10, 137–152.

Ashton, H.S., Beevers, C.E., Korabinski, A.A., & Youngson, M.A. (2006). Incorporating partial credit in computer-aided assessment of Mathematics in secondary education. *British Journal of Educational Technology, 37*(1), 93-119. https://doi.org/10.1111/j.1467-8535.2005.00512.x

Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Bodrova, E., & Leong, D. J. (1996). *Tools of the mind: The Vygotskian approach to early childhood education*. Englewood Cliffs, NJ: Merrill/Prentice Hall.

Boud, D. (2000). Sustainable assessment: rethinking assessment for the learning society. Studies in continuing education, 22(2), 151-167. Erişim adresi: http://dx.doi.org/10.1080/713695728

Bransford, J. C., Delclos, J. R., Vye, N. J., Burns, M., & Hasselbring, T. S. (1987). *State of the art and future directions. In C. S. Lidz (Ed.), Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 479–496). New York: Guilford Press.

Campione, J.C., & Brown, A.L. (1985). *Dynamic assessment: One approach and some initial data.* Technical report no. 361, Univ. of Illinois at Urbana-Champaign, Champaign, IL. (ERIC ED269735).

Campione, J. C., & Brown, A. L. (1987). *Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), Dynamic assessment: An international approach to evaluating learning potential* (pp. 82–115). New York: The Guilford Press.

Critchley, L. A., Kumta, S. M., Ware, J., & Wong, J. W. (2009). Web-based formative assessment case studies: role in a final year medicine two-week anaesthesia course. *Anaesthesia and Intensive Care, 37*(4), 637-645.

Costa, D. S., Mullan, B. A., Kothe, E. J., & Butow, P. (2010). A web-based formative assessment tool for Masters Students: A pilot study. *Computers & Education, 54*(4), 1248-1253.

Deniz Kan, Ü. (2007). Okul öncesi eğitimde değerlendirme aracı olarak portfolyo. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 27*(1), 169-178.

Elliott, J. G. (2003). Dynamic assessment in educational settings: Realizing potential. *Educational Review, 55*, 15–32.

Haywood, H. C., Brown, A. L., & Wingenfeld, S. (1990). Dynamic approaches to psycho-educational assessment. *School Psychology Review, 19*, 411–422.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice.* Clinical & educational applications. New York: Cambridge University Press.

Kılıç, G. B. (2001). Oluşturmacı fen öğretimi. *Kuram ve Uygulamada Eğitim Bilimleri Dergisi,* 1, 7-22.

Marinagi, C. (2011). Web-based adaptive self-assessment in Higher Education. *Education in a technological world: communicating current and emerging research and technological efforts*, 978-84.

Mayer, R. E. (1992). *Thinking, problem solving, cognition.* New York, NY: W.H. Freeman and Company.

Moore-Brown, B., Huerta, M., Uranga-Hernandez, Y., & Peña, E. D. (2006). Using dynamic assessment to evaluate children with suspected learning disabilities. *Intervention in School and Clinic, 41*, 209–217.

Richey, R. C., Klein, J. D., & Nelson, W. A. (2004). Developmental research: Studies of instructional design and development. *Handbook of research for educational communications and technology, 2,* 1099-1130.

Sternberg, R. J., & Grigorenko, E. L. (2001). All testing is dynamic testing. *Issues in Education, 7*, 137–170.

Van den Akker, J. (1999). Principles and methods of development research. *In Design approaches and tools in education and training* (pp. 1-14). Springer Netherlands.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wang, K. H., Wang, T. H., Wang, W. L., & Huang, S. C. (2006). Learning styles and formative assessment strategy: enhancing student achievement in Web-based learning. *Journal of Computer Assisted Learning, 22*(3), 207-217.

Wang, T. H. (2007). What strategies are effective for formative assessment in an e-Learning environment? *Journal of Computer Assisted Learning, 23*, 171–186.

Wang, T. H. (2008). Web-based quiz-game-like formative assessment: Development and evaluation. *Computers & Education*, *51*, 1247–1263.

Wang, T. H. (2010). Web-based dynamic assessment: Taking assessment as teaching and learning strategy for improving students' e-Learning effectiveness. *Computers & Education*, *54(4*), 1157–1166.

Wang, T. H. (2011). Implementation of web-based dynamic assessment in facilitating junior high school students to learn mathematics. *Computers & Education*, *56*, 1062-1071.

Wang, T. H. (2014). Developing an assessment-centered e-Learning system for improving student learning effectiveness. *Computers & Education*, *73*, 189-203.

Yıldırım, A. & Şimşek, H. (2011). *Sosyal bilimlerde nitel araştırma yöntemleri.* 8. Basım, Ankara: Seçkin Yayınları.

Zou, X. & Zhang, X. (2013). Effect of different score reports of Web-based formative test on students' self-regulated learning. *Computers & Education*, *66*, 54-63.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 657–673
https://doi.org/10.21449/ijate.714218
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Measurement Invariance Testing with Alignment Method: Many Groups Comparison

Gozde Sirganci [1], **Gizem Uyumaz** [2,*], **Alperen Yandi** [3]

[1]Yozgat Bozok University, Department of Educational Sciences, Yozgat, Turkey
[2]Giresun University, Department of Educational Sciences, Giresun, Turkey
[3]Bolu Abant İzzet Baysal University, Department of Educational Sciences, Bolu, Turkey

ARTICLE HISTORY

Received: Apr. 03, 2020
Revised: Sept. 30, 2020
Accepted: Oct. 15, 2020

KEYWORDS

Multi-group confirmatory factor analysis,
Validity,
Modification indexes,

Abstract: It is necessary to examine the measurement invariance (MI) among groups in studies where different groups are compared by using a measurement instrument. Most of the studies, measurement invariance is tested with multiple group confirmatory factor analysis. This model applies many model adjustments based on the modification indexes. Therefore, it is not practical due to too many large modification indexes while testing MI over many groups. Besides scalar model is a poor model fit when comparing many groups and so does not hold MI. In this study, the aim is to explain the basic concepts and processes of the alignment method which is offered as a new method for testing MI and illustrate an application on the real data set. In this study, measurement invariance among 56 countries including Turkey is tested with alignment method in order to set an example for researchers. For this purpose, the Instrumental Motivation Scale data, which is one of the psychological measurement instruments used in PISA 2015, was used. As a result of MG-CFA, it was found that configural invariance was ensured. The fit indexes of CFI and TLI were calculated as 0.982 and 0.946 respectively in this stage. After that, metric invariance was tested by considering the difference of fit indices obtained for the two stages. It was found that the metric invariance could not be provided. Alignment results show which countries hold MI and which do not. Besides it provides information which items have the most invariants for groups that hold MI.

1. INTRODUCTION

Validity is an important psychometric property that must be examined in every study that has been conducted with measuring instruments. Bias is one of the most important sources of systematical error that affect the validity (Messick,1995). Test bias, defined as a systematic error is the measurement, captures the idea that there are construct-irrelevant components that result in systematically higher or lower scores on the measurement for the groups under examination (American Educational Research Association [AERA] & National Council on Measurement in Education [NCME], 1999). In bias, the scores of individuals contain systematical error depending on their subgroup (Camilli & Shepard, 1994; Zumbo 1999). While

CONTACT: Gizem UYUMAZ ✉ gizemuyumaz@gmail.com Giresun University, College of Education, Department of Educational Sciences- Giresun, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

determining bias on item level, item response function of the item is analyzed. During the investigations made when determining bias, measurement invariance concept has been encountered. Measurement invariance is the statistical property of the correlation being the same between the observed variable (items) and latent variable (measured trait) among the subgroups (Drasgow & Kanfer, 1985; Widaman & Reise, 1997). In studies where different groups are compared by means of a measurement instrument, measurement invariance among the groups is needed to be investigated and the invariance must be proven. Measurement invariance is a validity issue; and if the measurement invariance cannot be obtained among the subgroups, we cannot make comparisons among groups.

For testing measurement invariance, there are methods based on two different theories. The first one is based on Item Response Theory (IRT), and the other is based on Structural Equation Modelling (SEM). One of the most used methods for examining measurement invariance under SEM is Multiple Group Confirmatory Factor Analysis (MG-CFA). MG-CFA is used in order to determine whether the factor structure of a scale is equal in multiple samples or in multiple subgroups (according to gender, socio-economic level, nation, religion, culture, etc.) (Jöreskog, Sörbom, Toit & Toit, 2001). Four hierarchical models are tested with MG-CFA: configural, metric (weak factorial), scalar (strong factorial), and strict invariance (Byrne, Shavelson & Muthen, 1989; Vandenberg & Lance, 2000). A hierarchical order exists among the models after the first model is confirmed the testing of the second model starts. While testing the models, the number of limited parameters is increased gradually. For configural invariance, no limitation is used in order to equalize any parameters between the groups. Metric invariance assumes that the factor loadings of the across groups are equal. In this way factor variances across groups and structural relations can be comparable. Scalar invariance assumes that both the factor loadings and the measurement intercept (thresholds with categorical items) are invariant among the groups, and only in this way, it becomes possible to compare factor means and variances among the groups. The strict invariance holds the value of the residual variances equal across groups (Muthén & Asparouhov, 2018). Equations related to models are presented in Equation 1-3.

Configural:
$$y_{ig} = v_g + \lambda_g f_{ig} + \varepsilon_{ig} \qquad (1)$$
$$E(f_g) = \alpha_g = 0,\ V(f_g) = \psi_g = 1$$

Metric:
$$y_{ig} = v_g + \lambda_g f_{ig} + \varepsilon_{ig} \qquad (2)$$
$$E(f_g) = \alpha_g = 0,\ V(f_g) = \psi_g$$

Scalar:
$$y_{ig} = v_g + \lambda_g f_{ig} + \varepsilon_{ig} \qquad (3)$$
$$E(f_g) = \alpha_g,\ V(f_g) = \psi_g$$

g: number of groups, i: number of independent observations in group g, f_g : latent variable, λ_g:factor loading, v_g: measurement intercept, α_g:factor mean ve ψ_g:factor variance

To provide model fit at any phase (model is usually rejected at strict invariance phase) may depend on large number of modifications. In circumstances where so many large modification indexes are presented MGFA fails. Because the presence of modification indexes enabling so many large valued changes shows that to achieve an acceptable model a long model modification line is needed. In this situation, the sources that ruin invariance cannot be defined suitably. Therefore, it is not guaranteed to achieve a suitable model at the end of the modifications (Asparouhov & Muthén, 2014). MGCFA is based on dual comparisons across groups. Since comparisons are made for each item when the number of groups is larger, the number of dual comparisons will increase exponentially; and this will increase the possibility of miscalculation of the measurement invariance, and will make the method unfavorable (Muthén & Asparouhov, 2018; Rutkowski & Svetina, 2014). On the other hand, for methods based on IRT, it is unlikely to talk about metric invariance. Because, while only the regression slope is constant, the regression intercept is not constant; it is hard to say that an item will be perceived the same by individuals. Especially in the analysis of many groups, scalar invariance rarely fits the data set (Muthén & Asparouhov, 2018). Because of the reasons mentioned here, while testing scalar invariance, comparison of factor means among groups is nearly impossible, either for SEM or IRT based conventional methods.

MG-CFA is impractical in comparing too many groups. In contrast, the alignment method automates and greatly simplifies the measurement invariance analysis. In addition, the alignment method can be used for determining invariance of parameters singly and which item provides the invariance mostly in measurement instrument. This situation is important for determining the best-fit CFA model that provides partial measurement invariance when estimating factor mean and variance of the groups. Also, the alignment method determines which group contributes to the measurement invariance by a single analysis. For this reason, in studies where measurement invariance is examined, when confronted with a large number of modifications, and especially when the number of groups is large; a new method is needed in order to investigate whether there is invariance across the groups or not. In this study it is aimed to explain basic terminologies and processes of the Alignment method suggested by Asparouhov and Muthén (2014) and explain an application example.

The Alignment method does not presume exact measurement invariance. It can estimate factor mean and variance parameters in each group while discovering the most suitable measurement invariance pattern. The strong aspect of the alignment method is that it is based on configural model and can predict the most suitable models for a large number of groups. In the configural model, since measurement intercepts and loadings are free across groups, factor means and variances cannot be defined. However, the model sets the metric of factor by fixing the factor mean to zero and the factor variance to 1. In the configural model, since the factor mean and variance are not defined latent characteristic (factor) cannot be compared across groups; that means it scales differently for each group. It is not possible to compare factorial scores of individuals situated in different groups and intergroup factor mean. The Alignment method can predict factor mean and variance for each group without assuming measurement invariance and by discovering the most suitable measurement invariance pattern. With this aspect, the method gives information about the level of measurement invariance along with intergroup factor mean and variance by calculating approximate measurement invariance. Thereby, which measurement parameters are approximately constant, and which are not specified (Asparouhov & Muthén, 2014; Kim, Cao, Wang & Nguyen, 2017). In other words, the alignment method can estimate factor loadings (λ_g), measurement intercepts (v_g), factor means (α_g) and variances (ψ_g) by predicting the number of variable item parameters and the model that can hold impaired measurement variance at the minimum level (Muthén & Asparouhov, 2018).

One advantage of the alignment method is that it has the same model fit with the configural model. The method minimizes the distortions of measurement invariance by predicting group-specific factor mean (α_g) and variance (ψ_g). Although these parameters cannot be defined without applying strong invariance, this is possible by using a series of constraints that optimize the simplicity function in the alignment method. The simplicity function "F" is optimized with few parameters that are not substantially invariant, and many parameters that are not nearly invariable, rather than many parameters that are not moderately invariant. This alignment method includes a simplification function similar to the rotation (Jennrich, 2006) used in Exploratory Factor Analysis (EFA) (Muthén & Asparouhov, 2014).

The measurement invariance is tested in two steps by the alignment method. In the first step, the configural model in which factor loading and measurement intercepts are free across groups and the factor means are fixed to 0 and the variances to 1 in all groups are estimated. This configural model, which is called the base model "M$_0$", is the best fit model among the multi-group factor analysis models since it does not contain parameter constraints across groups. In the second step, alignment optimization is done. At this stage, factor means and variances set free, and factor means and variances are calculated by a simplicity function that minimizes the distortions of measurement invariance. This simplicity function consists of the loss function (f) for each group pair, where each measurement intercepts and factor loadings values are components (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2018).

The model, which is not defined by the simplicity function, is defined by adding factor means and variances to the configural model. The simplicity function is presented in Equation 4.

$$F = \sum_p \sum_{g_1<g_2} w_{g_1,g_2} f(\lambda_{pg_1} - \lambda_{pg_2}) + \sum_p \sum_{g_1<g_2} w_{g_1,g_2} f(v_{pg_1} - v_{pg_2}) \qquad (4)$$

$$w_{g_1,g_2} = \sqrt{N_{g_1} N_{g_2}} \qquad (5)$$

w: factor weight, N: sample size of the group

The proposed final aligned model has the same fit as the M$_0$ model. Although the aligned model tries to minimize the amount of invariance, it does not compromise the model fit. The relationship between the M$_0$ model and the last aligned model is in line with the relationship between the non-rotated model in exploratory factor analysis which has the best fit between a fixed number of factors and all EFAs and the rotated model which has the same fit with the non-rotated model without compromising the fit of the model (Muthén & Asparouhov, 2014).

There are two different alignment optimizations in the alignment method: FIXED and FREE. In FIXED alignment optimization, the factor means of the first group is restricted to 0 and its variance to 1. In FREE optimization, there is no restriction on the factor mean and variance of the first group, and these parameters are considered as additional parameters that should be estimated (Kim, Cao, Wang & Nguyen, 2017).

2. METHOD

This study is a descriptive research which aims to assess the measurement invariance via alignment method of the Instrumental Motivation Scale (INSTSCIE) which is in science learning applied in PISA (The Programme for International Student Assessment) 2015 and to introduce of "alignment method" in this assessment. In the literature, there are some studies about exact MI which is done with PISA 2015 data. To debated this new MI method's results with the studies in the literature, PISA 2015 data is used in this study.

2.1. Study Group

PISA 2015 has been implemented in 72 countries, of which 35 are OECD members. In this study, which aims to explain the basic concepts and processes of the alignment method for measurement invariance and to introduce an example using the alignment method for researchers, the data of 406,961 participants from 57 countries which answered the INSTSCIE were used. Information about the countries in the study and the number of participants in the countries are shown in Table 1.

Table 1. *Frequencies and percentages related to the study group*

	Code	Country	f	%		Code	Country	f	%		Code	Country	f	%
1	Tur	Turkey	5608	.38	**0**	Grc	Georgia	5239	1.29	**9**	Prt	Portugal	6982	1.72
2	Aus	Australia	13377	.29	**1**	Hkg	Hong Kong	5158	1.27	**0**	Qat	Qatar	10682	2.62
3	Aut	Austria	6708	.65	**2**	Hun	Hungary	5341	1.31	**1**	Rus	Russia	5477	1.35
4	Bel	Belgium	8754	.15	**3**	Isl	Iceland	3150	0.77	**2**	Sgp	Singapore	5971	1.47
5	Bra	Brazil	18276	.49	**4**	Irl	Ireland	5473	1.34	**3**	Svk	Slovakia	5759	1.42
6	Bgr	Bulgaria	5143	.26	**5**	Ita	Italy	10815	2.66	**4**	Svn	Slovenia	5913	1.45
7	Can	Canada	18706	.60	**6**	Jpn	Japan	6404	1.57	**5**	Esp	Spain	6474	1.59
8	Chl	Chile	6731	.65	**7**	Kor	South Korea	5443	1.34	**6**	Swe	Sweden	5071	1.25
9	Tap	Taipei	7576	.86	**8**	Lva	Latvia	4678	1.15	**7**	Che	Switzerland	5545	1.36
10	Col	Colombia	11019	.71	**9**	Ltu	Lithuania	6047	1.49	**8**	Tha	Thailand	7856	1.93
11	Cri	Costa Rica	5686	.40	**0**	Lux	Luxembourg	4925	1.21	**9**	Are	United Arab Emirates	12940	3.18
12	Hrv	Croatia	5447	.34	**1**	Mac	Macau (China)	4414	1.08	**0**	Tun	Tunisia	4532	1.11
13	Cze	Czech Republic	6397	.57	**2**	Mex	Mexican	7209	1.77	**1**	Gbr	United kingdom	13082	3.21
14	Dnk	Denmark	6440	.58	**3**	Mne	Serbia	4945	1.22	**2**	Usa	USA	5414	1.33
15	Dom	Dominican Republic	3992	.98	**4**	Nld	Netherlands	5078	1.25	**3**	Ury	Uruguay	5412	1.33
16	Est	Estonia	5312	.31	**5**	Nzl	New Zealand	4239	1.04	**4**	Qch	B-S-J-G (China)	9564	2.35
17	Fin	Finland	5621	.38	**6**	Nor	Norway	5093	1.25	**5**	Qes	Spain B	31003	7.62
18	Fra	France	5312	.31	**7**	Per	Peru	6535	1.61	**6**	Quc	Massachusette	1534	0.38
19	Deu	Germany	5353	.32	**8**	Pol	Poland	4336	1.07	**7**	Que	North Carolina	1770	0.43
												Toplam	406961	100

The country with the highest number of participants is the Spain regions with 31.003 participants (7.62%). The statewith the lowest number of participants is Massachusette with 1,534 participants (0.38%). Turkey has participated the PISA 2015 with 5,608 students (1.38%).

2.2. Data Collection Tool

In this study, data of ST113 (INSTSCIE-Instrumental Motivation to Learn Science), one of the psychological measurement tools used in the PISA 2015, was used. According to OECD (2016):

> *"PISA 2015 focused on science learning in school by including several questions about the learning environment in the science classroom. They asked how often specific activities happened in the school science course."*

With ST113, it is aimed to measure the perspective of in-school scientific issues of the students.

The Instrumental Motivation to Learn Science is one of the subscale of the measurement tool related to the Disciplinary Climate in Science Classes. Items are presented in Table 2.

Table 2. *Items of Instrumental Motivation to Learn Science*

Code	Item
ST113Q01TA	Making an effort in my <school science> subject(s) is worth it because this will help me in the work I want to do lat
ST113Q02TA	What I learn in my <school science> subject(s) is important for me because I need this for what I want to do later on
ST113Q03TA	Studying my <school science> subject(s) is worthwhile for me because what I learn will improve my career prospects.
ST113Q04TA	Many things I learn in my <school science> subject(s) will help me to get a job.

2.3. Data Analysis

In the data analysis process, missing values and extreme values were examined and whether or not the assumptions of MG-CFA which were normality, linearity, homoscedasticity and multicollinearity were provided for every single country. The amount of missing values in all data set was below 5% thus it was deleted.

In this study, to set an example for practitioners, the approximate measurement invariance across the countries was determined by the alignment method. The analyzes of the Alignment method for approximate measurement invariance and MG-CFA that tested the exact measurement invariance were performed in the Mplus 7.1 program. The Mplus script of the alignment analysis is attached in the appendix.

In the analysis of measurement invariance with MG-CFA, model fit was examined by taking into consideration the Comparative Fit Index (CFI) and the Standardized Root Mean Square Residual (SRMR). Sokolov (2019) stated that RMSEA and TLI values give erroneous results in determining the measurement invariance with MG-CFA especially when testing metric invariance and CFI and SRMR values should be taken into consideration in such studies. Cut-off criteria for the fit indexes recommended by Sokolov (2019) for use in many group comparisons are presented in Table 3.

Table 3. *Cut-off criteria of Goodness of Fit Indexes for Multiple Group Measurement Invariance Comparisons*

	Cut-off Values			Relative Cut-off Values	
Fit Index	Configural inv.	Metric inv.	Scalar inv.	Configural inv.	Metric inv.
CFI	>0.985	>0.980	>0.970	>-0.010	>-0.010
TLI	-	-		-	>-0.005
RMSEA	-	-		-	<0.005
SRMR	<0.020	<0.040	<0.045	<0.010	<0.010

With the alignment procedure in the Mplus 7.1, the measurement invariance is tested with an algorithm based on the calculation of the largest number of groups where the difference between the parameters is not significant by making binary parameter comparisons. The table comparing factor means and variances for all groups shows on the top in the Mplus output file. The countries/groups which measurement invariance is not provided significantly are shown in bold in brackets. Another output is the table in which the factor means of all countries are ordered from high to low and the significant differences across them are shown through the z test. In addition, the contribution of each item's interceps and factor loading to the optimized simplicity function is calculated by the measure of R^2. The R^2 is a useful descriptive statistic that gives the degree of noninvariance that can be absorbed by group-varying factor means and variances (Muthén & Asparouhov, 2014; 2018). In the configural model, it shows how much of the parameter variation across groups for each measurement parameter can be explained by factor means and variation in factor variances. R^2 value close to "1" implies a high degree of invariance and to "0" a low degree of invariance.

3. FINDINGS

3.1. Results of Multi-Group Confirmatory Factor Analysis

The analyzes carried out with MG-CFA was stopped at the stage where measurement invariance were not provided. Therefore, the results given in Table 4 belong to the results of configural and metric invariance.

Table 4. *Configural and metric invariance results*

	Models	χ^2	df	p	CFI	SRMR
ST113	Configural	21055.716	114	0.000	0.982	0.018
	Metric	26378.988	282	0.000	0.978	0.046
	Configural vs metric	5323.272	168	0.000	-0.004	0.028

According to Table 4, the CFI is 0.003 points below the cut-off value proposed by Sokolov (2019). On the other hand, it is seen that SRMR is lower than the accepted cut-off value of 0.02. Considering these two values together, it can be said that the configural model is provided for 57 countries. After determining that configural model invariance was provided, analyzes were carried out for the metric invariance stage. The CFI and the SRMR, calculated at the metric invariance stage were outside the recommended cut-off value in Table 3 (CFI> 0.98, SRMS <0.04). When the relative cut-off values are examined, it is seen that the cut-off value of ΔSRMR is greater than 0.01 and the cut-off value of ΔCFI is less than -0.01. Considering the cut-off values of fit indexes and their relative cut-off values, it was concluded that the metric invariance was not provided. In addition, the chi-square test (Fan & Sivo, 2009), which is used as a complement to alternative fit indexes in testing measurement invariance, was also examined and it was found that the difference between chi-square fit index values was significant. Therefore, it was supported that the metric invariance was not provided with the chi-square test.

In the examinations, it has been determined that the measurement invariance is impaired in the weak invariance stage. However, for the 57 countries, it is not possible to determine across which countries the measurement invariance is impaired by MG-CFA. To reach this information, a double comparison of 57 groups (57x56/2=1.596) should be done, which will bring the work and time load to a high level. At this point, the alignment method was used to determine for which countries the measurement invariance was provided/impaired at the item level. With this method, the measurement invariance of the items on the scale has been revealed.

3.2. Results of the Alignment Method

In this section, it is shown how the alignment method solves the problem of comparing the factor means found by traditional multi-group factor analysis under scalar invariance. Maximum Likelihood estimation was used for measurement invariance analysis. In the alignment analysis, since the factor mean of the 29th country (Lithuania) is closest to zero, FIXED alignment optimization was used and this country was taken as the reference country whose factor mean was restricted to zero. Table 5 shows each item’s intercept and factor loading values for 57 countries.

Table 5. *Each item's Alignment results of 57 countries*

Item	Intercepts	Factor Loadings
ST113.1	1 2 3 **(4)** **(5)** 6 7 8 **(9)** 10 **(11)** 12 13 **(14)** 15 16 **(17)** **(18)** 19 20 **(21)** **(22)** **(23)** 24 **(25)** **(26)** **(27)** 28 **(29)** 30 31 **(32)** **(33)** 34 35 **(36)** 37 38 **(39)** 40 41 42 43 **(44)** **(45)** **(46)** 47 **(48)** 49 50 **(51)** **(52)** **(53)** 54 **(55)** 56 57	1 2 3 4 **(5)** 6 **(7)** 8 9 10 **(11)** 12 13 **(14)** 15 16 17 **(18)** 19 **(20)** **(21)** **(22)** 23 24 **(25)** **(26)** 27 28 **(29)** 30 **(31)** 32 **(33)** **(34)** **(35)** 36 **(37)** 38 **(39)** **(40)** **(41)(42)** 43 **(44)** **(45)** **(46)** 47 **(48)** **(49)** **(50)** **(51)** **(52)** **(53)** **(54)** **(55)** **(56)** 57
ST113.2	**(1)** **(2)** 3 4 5 6 **(7)** **(8)** **(9)** 10 **(11)** **(12)** **(13)** **(14)** **(15)** **(16)** **(17)** 18 19 20 21 22 23 **(24)** **(25)** **(26)** **(27)** **(28)** **(29)** 30 31 **(32)** **(33)** 34 **(35)** 36 **(37)** **(38)** 39 40 **(41)** 42 43 **(44)** 45 **(46)** 47 **(48)** 49 **(50)** **(51)** 52 53 54 55 56 57	**(1)** **(2)** 3 4 5 6 7 8 **(9)** 10 11 12 **(13)** **(14)** 15 **(16)** **(17)** 18 19 **(20)** **(21)** **(22)** **(23)** 24 **(25)** **(26)** **(27)** **(28)** **(29)** 30 31 32 **(33)** 34 35 **(36)** 37 38 **(39)** **(40)** **(41)** 42 43 **(44)** 45 46 47 **(48)** 49 50 **(51)** 52 53 **(54)** 55 56 57
ST113.3	**(1)** 2 3 4 **(5)** 6 7 8 9 10 **(11)** **(12)** **(13)** 14 15 16 **(17)** 18 19 20 21 **(22)** **(23)** 24 25 **(26)** 27 **(28)** **(29)** 30 **(31)** 32 **(33)** 34 35 36 37 **(38)** **(39)** **(40)** **(41)** 42 43 **(44)** 45 **(46)** 47 **(48)** **(49)** 50 51 **(52)** **(53)** **(54)** 55 56 **(57)**	1 2 3 **(4)** **(5)** **(6)** **(7)** **(8)** **(9)** 10 **(11)** 12 13 14 **(15)** 16 17 **(18)** 19 20 21 **(22)** 23 **(24)** **(25)** **(26)** 27 **(28)** **(29)** 30 **(31)** **(32)** **(33)** **(34)** 35 36 37 **(38)** **(39)** **(40)** **(41)** 42 **(43)** **(44)** **(45)** **(46)** 47 **(48)** 49 **(50)** **(51)** 52 **(53)** **(54)** **(55)** 56 57
ST113.4	**(1)** **(2)** 3 **(4)** 5 6 **(7)** 8 **(9)** **(10)** **(11)** 12 13 **(14)** 15 **(16)** **(17)** **(18)** 19 **(20)** **(21)** 22 **(23)** **(24)** 25 **(26)** **(27)** 28 **(29)** 30 31 **(32)** 33 34 **(35)** 36 **(37)** 38 **(39)** **(40)** 41 42 43 **(44)** **(45)** 46 **(47)** **(48)** **(49)** 50 **(51)** **(52)** 53 54 **(55)** 56 **(57)**	1 2 3 **(4)** 5 6 **(7)** 8 **(9)** **(10)** 11 **(12)** 13 14 15 **(16)** **(17)** **(18)** 19 20 21 22 23 24 **(25)** **(26)** 27 28 **(29)** 30 31 **(32)** **(33)** 34 35 **(36)** 37 38 39 **(40)** **(41)** 42 43 **(44)** **(45)** **(46)** **(47)** **(48)** **(49)** 50 51 **(52)** 53 54 **(55)** 56 57

Table 5 shows the findings regarding the invariance of the intercepts and factor loading values of the ST113 coded questionnaire for 57 countries. The results of the alignment analysis are interpreted as the fact that the intercepts and factor loading values differ significantly across the groups (countries) in parentheses. Thus, factor loadings and factor intercepts can be compared across countries which are within the parentheses. For example, the factor loading of the first item does not differ significantly for the countries coded as 1, 2, 3, 4, 6, 8, 9, 10, 12, 13, 15, 16, 17, 19, 23, 24, 27, 28, 30, 32, 36, 38, 43, 47, 57. Metric invariance has been provided for the first item across these countries. Therefore, factor variances and structural relationships can be compared across groups for this item. Also this item does not signiificantly differ from both the intercepts and loading among the countries coded as 1, 2, 3, 6, 8, 10, 12, 13, 15, 16, 19, 24, 28, 30, 38, 43, 47, 57. Accordingly, it can be said that scalar invariance that assuming both intercepts and factor loading parametres are equivalent across groups has been provided across these countries. Therefore, it is possible to compare factor means and intercepts across these countries.

Table 6 shows the coefficient of fit function and R^2 values of each item, which shows how much items contribute to the optimized simplicity function.

Table 6. *Alignment Fit Statistics*

Items	Intercepts		Factor Loadings	
	Fit Function Contribution	R^2	Fit Function Contribution	R^2
ST113.1	-536.658	0.964	-551.396	0.987
ST113.2	-535.435	0.970	-526.201	0.935
ST113.3	-530.307	0.972	-534.924	0.901
ST113.4	-557.961	0.960	-532.381	0.876

When Table 6 is examined, it can be said that all the items of the ST113 questionnaire contributed similarly to the simplicity function. This finding shows that the degree of noninvariance is similar. The R^2 results presented in Table 6 are interpreted in the configural model as able to explain the variation in the intercepts and factor loading values predicted for all groups with the variation in factor means and variances among all groups. The R^2 values of the item coded ST113.4 indicate that the item contributed the least to the simplicity function. In other words, this item has most degree of noninvariant across the groups. Table 7 shows the factor means estimated for all groups by the alignment method and groups that have factor means significantly different on the 0.05 level.

Table 7. *Comparison of Factor Means between Countries*

Ranking	Group	Factor Means	Groups with Significantly Smaller Factor Mean
1	19	.753	47 13 22 30 26 4 18 27 43 44 14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
2	3	.729	13 22 30 26 4 18 27 43 44 14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
3	34	.715	13 22 30 26 4 18 27 43 44 14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
4	47	.692	13 22 30 26 4 18 27 43 44 14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
5	13	.581	22 30 26 4 18 27 43 44 14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
6	22	.535	4 18 27 43 44 14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
7	30	.520	18 27 43 44 14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
8	26	.502	18 27 43 44 14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
9	4	.494	27 43 44 14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
10	18	.455	14 28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
11	27	.441	28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
12	43	.431	28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
13	44	.418	28 25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
14	14	.412	25 36 38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
15	28	.376	38 12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
16	25	.363	12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
17	36	.355	12 2 17 6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
18	38	.330	6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
19	12	.314	16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
20	2	.313	6 16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
21	17	.303	16 31 23 55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
22	6	.277	55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
23	16	.259	55 21 41 46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
24	31	.242	46 9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
25	23	.239	9 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
26	55	.230	9 45 56 20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
27	21	.213	20 53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50

Ranking	Group	Factor Means	Groups with Significantly Smaller Factor Mean
28	41	.206	53 39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
29	46	.200	39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
30	9	.192	39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
31	45	.191	39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
32	56	.169	8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
33	20	.168	39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
34	53	.158	39 52 8 57 10 33 24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
35	39	.117	24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
36	52	.116	24 1 35 51 11 29 7 5 37 42 32 54 40 48 49 15 50
37	8	.097	51 11 29 7 5 37 42 32 54 40 48 49 15 50
38	57	.094	11 29 7 5 37 42 32 54 40 48 49 15 50
39	10	.090	51 11 29 7 5 37 42 32 54 40 48 49 15 50
40	33	.079	11 29 7 5 37 42 32 54 40 48 49 15 50
41	24	.071	11 29 7 5 37 42 32 54 40 48 49 15 50
42	**1**	**.068**	**11 29 7 5 37 42 32 54 40 48 49 15 50**
43	35	.064	29 7 5 37 42 32 54 40 48 49 15 50
44	51	.052	29 7 5 37 42 32 54 40 48 49 15 50
45	11	.022	7 5 37 42 32 54 40 48 49 15 50
46	29	.000	7 5 37 42 32 54 40 48 49 15 50
47	7	-.032	37 42 32 54 40 48 49 15 50
48	5	-.050	37 42 32 54 40 48 49 15 50
49	37	-.086	48 49 15 50
50	42	-.091	48 49 15 50
51	32	-.096	49 15 50
52	54	-.100	49 15 50
53	40	-.107	15 50
54	48	-.119	15 50
55	49	-.127	50
56	15	-.161	
57	50	-.197	

For convenience of the presentation, the groups are ordered from high to low according to factor means and the groups that have factor means that differ on the 0.05 significance level are determined. For example, as seen in Table 7, the factor means of the 19th country estimated by the alignment method is 0.753 and this value of the 19th country is significantly higher than the countries whose codes written in the last column.

In exact MI framework, scalar invariance assumes that both the factor loadings and the measurement intercept are invariant among the groups if and only it is possible to compare factor means and variances among the groups. Especially, when there are many groups, scalar invariance rarely fits to the data set (Muthén and Asparouhov, 2018). In the Alignment method, the most appropriate measurement invariance pattern is discovered in which the factor means and factor variances of the groups are comparable. In this method, the maximum number of groups in which the factor means and factor variances across the groups do not differ statistically are estimated. In this way, it is revealed that among which groups that all items of the scale are comparable. Thus, MI are determined in not only item-based level but also scale-based level. Specifically for Turkey, is factor mean is 0.068, ranks 42nd out of 57 countries when ranked from high to low. When all statements taken together, while the factor means of the countries which coded 8 (Chile), 57 (North Carolina), 10 (Colombia), 33 (Serbia) and 24 (Ireland) coded countries' factor means are larger than Turkey, it is smaller for 35 (New Zealand) and 51 (United kingdom). However, the factor mean differences are not statistically significant between Turkey and these countries. This table shows that approximate measurement invariance is provided between Turkey and Chile, North Carolina, Colombia, Serbia, Ireland, New Zealand and the United kingdom. By means of the Alignment method, it is determined with a one-step analysis which countries factor means are comparable of each country included in the analysis.

Figure 1, Figure 2 and Figure 3 show that scatter diagrams between factor means obtained in the alignment method of countries and factor means obtained in the scalar invariance stage of MG-CFA. The scattering values for all countries are shown in Figure 1. In Figure 2 and Figure 3, the scattering values for the countries are presented on a larger scale by cutting-off the graph in Figure 1 at the level of (0.2, 0.1).

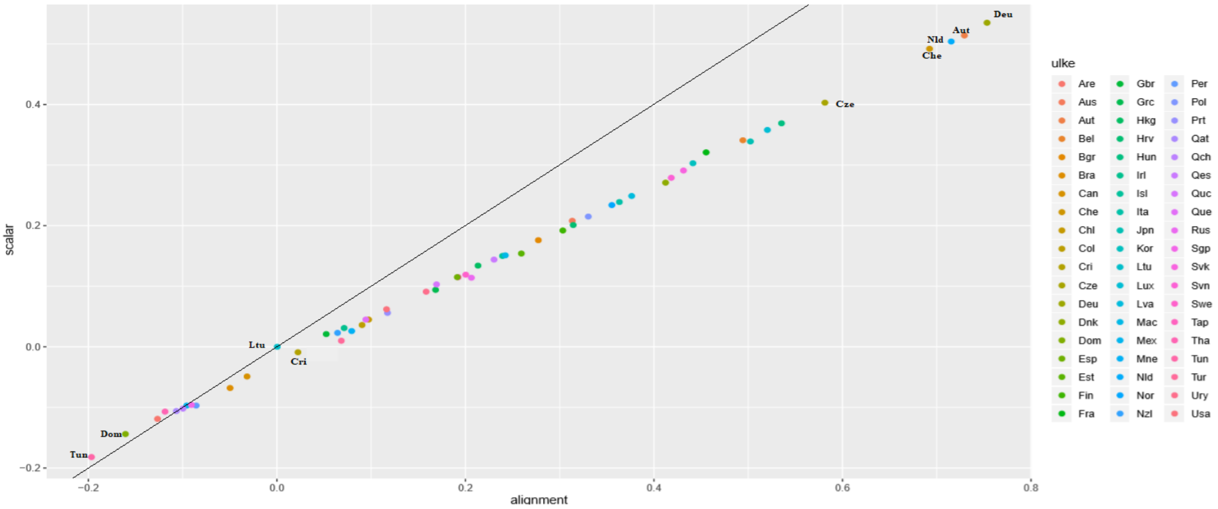


Figure 1. *Factor means obtained from the alignment method versus scalar model (57 countries)*

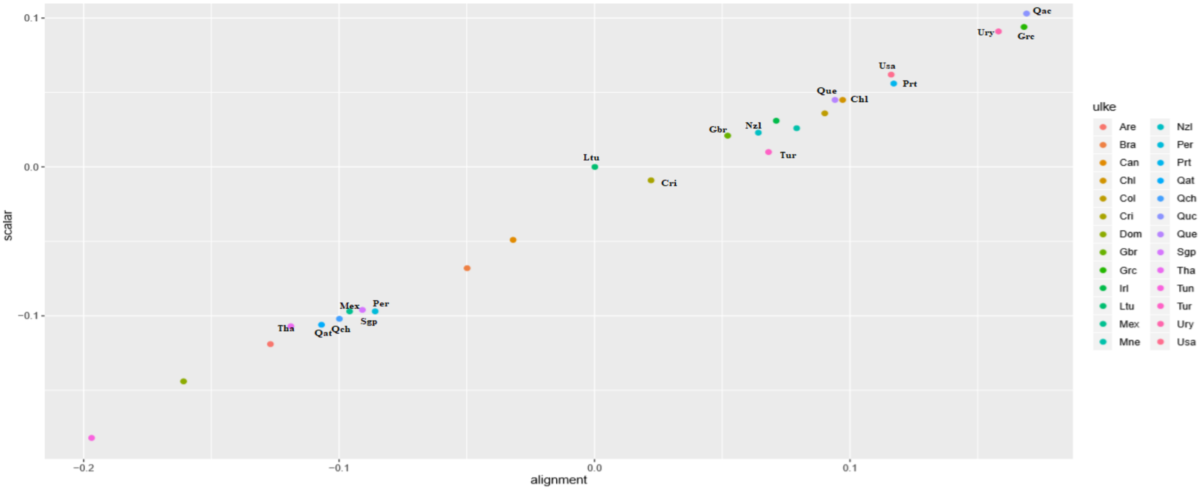


Figure 2. *Factor means obtained from the alignment method versus scalar model (26 countries)*

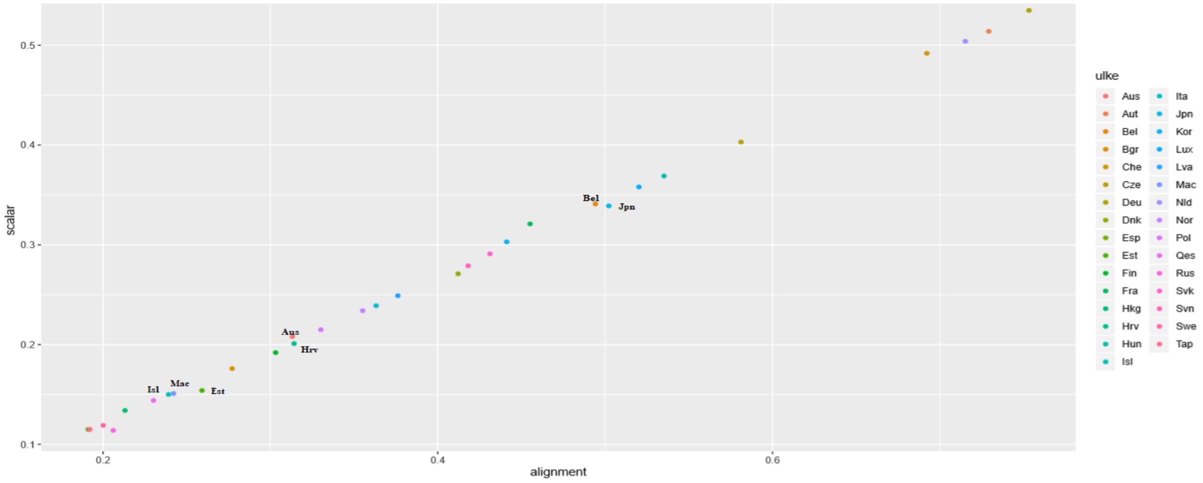


Figure 3. *Factor means obtained from the alignment method versus scalar model (31 countries)*

Figure 1 shows the scatter diagram between the factor means obtained in the alignment method of 57 countries and the factor means obtained from the scalar invariance model of MG-CFA. The fit is most impaired for Germany (Deu, 19); least impaired for Mexico (Mex, 32) and China (Qch, 57) in both methods. The correlation between the factor means obtained from the alignment method and scalar invariance was calculated as 0.999. Despite this high correlation, there are some differences between the two methods. For example; although there are no significant differences between the factor means of Turkey-Lithuania and Turkey-Costa Rica in scalar invariance, Turkey's factor mean calculated in the alignment method is significantly higher than factor means of Lithuania, and Costa Rica. Similarly, in the studies conducted by Asparouhov and Muthén (2014), Muthén and Asparouhov (2018) and Marsh, Guo, Parker, Nagengast, Asparouhov, Muthén and Dicke (2018), there was a high correlation between factor means predicted by strong invariance and the alignment method. But, there was no significant difference for factor means of some countries in scalar invariance while there was a significant difference between the factor means of these countries in the alignment method.

4. DISCUSSION and CONCLUSION

Traditional measurement invariance methods, also known as exact measurement invariance, are inadequate especially in studies comparing large numbers of groups. When the literature is reviewed, it is seen that MG-CFA is frequently preferred in the determination of measurement invariance based on structural equation modeling. However, in studies comparing many groups by MG-CFA, there is almost no study in which full measurement invariance is provided. In these studies, it is generally reported at which stage the measurement invariance is impaired, and the situations where partial measurement invariance is provided with the proposed modifications. While in some studies, no invariance was provided in any model (Gülleroğlu, 2016), in some it was observed that only configural invariance was provided (Hansson & Gustafsson, 2013; Sırgancı & Çakan, 2020). In some of them, it was determined that metric invariance was hold (Asil & Brown, 2015; Pauwels, 2018; İmrol, 2017; Luo, 2010). It has been reported that scalar invariance is held in a few studies (Uzun & Öğretmen, 2010) and rare of the studies, strict invariance is provided (Wu, Li & Zumbo (2007). The presence of many modifications in the MG-CFA both prolong the analysis time and there is no guarantee to determine the best-fit model. Data was collected from many countries with measurement applications such as PISA, TIMSS (The Trends in International Mathematics and Science Study), ESS (The European Social Survey). Due to the reasons mentioned above, in the studies comparing many cultures in large data, it is quite difficult to determine the exact measurement invariance or to determine in which cultures the measurement invariance is provided. Therefore, it is concluded that cross-cultural comparisons will not be valid since measurement invariance is not provided in the studies. However, the alignment method proposed in this study determines the approximate measurement invariance without requiring strict measurement invariance by estimating the group-specific factor mean and variance. The Alignment method is a powerful method to predict the model for multidimensional structures or multiple indicators (items). In this respect, it has essential advantages against MG-CFA method.

Alignment method has main three advantages compared other MI methods in the studies by using multicultural database such as TIMSS, PISA. Firstly, the maximum number of groups in which the measurement invariance is ensured at both item and scale levels can be determined by a single analysis. Secondly, in item-based analysis, the factor load and the intercept of each item can be pairly compared across countries in a one step analysis and also how much each item has contributed to the measurement invariance is determined. Thirdly, it is determined whether the factor means show a significant difference between which countries, and thus, cross-country comparability at the scale level is determined.

In this study, which was carried out using the data of 406.961 participants from 57 countries participating in PISA 2015, it was found that only configural invariance was provided in the investigations analyzed with MG-CFA. Uyar and Uyanık (2019) examined the measurement invariance of the science learning model constituted of measurement instruments in PISA 2015 questionnaires for Turkey and Singapore. As a result, it was found that only configural invariance was provided among these countries. It was concluded that the comparison of the item-scores obtained from the groups may be biased because they did not respond similarly to the items, thus the relationship between measured properties and dimensions of scale is not similar when compared Turkey and Singapore.In this study, the alignment analysis findings show that the intercept and load parameter of the items coded as (ST113.3) and (ST113.4) of the Instrumental Motivation Scale in science learning were equal between Turkey (1) and Singapore (42). In framework of exact MI, it means that the metric invariance is provided for these two items between Turkey and Singapore. Besides, there is no significant difference between the factor means of these two countries. This finding shows that scalar invariance is provided within the framework of exact measurement invariance between these two countries on the scale. Therefore, contrary to the findings of Uyar and Uyanık (2019), the measurement invariance results with the alignment method showed that the scale scores were comparable between these two countries. The findings of MG-CFA of this study showed that only configural invariance was provided for 57 countries. This finding is consistent with the study findings. However, the alignment analysis findings show that the metric invariance is also provided for in the third (ST113.3) and the fourth item (ST113.4) of the Instrumental Motivation Scale in science learning for Turkey (1) and Singapore (42). On the other hand, there is no significant difference between the factor means of these two countries.

In the study conducted by Tiryaki (2019), the measurement invariance of the scales measuring students' attitudes towards science for Turkey and the USA were investigated. When the model fit indexes for ST113 scale were examined, it was found that all invariance stages (configural, metric, scalar and strict invariance) were provided. In this study, the researchers reported that the intercepts and factor loadings were invariance for Turkey and the USA. Besides it was stated that the responses of the items were similar in terms of these two cultures, and the difference in the scores was due to the subgroups. In the same study, it was found that all the items in ST113 had DIF according to the Likelihood Ratio Test based on IRT. However, it was found that factor loadings of the item coded as ST113.3 is invariance for Turkey and the USA in this study differently from Tiryaki (2019). This means that the metric invariance is provided for this item. For other items in ST113, the measurement invariance is not provided for Turkey and the USA. It has also determined the factor means of the USA is significantly higher than Turkey's.

Gür (2019) compared in respect to measurement invariance England-Ireland, England-USA and England-Turkey by using generalized Mantel-Haenszel, poly-SIBTEST and ordinal logistic regression. For England-Ireland (same language-similar culture), the first item had DIF according to GMH, the first and second items had DIF according to OLR and the first, the second and the fourth items had DIF according to poly-SIBTEST. For England-USA (same language-different culture) all of the items had DIF according to OLR, the second, the third items had DIF according to poly-SIBTEST, and the first, the second and the third items had DIF according to GMH. For England-Turkey (different language and culture), it was detected that all items had DIF according to OLR and the second, the third and the fourth items had DIF according to poly-SIBTEST and GMH. In this study, it is concluded that the only factor loadings of the fourth item in ST113 is comparable for England-Ireland and England-Turkey. It means that the metric invariance is provided for this item for these samples. In addition, the factor loadings and measurement intercepts of the all items significantly differentiate for England-USA and the factor means of USA is significantly higher than England's. Besides scalar invariance is provided for the first item and metric invariance is provided for the fourth item for Turkey-Ireland. The factor means of these countries do not differ significantly.

When the findings of this study and the studies mentioned above are evaluated together, it was seen that the alignment analysis presented information that is more detailed as opposed to the traditional methods about measurement invariance between countries. Generally, the configural invariance is provided in studies where full measurement invariance is tested. This and other studies show that when the cultural differences increase, the measurement invariance is impaired (Asil & Gelbal, 2012; Ercikan & Koh, 2005; Kıbrıslıoğlu, 2015; Yandı, Köse & Uysal, 2017). In these studies, pairwise comparisons of the groups predicted to reflect cultural differences made by the researchers. However, it is discovered between which groups the measurement invariance is provided with an exploratory approach with the alignment method. In addition, information is provided on how much each item in the measurement instrument contributes to the invariance between comparison groups. In the case of exact measurement invariance studies, only in case of scalar invariance, which is very rare, factor means can be compared between groups. In the alignment method in which the approximate measurement invariance is tested, it is calculated between which groups the factor means differ statistically, so it is possible to compare factor means between the groups. Thus, the problem of comparability of factor means encountered in scalar invariance in traditional MG-CFA is solved with the alignment method thanks to the factor means estimated for all groups.

Acknowledgements

A part of this study was presented as an oral presentation at the 11th International Statistics Congress (4-8 October 2019, Turkey, Bodrum) with the title "Alignment Method: Concept and Application".

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).

ORCID

Gözde SIRGANCI https://orcid.org/0000-0003-4824-5413
Gizem UYUMAZ https://orcid.org/0000-0003-0792-2289
Alperen YANDI https://orcid.org/0000-0002-1612-4249

5. REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Asil, M., & Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği. *Eğitim ve Bilim, 37*(166), 236-249.

Asil, M., & Brown, G. T. L. (2015). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing, (16)*1, 71-93. https://doi.org/10.1080/15305058.2015.1064431

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495-508. https://doi.org/10.1080/10705511.2014.919210

Byrne, B.M., Shavelson, R.J., & Muthén, B.O. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (4th ed.). Thousand Oaks, California, USA: Sage Publications.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*(4), 662-80.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versionf of TIMSS. *International Journal of Testing, 5*(1), 23-35. https://doi.org/10.1207/s15327574ijt0501_3

Fan, X., & Sivo, S. A. (2009). Using Δ-goodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(1), 54-69, https://doi.org/10.1080/10705510802561311

Gülleroğlu, H. D. (2017). PISA 2012 matematik uygulamasına katılan türk öğrencilerin duyuşsal özelliklerinin cinsiyete göre ölçme değişmezliğinin incelenmesi, *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 37*(1), 151-175.

Gür, E. (2019). *PISA 2015 uygulamasındaki maddelerin kültüre göre değişen madde fonksiyonu acısından incelenmesi* (Master's thesis). Hacettepe University, Graduate School of Educational Sciences, Ankara, Türkiye.

Hansson, Å., & Gustafsson, J. (2013). Measurement invariance of socioeconomic status across migrational background, *Scandinavian Journal of Educational Research, 57*(2), 148-166. https://doi.org/10.1080/00313831.2011.625570

İmrol, F. (2017). Pisa 2012 *Türkiye örnekleminde matematiğe yönelik motivasyon ve öz-inanç yapılarının ölçme değişmezliğinin incelenmesi* (Master's thesis). Ankara University Graduate School of Educational Sciences, Ankara, Türkiye.

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika, 71*(1), 173-191. https://doi.org/10.1007/s11336-003-1136-B

Jöreskog, K.G., Sörbom, D., Du Toit, S.H.C., & Du Toit, M. (2001). *LISREL 8: New statistical features* (3rd ed.). Lincolnwood, IL: Scientific Software International.

Kıbrıslıoğlu, N. (2015*). PISA 2012 matematik öğrenme modelinin kültürlere ve cinsiyete göre ölçme değişmezliğinin incelenmesi: Türkiye, Çin (Şangay)-Endonezya örneği.* (Master's thesis, Hacettepe University, Graduate School of Educational Sciences, Ankara, Türkiye). Retrieved from http://openaccess.hacettepe.edu.tr:8080/xmlui/bitstream/handle/11655/1843/70bd9399-b70f-414a-a11f-b81218adb77c.pdf?sequence=1

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: a comparison of five approaches, *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524-544. https://doi.org/10.1080/10705511.2017.1304822

Luo, C. (2010). *Measurement invariance of Rosenberg Self-Esteem Scale between British and Chinese college students.* (Master's thesis The University of Edinburgh, Edinburgh, Scotland). Retrieved from https://pdfs.semanticscholar.org/274b/015fedaa0c835d94680df2a3c153aff36e5a.pdf

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). *What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. Psychological Methods, 23*(3), 524-545. https://doi.org/10.1037/met0000113

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *ETS Research Report Series, 2*, i-28. https://doi.org/10.1002/j.2333-8504.1994.tb01618.x

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: alignment and random effects. *Sociological Methods & Research, 47*(4) 637-664. https://doi.org/10.1177/0049124117701488

Pauwels, J. (2018) Do you really measure the same? (Doctoral dissertation, Ghent University, Gent, Belgium). Retrieved from: https://lib.ugent.be/fulltxt/RUG01/002/481/799/RUG01-002481799_2018_0001_AC.pdf

OECD (2016). *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264267510-en

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/ 0013164413498257

Sırgancı, G., & Çakan, M. (2020). Sirali lojistik regresyon ve poly-sibtest yöntemleri ile değişen madde fonksiyonunun belirlenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *20*(1), 705-717.

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/SOC/2019. Retrieved from https://wp.hse.ru/data/2019/07/09/1480015921/86SOC2019.pdf

Sörbom, D. (1989). Model modification. *Psychometrika, 54*, 371–384.

Tiryaki, F. (2019). *PISA 2015 öğrenci tutum anketlerinin değişen madde fonksiyonu ve ölçme değişmezliğinin incelenmesi* (Master's Thesis). Ankara University Graduate School of Educational Sciences, Ankara, Türkiye

Uyar, Ş. (2011). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi (Master's Thesis). Hacettepe University Graduate School of Educational Sciences, Ankara, Türkiye

Uyar, Ş., & Uyanık, G. K. (2019). Fen bilimlerine yönelik öğrenme modelinin ölçme değişmezliğinin incelenmesi: Pisa 2015 örneği. *Kastamonu Eğitim Dergisi, 27*(2), 497-507. https://doi.org/10.24106/kefdergi.2570

Uzun, B., & Öğretmen, T. (2010). Fen başarısı ile ilgili bazı değişkenlerin timss-r türkiye örnekleminde cinsiyete göre ölçme değişmezliğinin değerlendirilmesi. *Eğitim ve Bilim, 35*(155), 26-35.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70. https://doi.org/10.1177/109442810031002

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). Washington DC, USA: American Psychological Association. https://doi.org/10.1037/10222-009

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial ınvariance and updating the practice of multigroup confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*, 1-26. https://doi.org/10.7275/mhqa-cd89

Yandı, A., Köse, İ. A., & Uysal, Ö. (2017). Farklı yöntemlerle ölçme değişmezliğinin incelenmesi: PISA 2012 örneği. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 13*(1), 243-253. https://doi.org/10.17860/mersinefd.305952

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (Ordinal) item scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

6. APPENDIX

The Mplus script of the alignment analysis:

```
TITLE: align MODEL
DATA: file is C:\Users\Lenovo\Documents\GGA\2- ST113\st113.dat;
VARIABLE:
variable:
    NAMES ARE country clus u1 u2 u3 u4;
    USEVARIABLES ARE u1 u2 u3 u4;
    MISSING=ALL (999);
    classes = c(57);
    knownclass = c(country = 1 2 3 4 5 6 7 8 9 10 11 12
    13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
    31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
    49 50 51 52 53 54 55 56 57);
ANALYSIS:
    type = mixture;
    estimator = ml;
    alignment = fixed(29);
    astarts=100;
model:
    %overall%
    st113 by u1-u4;
output:
    tech1 tech8 align;
plot:
    type = plot2;
```

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 674–689

https://doi.org/10.21449/ijate.708169

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Study of The Validity and Reliability of Nanotechnology Awareness Scale in Turkish Culture

Zeki Ipek [1,*], Ali Derya Atik [2], Seref Tan [3], Figen Erkoc [4]

[1]Republic of Turkey, Ministry of National Education, Antalya, Turkey
[2]Kilis 7 Aralık University, Matematics and Science Education Department, Kilis Turkey
[3]Gazi University, Department of Measurement and Evaluation in Education, Ankara, Turkey
[4]Gazi University, Department of Biology Education, Ankara, Turkey

ARTICLE HISTORY

Received: March 24, 2020
Revised: Sept. 13, 2020
Accepted: Oct. 22, 2020

KEYWORDS

Nanotechnology,
Nanotechnology awareness,
Nanotechnology exposure,
Validity,
Reliability

Abstract: The aim of this study was to determine the validity and reliability of the Nanotechnology Awareness Scale (NAI) in Turkish culture. The study group consists of 624 biology, physics and chemistry teachers working in secondary schools in Antalya, Denizli, Burdur and Ankara. Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were conducted in order to determine its structural validity. Cronbach-alpha and stratified-alpha coefficient values were calculated for the reliability of the sub-dimensions and the whole of the scale, respectively. In EFA, Kaiser-Meyer-Olkin (KMO) value was found to be 0.92 and the Bartlett Test result (912 91 = 6519.27, p <.00) was found significant. In addition, in original scale, a two-factor structure was found. This two factors solution explains 59.192% of the total variance. In CFA, factor structure of the scale was tested for two-factor solution as it was designed. According to the findings, it was found that the scale, containing 14 items with two sub-dimensions, had sufficient goodness of fit indices such as $\chi 2$/sd =1.344, RMSEA = 0.07, GFI = 0.97, CFI = 0.97, NFI = 0.90 and AGFI = 0.83. These goodness of fit indices shows that we have good model-data fit. The Stratified-alpha coefficient was found as 0.942 for the whole scale. Cronbach alpha coefficient was found as 0.935 and 0.805 for the awareness sub-dimension and exposure sub-dimension, respectively. As a result of the research, it was concluded that the Turkish version of the scale can be used as a valid and reliable measurement tool.

1. INTRODUCTION

Nanoscience and nanotechnology (NSNT) is a very rapidly changing field and defined as one of the six key enabling technologies (KETs) by the European Commission, including advanced manufacturing technologies; advanced materials and nano-technologies; life science (as a broader definition of industrial-biotechnology); micro and nano-electronics and photonics; artificial intelligence; security and connectivity (Jackman et al., 2016). All are related to digital technologies (e-skills). Nanotechnology is the driving force behind a new industrial revolution with both the private and public sectors constantly spend more in research, innovation and commercialization. The expected growth rates and size of the market for nanotechnology

CONTACT: Zeki İpek ✉ ipekzeki@gmail.com Republic of Turkey, Ministry of National Education, Antalya, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

products is already comparable to the biotechnology sector (OECD, 2018). The initial nanoscale materials have been used in applications such as window glass, sunglasses, car bumpers and paints. Presently, the convergence of scientific disciplines (physics, biology, chemistry, electronics, engineering etc.) as a multidisciplinary field for wider scope and much diverse applications in materials manufacturing, computer chips, medical diagnosis and health care, energy, biotechnology, space exploration, security has introduced "NANO" to our daily lives and is expected to have a significant impact on economy and society, while generates great opportunities for cutting-edge research in science and for innovation in industrial production within the next decade. Furthermore, scientific and technology breakthroughs are expected in the long-term time scale. Nanotechnology increasingly takes the role of an enabling technology to biotechnology. The competitiveness of global industry depends on the effective use of new technologies and the knowledge, skills, competences and creativity of its workforce. Shortages, gaps and mismatches in high-tech skills negatively affect innovation, productivity growth, job creation and social cohesion. Public policies, education and training systems to react in time and funding programmes for high-tech skills development, to identify good practices and to make concrete recommendations for scaling up best practices and re-focusing funding programmes are necessary. (European Commission, 2019; OECD, 2018; Roco et al., 2011). Accordingly, scientists have raised concerns that the free particles smaller than one billionth of a meter as the basic building blocks of NBNT pose a potential new class of risk to health and the environment (Lauderwasser, 2005).

The motivation for continuous education and training and life-long-learning is the potential key for developing countries in the world NSNT market and industry workforce. Although NSNT is well established in academia and industry, the current generation of science teachers have insignificant exposure to NSNT, and few opportunities to understand the basic concepts (Andina et al., 2019; Hingant & Albe, 2010; Pas et al., 2019; Winkelmann & Bhushan, 2016). Since teachers are the force shaping students who will be the citizens of the future they need to be prepared for these developments in research, innovation and economy based on the new generation technologies (Jones et al., 2013; Laherto, 2010; Wansom et al., 2009).

Presently NSNT is not in neither the Primary School Science nor Secondary School Biology curricula. Nanotechnology topics are limited to 12th grade physics and chemistry courses in secondary schools with two lecture hours (Ministry of National Education Turkey (MoNE), 2013, 2017). There is a similar situation in higher education undergraduate courses in Turkish universities which offer courses as electives only in engineering and science faculties of few universities with limited content. However, increasingly more universities have NSNT education in their graduate level programs; but resources in Turkish are very rare (İpek et. al., 2017). NSNT education is still a relatively new field. Therefore, Turkey also needs to adapt such policies and changes for NSNT education and research. In addition, there is a necessity, at the international level, to update nanotechnology topics in secondary level physics, chemistry and biology curricula to incorporate scientific developments (Roco & Bainbridge, 2003). Having a good NSNT education at primary and secondary levels will be effective in the students' academic self-development, future career choices, citizen-science, and science communication (Karataş & Ülker, 2014). The need for research to measure teachers' awareness, attitudes and knowledge about NSNT was the driver behind our research. Level of awareness of secondary level school biology, physics and chemistry teachers is evaluated using a cultural adaptation of the NAI (Nanotechnology Awareness Instrument) developed by Dyehouse et al. (2008) with validity and reliability to Turkish culture. Our results would contribute to planning for science teachers' training needs and draw attention of academia and private sector, research institutes to NSNT education to start in pre-higher education levels to promote student academic self-development, future career choices, citizen-science, and science communication for future. There is no measurement tool to measure Nanotechnology Awareness in Turkish Language.

Therefore, the purpose of this article is to adapt NAI to Turkish language and to study the validity and reliability of this Turkish version obtained.

2. METHOD

Research model, study group, data analysis and cross-cultural adaptation of NAI instrument originally developed by Dyehouse et al. (2018) are presented in this section.

2.1. Study Design

Researchers use different methods for the translation, adaptation and cross-cultural validation of an original instrument and target culture by considering the differences between the original source and the target culture while maintaining equivalence of meaning. The process of translation, adaptation, and cross-cultural validation of an instrument for use on other cultures, languages, and countries requires careful planning and adaptation of comprehensive methodological approaches (Sousa & Rojjanasrirat, 2011). The purpose of translation is to achieve equivalence between the instrument in the original language and the scale in the target language (Chapman & Carter, 1979). Pilot testing of the pre-final version of the scale is used to be easily understood by the target population prior to psychometric testing (cognitive debriefing). Full psychometric testing of the pre-final version of the scale among individuals from the target population used to establish internal consistency reliability (or sensitivity and specificity), stability reliability, homogeneity, construct-related validity, criterion-related validity, factor structure and model fit of the instrument.

2.2. Study Group

The study group consisted of physics, chemistry, and biology teachers participate in distinct cities. The respondents who were participated in in-service training or a course from a faculty member on this subject or following scientific journals, documentaries, etc. were selected as far as possible. The interviews were conducted face to face, and the teachers were asked to write their own answers to the written documents. The respondents participated in the study voluntarily. The require ethical approval was taken from Gazi University Ethical Committee (Date/Number: 19.02.2016 / 81576613/605/1955049).

46.2% (n=288) of the respondents were female and 53% (n=336) were male. The seniority in the profession distribution was: 1-5 years 17.6% (n=110); 6-10 years 13.9% (n=87); 11-15 years 15.9% (n=99); 16-20 years 18.3% (n=114); 21-25 years 19.7% (n=123); 26 years and over 14.6% (n=91). When the respondents' branches were examined, 30.6% (n=191) was physics, 25.6% (n=160) was chemistry, and 43.8% was biology. 26% (n=162) of the respondents have Master's degree and 2.4% (n=15) have PhD degree. The respondents of 10.1% (n=63) teacher worked in Science High School, 54.2% (n=338) in Anatolian High School, and 35.7% (n=223) Vocational High School. The cities where the respondents work were~~-~~Antalya (36.2%), Denizli (34.8%), Burdur (12.7%), and Ankara (16.3%) during 2015/2016 acedemic year.

2.3. The Procedure of the Cross-cultural Validity and Reliability Testing

The process of study is given below;

a. Translation of the original instrument into the Turkish (forward translation or one-way translation):

Permission to adapt NAI to Turkish culture and use was obtained via e-mail communication from Dr. Dyehouse (Learning Systems Institute, Florida State University, Tallahassee, FL, USA).

The original instrument NAI consists of two parts. Part A: Awareness (eight items), Exposure (six items), and Motivation (five items). The *Awareness* subscale aims to measure how aware respondents are of the impact and application of nanotechnology. The *Exposure* subscale measures the respondents' exposure and experiences with nanotechnology. The *Motivation* subscale contains five statements that describe types of related nanotechnology activities. Part B: Nano-Knowledge (eight items), Nanotechnology Uses and Equipment (seven items). The *Knowledge* subscale items included a correct answer and measures knowledge about nanotechnology facts. The second subscale provides information about how familiar respondents are with nanotechnology uses and equipment (Dyehouse et al., 2008). We used the subscales of *Awareness* and *Exposure*, with 14 items. Furthermore, İpek (2017) developed the additional third subscale with 10 items to measure knowledge about NSNT. These were reduced to five items after opinions of a field, an education and a measurement and assessment experts were obtained, and the final version was elaborated. This knowledge subscale aims to increase reliability of the data collected from the instrument in awareness subscale by improving sincerity of the responses. In subscale C (knowledge), respondents answer some questions related to awareness subscale. These questions are presented in the form of fill in the blank and "*open-ended"* questions. Responses of teachers to these questions are assessed to determine their real knowledge level of NSNT.

The NAI in the original language is forward translated to Turkish by two independent translators. The translators (two academicians and one language editor) speak English fluently and have in-depth experience in the culture of the source. And, translators have distinct backgrounds. The first translator is the physics professor who studies in the field of nanotechnology. So, he has knowledge about nanotechnology terminology, and the subject area of the construct of the instrument. The second translator is familiar with assessment and evaluation in education. To achieve equivalence between the instrument in English and the instrument in the Turkish, well-qualified translators were chosen. The items were first translated by two with specialty in translation (Version 1 translation a: V1a and Version 1 translation b: V1b).

b. Comparison of the two translated versions of the instrument

The third bilingual independent expert in grammar and language compared the items of the two forward translated versions (V1a and V1b) of the instrument (V1a and V1b) and both the V1a and the V1b with the original version of the instrument (NAI) words, sentences and meanings.

c. Blind back-ward (double) translation and comparation of the two back-translated versions of the instrument

The comparison of the two translated versions of the instrument translated back into English by four other independent another translator that had never seen the original version of the instrument with the same qualifications and characteristics. These four experts had undergraduate degrees in English language and literature. They produced two back-translated versions of the instrument. This step allowed for clarification of words and sentences used in translations. The experts compared the translations with the items of the original instrument NAI regarding format, the grammatical structure of the sentences, the similarity in meaning, and relevance. The two translated texts were merged, and experts agreed that it was consistent with the original instrument NAI. A native English speaker expert, who also has good command of Turkish reviewed the original and translated versions of the instrument and confirmed consistency of the Turkish version with the original. The equivalence of the items in the scale in terms of semantic, idiomatic, experiential, and conceptual aspects were reviewed by two experts in the field. Some items were amended based on opinions of these experts. Thus, pilot testing of the pre-final version (PF-V) of the instrument was prepared.

d. Pilot testing of the pre-final version of the instrument

The participants whose language is Turkish tested the PF-V of the instrument evaluated the items of the instrument for clarity. A sample size of 10-40 individuals is recommended for pilot testing the instrument (Beaton et al., 2000). Participants response format or any item of the instrument as unclear asked to provide suggestions as to how to write the statements to make the language clearer. The sample is 20 students, seniors of English Teacher Program of Faculty of Education, Akdeniz University were given the Turkish and English versions with a one-week interval and to confirm consistence/agreement and no problems were reported in comprehension of the items. The final version (FV) was termed Nanoscience and Nanotechnology Awareness Scale (NSTAS).

e. Testing the FV of the translated instrument in a sample of the target population

This last step used to establish the initial full psychometric properties of the newly translated, adapted and cross-validated instrument with a sample of the target population of interest. The sample depends on the types psychometric approaches that will be used. In general, it is highly recommended to use at least 10 subjects per item of the instrument scale and item analysis and exploratory factor analysis. If there is a plan to use confirmatory factor analysis to test the factor structure of the instrument, the recommendation per rule of thumb is approximately 300–500 subjects per item of the instrument (Tabachnick & Fidell, 2001). To testing the NSTAS in a sample of the target population was participated 624 teachers form distinct branches (physics, chemistry, and biology) and cities. Sample size significance for factor analysis was satisfactory.

2.4. Data Analysis

For the FV, factorial structure of the instrument, structural validity, reliability of scale scores and item discrimination indexes were assessed. EFA and CFA were carried out for construct validity of the scale. Reliability of the subscales were calculated using Cronbach's alpha and stratified alpha coefficient for overall reliability of the scale.

3. RESULT / FINDINGS

3.1. Results of Exploratory Factor Analysis

Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) statistic indicating the proportion of variance in our variables which might be caused by underlying factors for applicability of items in Nanoscience and Nanotechnology Awareness Scale (NSTAS) was .92; values around .90 indicating that factor analysis will be useful (Schermelleh-Engel et al., 2003; Young & Pearce, 2013). Bartlett's Test of Sphericity was found to be significant ($\chi^2_{91} = 6519.27$, $p <.00$); showing that correlation matrix is not an identity matrix, indicating our variables to be related and therefore suitable for dimension reduction for creating factors.

Exploratory factor analysis was the first procedure applied to our data set to examine the factorial structure. Since theoretically only two factors were in the original Instrument, we limited factors to two in exploratory factor analysis, and in this case, principle axis factoring method was used as factor extraction method (Tan, 1999). The factor loadings were found to be above .50 each, indicating strong loadings (Seçer, 2013).

Both factors have Eigenvalues above 1 (Table 1), the items gather under the two factors and explain 59.191% of the total scale variance (Zwick & Velicer, 1986). Our analysis shows that two factor analysis is appropriate since explained total variations by two factors is very close to 60% (Bektaş, 2015; Deniz et al., 2013).

Table 1. *Total Variances Explained for NSTAS*

Items	Total Variances Explained		
	Eigen Value	Variance (%)	Total (%)
1	7.019	50.138	50.138
2	1.267	9.053	59.191

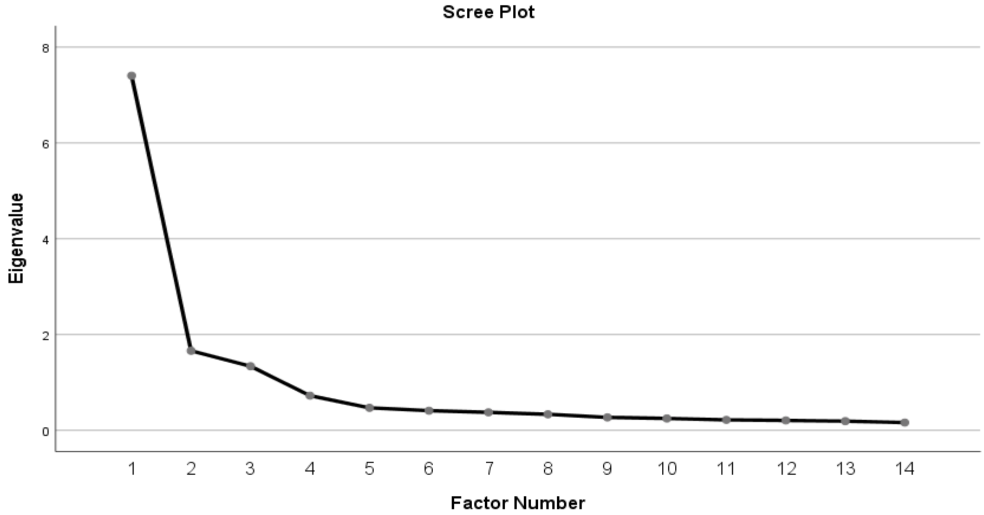


Figure 1. *Scree plot of the eigenvalues of the factors for NSTAS.*

The scree plot graph based on Eigenvalues of the factors is depicted in Figure 1, this enables us to give a final decision on number of factors (Henson & Roberts, 2006). The slope of the scree plot is in the same direction after factor 2 (Figure 1), verifying the number of factors as two (Thompson, 2004; Zwick & Velicer, 1986). The first factor is named "Awareness" and the second "Exposure" keeping the structure of the original Instrument.

Table 2 depicts results of the direct oblimin method where distribution of the items to factors are investigated since the factors are correlated this oblique rotation employed (Osborne & Costello, 2005).

Table 2. *Distribution of Factor Loadings by Factors of NSTAS Items*

Items	Factors	
	1	2
A_1	.88	
A_2	.87	
A_3	.76	
A_4	.89	
A_5	.86	
A_6	.56	
A_7	.53	
A_8	.75	
B_1		.39
B_2		.49
B_3		.52
B_4		.82
B_5		.82
B_6		.73

Table 2 shows cluster of eight items in the Awareness factor, and six items in the Exposure factor. Factor loadings of the items in the first factor ranged between .52 to .87, the second factor had .39 to .82. There is a positive but moderate correlation between the first and second factors (r = .58, $p < .05$). These results are the same as the original factorial structure of NAI (The two factors of NAI were moderately correlated at 0.52 and NanoAwareness and Nano-Exposure items, loaded onto factor 2. Pattern coefficients for the second factor ranged from 0.26 to 0.85) (Dyehouse et al., 2008).

3.2. Reliability and Validity of NSTAS Scores

Awareness subscale items have corrected total score correlation coefficients of .69 to .82 in Table 3.

Table 3. *Corrected Item Total Score Correlations for Awareness Sub-scale Items of NSTAS*

Items	Corrected Item Total Correlation	Cronbach's Alpha If Item Deleted
A_1	.78	.93
A_2	.79	.93
A_3	.77	.93
A_4	.82	.92
A_5	.82	.92
A_6	.72	.93
A_7	.69	.93
A_8	.78	.93

Exposure subscale items have corrected total score correlation coefficients of .54 to .71 in Table 4. We conclude that total corrected item correlations are satisfactory (very high). In other words, all items in the sub-scales have very high item discrimination power.

Table 4. *Corrected Item Total Score Correlations of B (Exposure) Sub-Dimension Items of NSTAS*

Items	Corrected Item Total Correlation	Cronbach's Alpha If Item Deleted
B_1	.54	.84
B_2	.71	.81
B_3	.69	.82
B_4	.69	.81
B_5	.64	.82
B_6	.54	.84

Table 5. *Internal Consistency Coefficients for NSTAS*

Sub-dimension	Variance	Cronbach's alpha Coefficient	Stratified Cronbach's alpha Coefficient
A: Awareness	67.66	0.93	-
B: Exposure	28.98	0.85	-
NSTAS	152.09	-	0.94

As it given in Table 5, stratified Cronbach**'s** alpha coefficient (Feldt & Qualls, 1996) of NSTAS (overall reliability) was found as .94. Calculation of the stratified Cronbach's alpha coefficient is the preferred method of reporting reliability estimates when multifactors are concerned (Tan, 2009). Cronbach's alpha coefficients were found to be .93 for Awareness and .85 for Exposure subscales, therefore the sub-scales and NSTAS have very high reliability levels.

3.3. Results of Confirmatory Factor Analysis

CFA was carried out to decide whether the two factors model is confirmed or not. Goodness of fit indices are depicted in Table 6: χ2/df =1.344, RMSEA = .07, GFI = .97, CFI = .97, NFI = .90 and AGFI = .83, all findings are at good and excellent level (Hu & Bentler, 1999; Schermelleh Engel et al., 2003; Kline, 2005).

Table 6. *Comparison of Standard Goodness of Fit Measures and Research Results*

Goodness of Fit Measure	Good Fit	Acceptable Fit	Goodness of fit Values Obtained in the Research
χ2/df	0≤χ2/df≤2	2≤χ2/df≤3	1.344
RMSEA	0≤RMSEA≤0.05	0.05≤RMSEA≤0.08	0.07
NFI	0.95≤NFI≤1.00	0.90≤NFI≤0.95	0.90
CFI	0.97≤CFI≤1.00	0.95≤CFI≤0.97	0.97
GFI	0.95≤GFI≤1.00	0.90≤GFI≤0.95	0.97
AGFI	0.90≤AGFI≤1.00	0.85≤AGFI≤0.90	0.83

AGFI value is .83, below the acceptable limit of .85. However, since it is close to the acceptable limit value of .85 and χ2/df value calculated as 1.344, smaller than 3, when we consider the values given in Table 6, we conclude that the scale has good fit indices (Schermelleh-Engel et al., 2003). The results confirm the measurement model of NSTAS, given in Figure 2.

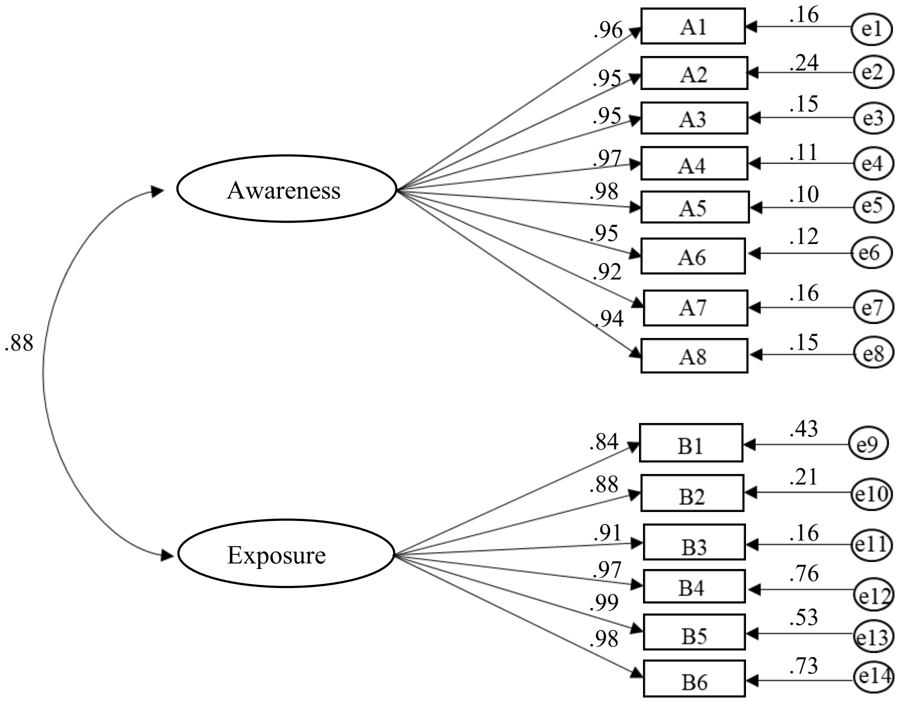


Figure 2. *Confirmatory Factor Analysis Structural Modeling of NSTAS to Awareness (A) and Exposure (B) Sub-dimensions*

As it seen in Figure 2, the Awareness and Exposure factors (latent variables) correlated with each other. Each observed variable is loaded on the related factor and all factor loadings and factor correlations in the model were found to be significant. The value of .88 depicting correlation between the two latent variables confirms that Awareness and Exposure latent variables are correlated. Standardized coefficients of the measurement model provides view that each item (observed variable) is a good representative of its' latent variable (Çelik & Yılmaz, 2013). In conclusion, the results of CFA show that this measurement model confirms the structural validity of the NSTAS scores. So, the measurement of applying the NSTAS to 624 teachers for cross-cultural adaptation to Turkish are statistically proven to be valid and reliable.

4. DISCUSSION and CONCLUSION

The rapid development of nanotechnology and its impact on the economy has led to the focus on nanotechnology education. Nanotechnology is a relatively new field, and as such is not yet widely understood by the society. Because it is difficult to train qualified manpower of the mentioned size easily and in a short time (Laherto, 2010). Many educational interventions are being implemented to address workforce issues in the field of nanotechnology. When the literature is reviewed, it is seen that nanotechnology education has started to take its place at the secondary education level in many developed countries. According to the new researches, nanotechnology education must start earlier grades such as middle and primary schools. Many countries, several nanotechnology training materials for the K-12 level have already been developed. Although the researches and planning the activities on NST education started from the primary school in USA, Europe and developed countries, NST educational researches are limited and curricular planning studies are not sufficient in Turkey. Nanotechnology is not mentioned at all grade levels in the curricula of science courses in primary schools and Biology courses in secondary schools, currently. The public and the next generations are informed about nanotechnology are crucial.

To increase the awareness of NST, education and training activities from primary education to university level should be organized in a conscious and programmed manner. The most important and first step in this process is to determine the current situation. To determine the awareness and knowledge levels of biology, physics, and chemistry teachers are working in secondary schools is a critical and important beginning to show the current situation. However, there is no Turkish instrument to assess the teachers/students' awareness of and exposure to nanotechnology. To address this need, the study of the validity and reliability of NAI in Turkish culture was conducted. This research is expected to be a source for researchers who will work on NST in Turkey. The research has been conducted considering the following limitations: The research calendar is limited to the 2015/2016 academic year. The research is limited to 624 biology, physics and chemistry teachers working in secondary education institutions located in the city center of Antalya, Ankara, Denizli and Burdur. The content of the research is limited to the awareness of teachers about the NSTAS subject, its teaching in secondary education, and the NST Awareness Scale. Research is limited with the resources available.

When developing the original instrument NAI, an explanatory factor analysis (EFA) for the Awareness, Motivation, and Exposure subscales was conducted to determine what underlying constructs were being measured. A principal factors analysis with a promax rotation (power=3) was used. Originally, five factors were retained. However, only four variables loaded onto the fifth factor. The scree plot revealed a levelling off after three factors. So a 3-factor solution was obtained. The factor solution indicated that Motivation can be considered a separate construct from Awareness and Exposure, although the constructs are moderately related. Awareness and Exposure subscales are similar enough to be considered a single construct. Because being aware of nanotechnology probably means that one has also been exposed to nanotechnology (Dyehouse et al., 2008).

Firstly, EFA applied in this study that was investigated the validity and reliability of NSTAS in Turkish culture. As a result of the eigenvalue, scree plot and explained total variance proportion examinations, it was determined that the number of scale factors was two as it was in the original scale and these two factors explained 59.19% of the total variance of the data. Also, the items have high factor loadings with related factor in the scale. These factor loading values and the CFA results show that the adapted instrument provides structural validity in Turkish culture.

The high "corrected item-total score correlations" for items in both sub-dimensions of NSTAS indicate that the item discrimination indexes are very high. Also, as in the original scale, moderate positive correlation (r = .58 p <.05) between the awareness and exposure factors of NSTAS is another proof that compliance with the original scale structure is achieved.

The original instrument's (NAI) internal consistency Cronbach's alpha coefficient values were calculated for the Awareness scale was 0.91 and for the Exposure scale was .82. These coefficients showed a satisfactory level of internal consistency reliability (Dyehouse et al., 2008). The stratified alpha coefficient (.94) for the reliability of the NSTAS scores, Cronbach-alpha coefficient for the awareness (.93) sub-scale scores and Cronbach-alpha coefficient for the exposure (.85) sub-scale scores showed that the adapted scale was highly reliable (Özdamar, 1999). For this reason, the Turkish version of the scale, composed of 14 items, gives valid and reliable results.

Furthermore, the model fit indices were tested with CFA and model goodness of fit indices were found to be good and excellent (see Table 6) (Schermelleh Engel et al., 2003). During the data collection process and the interviews with teachers, the participants stated that the newly developed knowledge sub-scale (C) section by the authors was a very effective approach to increase the reliability of the scale scores. In the scale development to measure especially affective domain, it is considered to be useful to use such sub-scale (C) that will verify the data obtained to increase the reliability of the scale or sub-scale scores.

As a result of the validity and reliability findings, the adopted scale can be used to determine the nanoscience and nanotechnology awareness of the secondary school biology, chemistry and physics teachers in Turkey.

The practical implication of this instrument may be able to determine the teachers and students' awareness of and exposure to nanotechnology. And it may be able to use to gather information about whether a program is effective for increasing their awareness and knowledge. Knowing the degree to which programs increase teachers/students' awareness, knowledge, and motivation will also aid in making curricular design decisions. Nanotechnology as a field of study or career provides opportunities for the next generation. This instrument may also be a valuable tool for students to draw attention nanotechnology as a field of study and carrier.

Acknowledgements

Preliminary results of this work have been presented in C-MEEP 2018 6th International Congress on Measurement and Evaluation in Education and Psychology, September 5-8, 2018, Prizren, Kosovo.

Declaration of Conflicting Interests and Ethics

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

ORCID

Zeki İpek https://orcid.org/0000-0002-8097-5849
Ali Derya Atik https://orcid.org/0000-0002-5841-6004
Şeref Tan https://orcid.org/0000- 0002-9892-3369
Figen Erkoç https://orcid.org/0000-0003-0658- 2243

5. REFERENCES

Andina, R. E., Rahmawati, Y. & Budi, S. (2019). Improved learning designs for shaping Indonesia's future science teachers applied in a nanoscience project. *Issues in Educational Research*, *29*(4), 997-1015. http://www.iier.org.au/iier29/andina.pdf

Bektaş, H. (2015). *Factor analysis for binary variables: An application on the quality of working life.* Doctoral dissertation, İstanbul University, İstanbul, Turkey. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Beaton, D. E., Bombardier, C., Guillemin, F. & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine, 25*(24), 3186-3191.

Chapman, D. W. & Carter, J. F. (1979) Translation procedures for the cross-cultural use of measurement instruments. *Educational Evaluation and Policy Analysis*, *1*(3), 71-76. https://doi.org/10.3102/01623737001003071

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), 1-9.

Çelik, H. E., & Yılmaz, V. (2013). *Yapısal eşitlik modellemesi: temel kavramlar, uygulamalar, programlama* (2nd ed.). Anı Yayıncılık.

Deniz, M. E., Özer, E., & Işık, E. (2013). Duygusal zekâ özelliği ölçeği–kısa formu: Geçerlik ve güvenirlik çalışması [Trait Emotional Intelligence Questionnaire–Short Form: Validity and reliability studies]. *Education and Science, 38*(169), 407-419.

Dyehouse, M. A., Diefes-Dux, H. A., Bennett, D. E., Imbrie, Æ P. K. (2008). Development of an instrument to measure undergraduates' nanotechnology awareness, exposure, motivation and knowledge. *J Sci Educ Technol.*. 17, 500-510.

European Commission (2019). Skills for Industry. High-Tech Skills: Scaling up best practices and re-focusing funding programmes and incentives, Final Report. Executive Agency for Small and Medium-sized Enterprises (EASME), EASME/COSME/2018/016, 2019. Luxembourg: Publications Office of the European Union, 2019. PDF ISBN 978-92-9202-548-9. https://doi.org/10.2826/024306 EA-01-19-571-EN-N

Hingant, B. & Albe, V. (2010). Nanosciences and nanotechnologies learning and teaching in secondary education: A review of literature. *Studies in Science Education, 46*(2), 121-152. https://doi.org/10.1080/03057267.2010.504543

Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education*, *9*(3), 277-286.

Henson, R. K. & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416.

Hu, L. & Bentler, P. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

İpek, Z. (2017). *Research on awareness levels of physics, chemistry, and biology teachers about nanoscience and nanotechnology.* Doctoral Dissertation, Gazi University, Ankara. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

İpek, Z., Atik, A. D., Tan, Ş. & Erkoç, F. (2020). Awareness, exposure, and knowledge levels of science teachers about nanoscience and nanotechnology. *Issues in Educational Research*, *30*(1), 134-155. http://www.iier.org.au/iier30/ipek.pdf

Jackman, J. A., Cho, D. J., Lee, J., Chen, J. M., Besenbacher, F., Bonnell, D. A., … Cho, N. J. (2016). Nanotechnology education for the global world: Training the leaders of tomorrow. *ACSNano*, 10, 5595−5599. https://doi.org/10.1021/acsnano.6b03872

Jones, M. G., Blonder, R., Gardner, G. E., Albe, V., Falvo, M., & Chevrier, J. (2013). Nanotechnology and Nanoscale Science: Educational Challenges. *International Journal of Science Education, 35*(9), 1490-1512.

Kalaycı, N. (2008). *Yükseköğretimde uygulanan toplam kalite sürecinde göz ardı edilen unsurlardan "TKY merkezi" ve "eğitim programları"*. Tekışık.

Karataş, F. Ö., & Ülker, N. (2014). Undergraduate chemistry students' understanding level of nano-science and nano-technology. *Journal of Turkish Science Education, 11*(3), 103-118. https://doi.org/10.12973/tused.10121a

Kline, R. B. (2005). *Principles and practice of structural equation modeling (second edition)*. Guilford Publications, Inc.

Laherto, A. (2010). An Analysis of the Educational Significance of Nanoscience and Nanotechnology in Scientific and Technological Literacy. *Science Education International, 21*(3), 160-175.

MoNE. Talim ve Terbiye Kurulu Başkanlığı (2013). *İlköğretim Kurumları (İlkokullar ve Ortaokullar) Fen Bilimleri Dersi (3, 4, 5, 6, 7 ve 8. Sınıflar) Öğretim Programı*. Doc Player. https://docplayer.biz.tr/1747250-Fen-bilimleri-dersi-3-4-5-6-7-ve-8-siniflar.html

MoNE. Talim ve Terbiye Kurulu Başkanlığı (2018). *Ortaöğretim Biyoloji, Fizik ve Kimya Dersi (9, 10, 11 ve 12. Sınıflar) Öğretim Programı.* http://mufredat.meb.gov.tr/Programlar.aspx

Lauterwasser, C. (Ed.) (2005). *"Small sizes that matter: Opportunities and risks of Nanotechnologies", Report in co-operation with the OECD International Futures Programme.* http://www.oecd.org/dataoecd/32/1/44108334.pdf

OECD (2018). Report on statistics and indicators of biotechnology and nanotechnology. OECD Science, Technology and Industry Working Papers 2018/06. Paris, France. https://dx.doi.org/10.1787/3c70afa7-en

Özdamar, K. (1999). *Paket programlar ile istatistiksel veri analizi*. Kaan.

Pas, M., Vogrinc, J., Raspor, P., Knezevic, N. U. & Zajc, J. C. (2019). Biotechnology learning in Slovenian upper-secondary education: Gaining knowledge and forming attitudes. *Research in Science & Technological Education, 37*(1), 110-125. https://doi.org/10.1080/02635143.2018.1491473

Roco, M. C. & Bainbridge, W. (2003). *Societal implications of nanoscience and nanotechnology*. Kluwer.

Roco, M. C., Mirkin, C. A. & Hersam, M. C. (2011). Nanotechnology research directions for societal needs in 2020: Summary of international study. *Journal of Nanoparticle Research, 13*(3), 897-919. https://doi.org/10.1007/s11051-011-0275-5

Schermelleh-Engel, K., Moosbrugger, H. & Müler, 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. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.509.4258&rep=rep1&type=pdf

Seçer, İ. (2013). *SPSS ve LISREL ile pratik veri analizi*. Anı.

Sousa, V.D. & Rojjanasrirat, W. (2011). Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *International Journal of Evaluation in Clinical Practice, 17*, 268-274. https://doi.org/10.1111/j.1365-2753.2010.01434.x

Tabachnick, B. G. & Fidell, L. S. (2001) Using multivariate statistics (4th edn.). Allyn & Bacon.

Tan, Ş. (1999). Psikolojik Test Geliştirmede Faktör Analizinin Kullanımı. *Çağdaş Eğitim*, *255*, 32-38.

Tan, Ş. (2009). Misuses of KR-20 and Cronbach's alpha reliability coefficients. *TED Education and Science, 34*(152), 101-112.

Tan, Ş. (2015). *Uygulamalı temel istatistik-1*. Ankara: Pegem Akademi.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: understanding concepts and applications*. American Psychological Association.

Turgut, M. F. & Baykul, Y. (1992). *Ölçekleme teknikleri*. Ankara: ÖSYM Yayınları.

Ural, A. & Kılıç, İ. (2005). *Bilimsel araştırma süreci ve SPSS ile veri analizi*. Ankara: Detay Yayıncılık.

Wansom, S., Mason, T. O., Hersam, M. C., Drane, D., Light, G., Cormia, R., Stevens, S., Bodner, G. M. (2009). A Rubric for Post-Secondary Degree Programs in Nanoscience and Nanotechnology. *International Journal of Engineering Education, 25*(3), 615-627.

Winkelmann, K. & Bhushan, B. (Eds.) (2016). *Global perspectives of nanoscience and engineering education*. Science Policy Reports. Springer. https://link.springer.com/book/10.1007%2F978-3-319-31833-2

Young, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9*(2), 79-94.

Zwick, W.R. & Velicer, W. F. (1986). Factor Influencing Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin, 99*, 432-442.

6. APPENDIX

Nanotechnology Awareness Instrument (NAI, Turkish version)

Sayın Katılımcı:

Aşağıdaki farkındalık ölçeğinde Nanobilim ve Nanoteknoloji (NBT) ile ilgili ifadeler bulunmaktadır. Her ifadenin karşısında beş (5) cevap seçeneği vardır. Her cümleyi okuduktan sonra cümledeki ifadeye ne düzeyde katılıyorsanız, o cevap seçeneğini (X) işaretleyiniz. Cevap seçenekleri arasında doğru ya da yanlış cevap bulunmamaktadır. Lütfen hiçbir ifadeyi boş bırakmayınız. Katkılarınız için teşekkürler.

Farkındalık ölçeğini cevaplamadan önce aşağıdaki kişisel bilgilerinizi doldurmanızı rica ederiz.

1. Cinsiyetiniz: E ( ), K ( ).	6. Görev Yaptığınız Okul Türü: Fen Lisesi ( ) Anadolu Lisesi ( ) Meslek Lisesi ( )
2. Meslekteki Kıdem Yılınız: 1-5 yıl: ( ) 6-10 yıl: ( ) 11-15 yıl: ( ) 16-20 yıl: ( ) 25 yıl ve üzeri: ( )	7. Nanobilim ve Nanoteknoloji ile ilgili bir hizmet içi eğitim aldınız mı? Evet ( ) Hayır ( )
3. Branşınız: Fizik ( ), Kimya ( ), Biyoloji ( ).	
4. Mezun Olduğunuz Okul: Yüksekokul ( ) Fakülte ( ) Eğitim Enstitüsü ( ) Yüksek Öğretmen Okulu ( ) Diğer ( ) ……………………………….	8. Takip ettiğiniz bir bilimsel yayın (“Bilim ve Teknik Dergisi” vb.) var mı? Evet ( ) Hayır ( )
5. Öğrenim Durumu: Ön Lisans ( ) Lisans ( ) Yüksek Lisans ( ) Doktora ( )	9. Bilimsel alanda belgesel yayını veya programı takip etme sıklığınız nedir? Her zaman ( ) Çok sık ( ) Ara sıra ( ) Nadiren ( ) Hiçbir zaman ( )
6. Görev Yaptığınız Okul Türü: Fen Lisesi ( ) Anadolu Lisesi ( ) Meslek Lisesi ( )	10. Görev Yaptığınız Şehir: Antalya ( ) Denizli ( ) Burdur ( ) Ankara ( )

A. Aşağıdaki ölçeği kullanarak, ölçekte yer alan ifadelere katılım düzeyinizi her bir madde için lütfen belirtiniz.

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

B. Aşağıdaki ölçeği kullanarak, ölçekte ifade edilen faaliyetlere katılım düzeyinizi her bir madde için lütfen belirtiniz.

Nanoteknoloji deneyiminiz (ile etkileşiminiz) nedir?	Hiçbir zaman	Nadiren	Ara sıra	Çok sık	Her zaman
Nanoteknoloji terimini duydum.					
Nanoteknoloji hakkında bir şeyler okudum.					
Nanoteknoloji hakkında bir program izledim.					
Sınıfta bir (veya daha fazla) öğretmen/öğretim elemanının nanoteknoloji hakkındaki konuşmalarını dinledim.					
Nanoteknoloji konusunun işlendiği bir etkinliğe katıldım (laboratuvar çalışması, proje, seminer, konferans).					
Nanoteknoloji hakkında bir ders aldım.					

C. Aşağıdaki sorular Nanobilim ve Nanoteknoloji hakkında bilgi düzeyinizi belirlemek amacıyla hazırlanmıştır. Nanobilim ve Nanoteknoloji konusu son yıllarda oldukça güncel olmasına rağmen bu kavramlar ve uygulama alanları olarak oldukça yeni olduğundan soruları eksiksiz ve doğru olarak yanıtlamanız beklenmemektedir. Amaç sizleri test etmek ve değerlendirmek değildir. Amacımız sizlerin Nanoteknoloji hakkında varsa bilgi düzeyinizi belirlemektir. Araştırmada sizlerin kimliklerini belirleyecek ifadelere ve sorulara yer verilmemiştir. Katılımınız ve sorulara verdiğiniz samimi yanıtlardan dolayı teşekkür ederiz.

1. Nanometre, metrenin ……………………… birine denk gelir.
2. Nanometre boyutunda nesnelere örnek olarak …………………………………… verilebilir.
3. Nanoölçek boyutunda nesnelerin ölçülmesinde kullanılan araçlardan biri ……………………………………………………dir.
4. Nanoteknolojinin birçok uygulama alanı bulunmaktadır. Ayrıca farklı alanlarda kullanılan Nanoteknoloji yöntemleri ile geliştirilen malzemeler ve araçlar bulunmaktadır. Aşağıya bildiğiniz Nanoteknolojinin uygulama alanları ile bu alanda geliştirilen malzeme veya araç adı yazabilir misiniz?

   Uygulama alanı — Geliştirilen malzeme veya araç

   a. …………………………………… ……………………………….....
   b. …………………………………… ……………………………….....
   c. …………………………………… ……………………………….....
   d. …………………………………… ……………………………….....
   e. …………………………………… ……………………………….....
5. Nanoteknoloji uygulamaları tarafından geliştirilerek gelecekte insanları doğrudan veya dolaylı olarak etkileyecek bir malzeme veya araca örnek verir misiniz?

…………………………………………………………………………………………………
…………………………………………………………………………………………………
…………………………………………………………………………

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 690–714

https://doi.org/10.21449/ijate.744640

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The Assessment of the Fifth-Grade Students' Science Critical Thinking Skills through Design-Based STEM Education

Ayse Savran Gencer [1,*], **Hilmi Dogan** [2]

[1]Department of Science education, Faculty of Education, Pamukkale University, Denizli, Turkey.
[2]Ministry of National Education, Antalya, Turkey.

ARTICLE HISTORY

Received: May 28, 2020
Revised: Sept. 10, 2020
Accepted: Oct. 23, 2020

KEYWORDS

STEM education,
Science critical thinking,
Design-based,
Living things,
Force and friction

Abstract: Critical thinking has been one of the 21st-century skills consistently associated with students' future career advancement as a positive student outcome of STEM education. The aim of the study is to develop and validate science critical thinking skill instruments to assess the improvement in the subject of living organisms and force and friction through design-based STEM education. In this design-based research study, the student's modules were developed by the integrated STEM education principles involving the activities and worksheets in line with the frame of critical thinking approach. The kappa statistics for content validity, exploratory and confirmatory factor analysis for construct validity, and item and reliability analysis for the quality of items were used in the development stage of instruments. The results of these analyses endorsed the 15 two-tier item for each test of Living Things Critical Thinking (LTCT) and Measuring Force and Friction Critical Thinking (MFFCT) as unidimensional constructs to produce valid and reliable data to measure the fifth grade students' critical thinking skills in the related science content. Comparing the pre and post applications of instruments in the study group indicated that STEM modules improved the students' science critical thinking skills such as interpretation, analysis, and inference. In this respect, developing and validating instruments to assess the integrated critical thinking skills will contribute to the empirical examination of this construct within the context of school science learning.

1. INTRODUCTION

Rapid changes in the flow of knowledge in today's world have led to nations to revise science education programs and science teaching goals in such a way of cultivating individuals who are able to produce knowledge and use it functionally in their lives by contributing to society and culture with the skills of problem-solving, critical thinking, entrepreneurship, decision making, collaboration, communication, and empathy (e.g., Ministry of National Education [MoNE], 2018). To accomplish these goals, the initiatives of integrating science, technology, engineering, and mathematics (STEM) have been appropriated as interdisciplinary approach by involving learning about knowledge, skills, beliefs and values from more than one STEM discipline through the collaborative efforts of students and teachers (Baharin, Kamarudin, & Manaf, 2018; Çorlu, Capraro, & Capraro, 2014; Ergün & Külekci, 2019; Öner et al., 2014;

CONTACT: Ayşe Savran Gencer ✉ asavran@pau.edu.tr 🖃 Department of Science education, Faculty of Education, Pamukkale University, Denizli, Turkey.

ISSN-e: 2148-7456 /© IJATE 2020

Wang, Moore, Roehning, & Park, 2011). In particular, STEM teaching can be more meaningful when embedded in real-life problems with challenges in a manner of integrity for extending students' motivation and persistence to learn and succeed in science (Honey, Pearson, & Schweingruber, 2014).

Critical thinking has been one of the 21st-century skills consistently associated with students' future career advancement as a positive student outcome of STEM education (Next Generation Science Standards [NGSS] Lead States, 2013). A great deal of literature from many countries provide insight on how design based STEM learning activities engage students to solve real-world problems through investigating and collaborating with their peers in establishing an effective learning environment to foster critical thinking skills (Baharin et al., 2018; Duran & Şendağ, 2012; Mutakinati, Anwari, & Yoshisuke, 2018; Oonsim & Chanprasert, 2017; Rahmawati, Ridwan, Hadinugrahaningsih, & Soeprijanto, 2019; Waddell, 2019). For instance, a study with Japanese middle school students by using STEM education through project-based learning to solve the need for clean water in the future reported that students' overall critical thinking skills developed up to the category of the average thinker (Mutakinati et al., 2018). In the Indonesia context, Rahmawati et al. (2019) explored that integrating STEAM approach into chemistry learning within real-life problems provided opportunities for students to improve their critical thinking skills. In Thailand, Oonsim & Chanprasert (2017) indicated an average increase of critical thinking skills by using STEM education in the subject of physics for secondary school students. For the United States, Duran & Şendağ (2012) reported a significant effect of STEM experiences enhanced with information technology on the improvement of urban high school students' critical thinking.

1.1. The Theoretical Framework

Critical thinking is the process of mentally acting on something by "making reasoned judgments" (Beyer, 1995, p.8). Facione (1990) defines critical thinking as "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based" (p. 2). Thinking skills in education settings generally involve activities of comparing and contrasting, classifying, predicting, generating original ideas, cause and effect, decision making, uncovering assumptions, and determining the reliability of sources of information (Swartz, Costa, Beyer, Reagan, & Kallick, 2008). Critical thinking includes process skills of "analysing, evaluating, or synthesizing relevant information to form an argument or reach a conclusion supported with evidence" (Reynders, Lantz, Ruder, Stanford, & Cole, 2020, p.4). Critical thinking enables individuals to develop their way of thinking about any subject, content, or problem by skilfully handling thought-specific structures and assigning intellectual standards to them. Then, these individuals can use the principles that help them to improve their thinking while analysing and evaluating the problems or their thoughts (Gencer & Boran, 2017). In addition, critical thinking skills are required in the process of analysing possible solutions during the problem solving, evaluating the consistency between alternatives during decision making or predicting the results of the decision (Dilekli, 2019).

Beyond its wide range of definitions, there has been a dichotomy in the construction of critical thinking as domain-general versus domain-specific (Ennis 1989, Facione 1990; National Research Council [NRC], 2011; Swartz et al., 2008; Willingham, 2008). According to the report by NRC (2011), the predominant view on domain-specific construct advocates that critical thinking coevolves with the increasing content knowledge and cannot be transferred spontaneously from one subject matter to another. Willingham (2008) ascertains specific types of critical thinking to the extent to which they are characterized by different subject matters. Bailin (2002) points to the contextual nature of critical thinking in science education due to the fact that focusing on the concepts, tasks, problems, and issues in the science curriculum initiate critical thinking by collecting knowledge from observation, classification, correlation, causation, hypothesis, inference and prediction as well as background knowledge of students for critical analysis, interpretation, and evaluation. Students cannot learn spontaneously how to think in each subject matter and therefore the ways should be modelled with the characteristic of the different subject matter by giving opportunities to practice in the context of the related classroom tasks (Willingham, 2008).

Such specific types of skilful thinking and mental behaviours need to be taught students explicitly and by direct instruction to be effective thinkers (Swartz et al., 2008). Regarding the instruction of critical thinking skills, Ennis (1989) classifies approaches into four types. The general approach involves teaching critical thinking skills in a separate course without a specific subject matter. According to the infusion approach, students are involved in the explicit teaching of critical thinking skills process in a specific subject matter. In the immersion approach, a subject course is organised to teach critical thinking, but critical thinking principles are not given explicitly. The mixed model approach combines a general approach with infusion or immersion approach. In this research, we have adopted the infusion approach to teach critical thinking skills in science-domain. The student's modules for this study were designed by the integrated STEM education principles involving activities and worksheets in reference to Swartz et al.'s (2008) frame of critical thinking approach.

1.2. The Significance and Purpose of the Study

A limited research has been recently conducted to develop and assess critical thinking skills of students in the specific content knowledge such as mathematics (Harjo, Kartowagiran, & Mahmudi, 2019; Kuş & Çakıroğlu, 2020), chemistry (Reynders et al., 2020; Sadhu & Laksono, 2018), physics (Asysyifa, Jumadi, Wilujeng, & Kuswanto, 2019; Mabruroh & Suhandi, 2017), and science (Mapeala & Siew, 2015; Sya'bandari, Firman, & Rusyati, 2017). At primary level, such a study by Mapeala and Siew (2015) developed a science critical thinking test to measure the critical thinking skills of the fifth-grade students in the theme of physical sciences. At secondary school level, Sya'bandari et al. (2017) constructed a science virtual test to measure the seventh-grade students' critical thinking in the matter and heat topic. For high school students, there were integrated assessment instruments to measure critical thinking skills in the concepts of chemical equilibrium (Sadhu & Laksono, 2018) and sound waves (Mabruroh & Suhandi, 2017). In reaction to the shortage of subject-specific construct of critical thinking skills in STEM fields, this study will contribute the broadening the scope of science subjects to be taught for integrated critical thinking skills.

Another issue associated with teaching critical thinking skills is the importance of developing students' critical thinking skills at an early age. In essence, critical thinking skills should be embedded in the science curriculum from beginning in the early grades of schooling and growing in complexity and sophistication throughout the grades (Bailin, 2002; Wicaksana, Widoretno, & Dwiastuti, 2020). In doing so, the current study can contribute to the development and assessment of the early grade students' science critical thinking skills in informing science educators and teachers about how to design an effective learning environment to teach science critical thinking skills in their classroom. Due to the fact that the importance of critical thinking skill has been appreciated as one of the higher-order thinking skills to be assessed in international exams (e.g., Programme for International Student Assessment [PISA]) and national exams in Turkey (e.g., High School Pass Exam), further investigations need to be done to construct and measure more accurately integrated critical thinking skills in science learning. In an effort to attain these goals, the present study aims to develop and validate science critical thinking skill instruments to assess the improvement in the subject of living organisms and force through design-based STEM education.

2. METHOD

The current research is a part of a larger dissertation study based on design-based research consisting of the preliminary research phase, the prototyping phase (the iterative design phase), and the assessment/reflective phase as proposed by Abdallah & Wegerif (2014). Based on the preliminary phase, STEM modules were developed to provide students the learning opportunities to explore both engineering design principles and learning outcomes in the units of "Living Things World" and "Measuring Force and Friction". Worksheets in the related STEM modules were constructed in line with the critical thinking frame of Swartz et al. (2008) including analysing ideas in terms of comparing and contrasting, classifying, sequencing, determining parts/whole relationship, identifying causal relationship and further analysing arguments, drawing a conclusion, making a decision and problem solving to integrate learning tasks with critical thinking skills. Appendix I indicates examples of worksheets to integrate specified critical thinking skills with learning tasks in the modules.

The iteration cycles include the eight-step engineering design process of the Massachusetts Department of Education (2006). The engineering design cycle consists of identifying the need or problem, researching the need or problem, developing a possible solution, selecting the best possible solution, constructing a prototype, testing and evaluating the solution, communicating the solution, and redesigning. The study was conducted for sixty-one hours during the science lessons in the first term of the school year of 2018-2019 by the second author of this research.

The students' learning modules which were designed with STEM education approach applied to the study group as an intervention. The assessment/reflective phase involved an assessment of the instruments as a pre- and post-test to collect data about the effectiveness of the STEM modules on the student's integrated critical thinking skills.

2.1. Study Group

In the first stage of the preparation phase, the sample for the pilot study was needed for developing the science domain instruments. The data were obtained from the sixth grade students of (N = 147) for Living Things Critical Thinking (LTCT)-Test and (N = 116) for Measuring Force and Friction Critical Thinking (MFFCT)-Test studying in three different public secondary schools located in Antalya.

In the second phase of the prototyping phase, the study group was chosen with a convenience sampling method (Patton, 2014) by considering the school where the second researcher worked as a teacher. The study was carried out with 22 students (10 girls, 12 boys) who were 10-11 years old attending the fifth grade at the public secondary school in Antalya, Turkey.

2.2. The Instruments

The instruments were developed in order to evaluate the integrated critical thinking skills within school science learning through design-based STEM education. LTCT and MFFCT tests focused on the three elements of the critical thinking skills including interpretation, analysis, and inference as proposed by Facione (1990) within the contents of the current science education curriculum in Turkey (MoNE, 2018). Each of the final version of instruments consists of 15 two-tier multiple-choice items. The first-tier item includes content questions with four choices and the second tier includes a blank for the first part to allow students to explain the reason why they choose the option in the first tier (Griffard & Wandersee, 2001). The items of open-ended two-tier multiple-choice tests were scored as 3 (the right answer -the right reason), 2 (the right answer- partly correct reason), 1 (the right answer- the wrong reason), 2 (the wrong answer-the right reason), 1 (the wrong answer- partly correct reason), and 0 (the false answer-the wrong reason). In this study, one point is given for the students who can write a partially correct reason despite their wrong answers in addition to the commonly used scoring in the literature. A guideline was prepared for students and practitioners regarding the duration of the test, scoring method and what they are expected to explain in the second tier.

2.2.1. *The Instrument Development Process*

In the initial versions, 19 two-tier items for LTCT-Test and 20 two-tier items for MFFCT-Test with four options were developed. Some of the cognitive critical thinking skills and sub-skills compiled by Facione's (1990) Delphi report and learning objectives in the science curriculum (MoNE, 2018) in Turkey were taken into consideration as a guide for the development of the science-domain critical thinking tests. In the unit of Living Things World, students are expected to give examples of living things and classify them according to their similarities and differences as microscopic organisms, fungi, plants, and animals. In the unit of Measuring Force and Friction, students are expected to measure the magnitude of the force with a dynamometer, give examples of friction force from daily life, discover the effect of friction force on motion in various environments, do experiments about the effect of friction force on motion on rough and slippery surfaces, and generate new ideas to increase or decrease friction in everyday life (MoNE, 2018).

LTCT-Test contains critical thinking constructs of interpretation (4 items), analysis (4 items), and inference (7 items). Table 1 indicates the core and sub-skills of the critical thinking constructs within the science content for LTCT-Test. MFFCT-Test contains critical thinking constructs of interpretation (5 items), analysis (3 items), and inference (7 items). Table 2 indicates the core and sub-skills of the critical thinking constructs within the science content for MFFCT-Test. A sample item was given in Figure 1 and Figure 2 for LTCT-Test and MFFCT-Test, respectively.

Table 1. *The integrated critical thinking skills and the science content for LTCT-Test*

Item	Core Skills of Critical Thinking	Sub-skills	Scientific content
1-3-7-10	Interpretation	Categorization	Identifying the distinctive and/or common features of living things from image/diagram/text to classify them.
2	Analysis	Examining ideas	Comparing and contrasting living things by classifying them in terms of criteria and revealing the relations using the data presented in the graph.
8		Analysing arguments	Recognizing the difficulties in classifying living things and distinguishing the rejecting or supporting reasons for the claims regarding the classification.
11-14			Distinguishing the rationale for rejecting or supporting the claim regarding the classification of living things.
4-12-15	Inference	Drawing conclusion	Drawing a conclusion about the function of the structure by observing the structure of living things.
5			Drawing a conclusion that scientific knowledge is tentative by using relevant information/data
6			Drawing a conclusion about how scientific knowledge is formed and the process through which knowledge is passed.
13		Querying evidence	Deciding the accuracy of classification of living things based on evidence.
9		Conjecturing alternatives	Identifying the hypothesis tested by obtaining the variables that affect the growth of bacteria from a given experimental setup.

Item 11. **Core skill: Analysis** **Sub-skill: Analysing arguments**

Melek and Cemre, who have visited the zoo, come to the section with seals. Melek carefully examines this creature because she has seen those animals for the first time and says to Cemre:

Melek: It can breathe on land. I think this is a mammal.
Cemre: Frogs can breathe on the land too, but they are not a mammal. It has got fins and swims like a fish. I think that it must be a fish.
Melek: But it has not got scales. Instead of them, it seems to have short and stiff hairs

Melek and Cemre cannot decide whether the seal is a mammal or fish. They then decide to ask a zoologist working at the zoo. After the zoologist gives them information, Melek and Cemre are convinced the seal is a mammal.
According to the text, what might the zoologists have told Melek and Cemre?
A. Seals feed on other fishes
B. Seals have skeletons
C. Seals feed their offsprings with milk
D. Seals do not have gills
How did you decide that the option you marked was correct? Please explain.
………

Figure 1. *A sample item for an open-ended two-tier multiple-choice question in LTCT-Test*

Item 14. **Core skill: Inference** **Sub-skill: Drawing conclusion**

Predator birds can see their prey while flying high thanks to their sharp eyes. While they are searching for prey, they spread their wings as wide as they can, and they do not have to flap them for a long time. As soon as they see their prey, they close their wings slightly and dive to catch the prey.

In the images below, Figure I shows a predator bird searching its prey, and Figure II shows the bird diving to catch its prey.

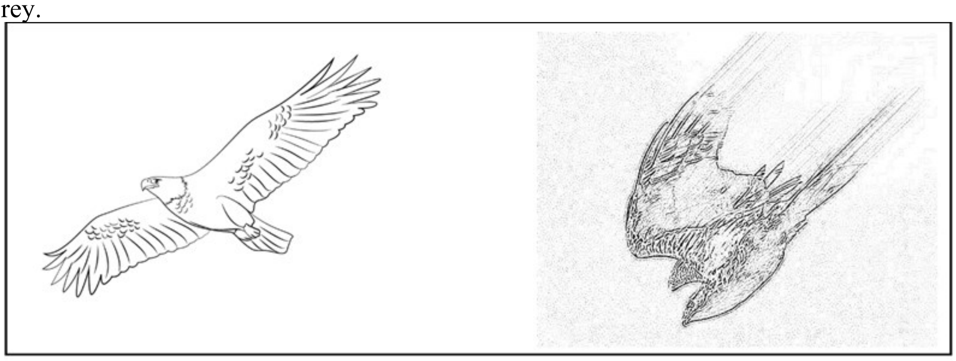


Figure I Figure II

Which of the following can be reached for the two given conditions accordingly?
A. The air resistance affecting the predator bird is greater when its wings are wide open.
B. The bodyweight of the predator bird decreases while it flies with open wings.
C. The gravity affecting the predator bird increases when the bird closes its wings.
D. The air resistance exerted on the predator bird increases when the bird closes its wings.

How did you decide that the option you marked was correct? Please explain.

Figure 2. *A sample item for an open-ended two-tier multiple-choice question in MFFCT-Test*

Table 2. *The integrated critical thinking skills and the science content for MFFCT-Test*

Item	Core Skills of Critical Thinking	Sub-skills	Scientific content
2	Interpretation	Clarifying meaning	Obtaining the magnitude of the force acting on the objects from the given images, display the data with a graph.
12		Categorization	Explaining by obtaining the magnitude of the force acting on the objects from the given graph.
9-13			Classify the applications that increase or decrease the friction force in daily life according to their similar and different features.
5	Analysis	Analysing arguments	Distinguishing the justifications for rejecting or supporting the argument regarding the relation between air friction with the surface area.
8-15			Distinguishing the rationale for rejecting or supporting the argument about the effects of friction force in daily life.
10			Determination the interrelation between the supplied parts and each other in an experiment on the relation between the spring thickness and the sum of its extension
1	Inference	Conjecturing alternatives	Obtaining the magnitude of the forces from the given evidence, making inference by comparing the data.
3		Drawing conclusion	Drawing a conclusion by using empirical data confirming or falsifying the claims regarding the measurement of the magnitude of the force.
4			Obtaining the data about the magnitude of the forces acting on the objects from the visuals, comparing weights and determinate the relationship between the magnitudes.
6			Drawing a conclusion about the tested hypothesis from the result of the experiment that friction force depends on the type of surface/surface area.
7			Identifying and distinguish the causes that help to support the outcome of friction-induced events in everyday life.
11			Drawing a conclusion by using empirical data confirming or falsifying the claims regarding the factors affecting air friction.
14			Identify and distinguish the causes that help support the outcome of friction-induced events in everyday life.

2. 3. Data Analysis

While the instruments were being developed, validity, reliability and item analysis were conducted in order to obtain information for each item whether to use, revise or eliminate the faulty items (Whiston, 2012). For content validation, the written items were examined by a panel of experts (n = 6) consisting of two science teachers, one academician in the field of science education, two academicians in curriculum specialized in critical thinking and one academician in the field of measurement and evaluation. Kappa statistics were used to assess the opinions of experts on the items in terms of the relevance to the content, construct, grade level, and clarity. The kappa statistics were also used to assess the coding of the open-ended parts of two-tier questions.

For construct validity, FACTORv.10.10.01 software was used to determine dimensionality and structure testing with regard to Exploratory Factor Analysis (EFA) carried with optimal implementation of Parallel Analysis (PA) (Timmerman, & Lorenzo-Seva, 2011) based on Polychoric Correlations Matrix (PCM). In order to confirm the unidimensionality of the data, the Confirmatory Factor Analysis (CFA) was applied by using LISRELv.8.80 software.

The obtained data were analysed by TAP 19.1.4 software to carry out item statistics for the first-tier items scored dichotomously (0 and 1) of both LTCT-Test and MFFCT-Test. In this research, the difficulty index (p value) and item discrimination point biserial coefficient (r_{pb}) values were calculated. For the reliability assessment to examine the internal consistency of an instrument, both Kuder–Richardson formula known as *KR-20* was used to calculate reliability for the first tier of the test items scored as dichotomous (0 and 1) and Cronbach's alpha coefficient was used to calculate for the integrated assessment of items with open-ended second tier scored as polytomous (0, 1, 2, and 3).

After the tests were applied on the study group of this research as a pre- and post-test in order to determine the impact of the intervention, the collected data were analysed by Wilcoxon Test by using SPSS v.22 software programme.

3. RESULT / FINDINGS

In this section, the results of kappa statistics, item statistical analysis, exploratory and confirmatory factor analysis, and Wilcoxon test are presented.

3.1. Content Validity of the Instruments

In the preliminary version of the instruments, 19 two-tier items for LTCT-Test and 20 two-tier items for MFFCT-Test were validated by a panel of expert judges. Each item of the tests was evaluated by the experts considering a) relevance of the item with the content b) relevance of the item with the critical thinking skills c) clarity of the item d) relevance of the item with the grade level. The three attributes of the items were rated in a three-point Likert scale format (1 = not relevant; 2 = partly relevant; 3 = relevant). Also, a blank section for each item is allocated for experts to comment on each item. The modified kappa statistic was computed to estimate the agreement between the experts indicated beyond the chance on item level content reliability (Polit, Beck & Owen, 2007). The probability of chance agreement (P_c) is first computed with formula 1 and to compute modified kappa statistic (k*) inserted into formula 2.

$$PC = \left[\frac{N!}{N_G\,!\,(N - N_G\,)!}\right] \cdot \left[\frac{1}{2}\right]^N \qquad (1)$$

N: Number of experts
N_G: Number of agreements rated relevant

$$Kappa = \frac{\left(\frac{N_G}{N}\right) - P_C}{1 - P_C} \qquad (2)$$

The calculated *PC* and the k^* values for LTCT-Test and MFFCT-Test are displayed in the Table 3 and Table 4, respectively. If the kappa value is between (.60 ≤ kappa ≤ .74), the agreement among experts is good. If the kappa value is greater than .75, the agreement among experts is perfect (Fleiss, 1981, as cited in Yurdugül ve Bayrak, 2012). As regards to these criteria, the modified kappa values of items (Items 4, 9, 15, and 19) which were lower than .60 for LTCT-Test were eliminated from the test. The rest of the items all had modified kappa values which were greater than .75. Consequently, it can be interpreted that the agreement among experts was perfect for these items.

Table 3. *The kappa statistics for content validity of LTCT-Test*

Item	a. Relevance of the item with the content					b. Relevance of the item with the critical thinking skills					c. Clarity of the item					d. Relevance of the item with the grade level				
	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*
	R	PR	NR			R	PR	NR			R	PR	NR			R	PR	NR		
1	6	0	0	0.02	1.00	6	0	0	0.02	1.00	5	1	0	0.09	0.82	6	0	0	0.02	1.00
2	5	1	0	009	0.82	5	1	0	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
3	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
4	4	0	2	0.23	0.56	3	1	2	0.31	0.27	6	0	0	0.02	1.00	5	1	0	0.09	0.82
5	6	1	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
6	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
7	6	0	0	0.02	1.00	5	1	0	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
8	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
9	5	1	0	0.09	0.82	3	1	2	0.31	0.27	5	1	0	0.09	0.82	6	0	0	0.02	1.00
10	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
11	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
12	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
13	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
14	3	2	1	0.31	0.27	4	1	1	0.23	0.56	6	0	0	0.02	1.00	6	0	0	0.02	1.00
15	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
16	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
17	5	1	0	0.02	1.00	5	0	1	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
18	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
19	4	0	2	0.23	0.56	4	1	1	0.23	0.56	5	1	0	0.09	0.82	5	1	0	0.09	0.82

R: relevant, PR: partly relevant NR: not relevant, $P_{c:}$ probability of chance relevant, k*: modified kappa value

Similarly, the modified kappa values of items (Item 7, 12, 13, 16, and 20) were lower than .60 for MFFCT-Test were eliminated from the test. The rest of the items had modified kappa values which were greater than .75. Consequently, it can be interpreted that the agreement among experts was perfect for these items.

Table 4. *The kappa statistics for content validity of MFFCT-Test*

Item	a. Relevance of the item with the content					b. Relevance of the item with the critical thinking skills					c. Clarity of the item					d. Relevance of the item with the grade level				
	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*	Number of expert reviews			P_c	k^*
	R	PR	NR			R	PR	NR			R	PR	NR			R	PR	NR		
1	6	0	0	0.02	1.00	5	1	0	0.09	0.82	5	1	0	0.09	0.82	6	0	0	0.02	1.00
2	5	1	0	009	0.82	5	1	0	0.09	0.82	6	0	0	0.02	1.00	5	1	0	0.09	0.82
3	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
4	6	0	0	0.02	1.00	5	1	0	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
5	6	1	0	0.02	1.00	5	1	0	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
6	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
7	3	2	1	0.31	0.27	4	1	1	0.23	0.56	6	0	0	0.02	1.00	6	0	0	0.02	1.00
8	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
9	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
10	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
11	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
12	3	2	1	0.31	0.27	4	1	1	0.23	0.56	6	0	0	0.02	1.00	5	1	0	0.09	0.82
13	4	2	0	0.23	0.56	3	1	2	0.31	0.27	5	1	0	0.09	0.82	5	1	0	0.09	0.82
14	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
15	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
16	3	2	1	0.31	0.27	4	1	1	0.23	0.56	6	0	0	0.02	1.00	6	0	0	0.02	1.00
17	5	1	0	0.02	1.00	5	0	1	0.09	0.82	6	0	0	0.02	1.00	6	0	0	0.02	1.00
18	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
19	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00	6	0	0	0.02	1.00
20	4	0	2	0.23	0.56	3	1	2	0.31	0.27	6	0	0	0.02	1.00	5	1	0	0.09	0.82

R: relevant, **PR**: partly relevant **NR**: not relevant, $\boldsymbol{P_{c:}}$ probability of chance relevant, $\boldsymbol{k^*}$**:** modified kappa value

In addition, kappa statistics were calculated for the answers of the open-ended parts of two-tier questions. For this purpose, the second author of the study scored the students’ answers at two different times to provide interrater reliability for the consistency of between two scores by a single rater. The kappa values of .883 and .886 were calculated for LTCT-Test and MFFCT-Test, respectively. The level of kappa coefficient indicated that there is excellent consistency of the scores.

3.2. Item Statistical Analysis

The multiple choice first-tier items were scored as a dichotomous variable (0 and 1), and analysis was carried out with the Test Analysis Program (TAP) (Brooks & Johanson, 2003). Coaley (2010) defined that “the difficulty indicator, known as the *p* value, represents the percentage of participants who have answered an item correctly and is calculated by dividing the number of people getting it right by the total number who attempted it” (p.38). An item difficulty index can range from .00 (meaning no one got the item correct) to 1.00 (meaning everyone got the item correct). Whiston (2012) points out that “item difficulty does not really indicate difficulty; rather, because it provides the proportion of individuals who got the item correct, it shows how easy the item is” (p.71). According to the Coaley (2010), “if all is well, the mean item *p* value is about .50 indicates moderate difficulty level…But it does not mean that a mean *p* value of .50 is always appropriate because a high level assessment of cognitive ability may need more difficult items and, therefore, a lower mean value (is preferable)” (p.38).

The item discrimination analysis indicates that each item of the test is related to the overall test performance (Haladayna, 1999; Nunnally & Bernstein, 1994). The discrimination value can be decided by using the point biserial coefficient (r_{pb}) that compares correct and incorrect answers for each item statistically with overall test score performance (Polit & Hungler, 1999). If the item discrimination value is greater than ($r_{pb} \geq .40$), item is very good or perfect; between ($.30 \leq r_{pb} \leq .39$), it is reasonable good; between ($.20 \leq r_{pb} \leq .29$), it is marginal but acceptable; and lower than ($r_{pb} \leq .19$), it is weak and should not be included in the test (Crocker & Algina, 1986; Ebel & Frisbie, 1991; Wiersma & Jurs, 2005).

The result of the TAP analysis produced the item difficulty index (*p*) and point biserial coefficient (r_{pb}) value to determine the discrimination index of items which are displayed for LTCT-Test and MFFCT-Test in Figure 3 and Figure 4, respectively.

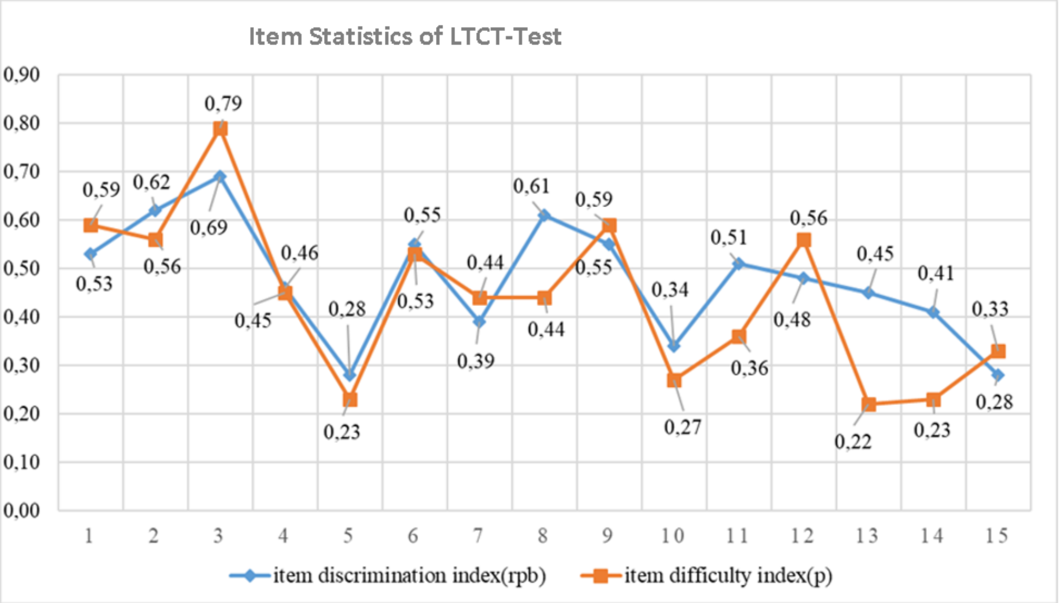


Figure 3. *The graph of item statistics of LTCT-Test*

Figure 3 indicates item difficulty (p) values ranging from .22 to .79 for LTCT-Test. The result of the analysis indicates that the test was formed with different difficulty levels of items. The average item difficulty value was calculated as .44 for LTCT-Test. Figure 3 indicating point biserial coefficients for 11 items were greater than .40 and for two items were greater than .30 that these items had perfect and good discrimination values. Item 5 and 15 had a point biserial value of .28, then these two items were decided to conserve in the test, but they should be revised as regards the criteria of ($.20 \leq r_{pb} \leq .29$). Overall, all the items with the value of ($r_{pb} \geq .20$) were determined to be included in LTCT-Test with the average discrimination value of .48.

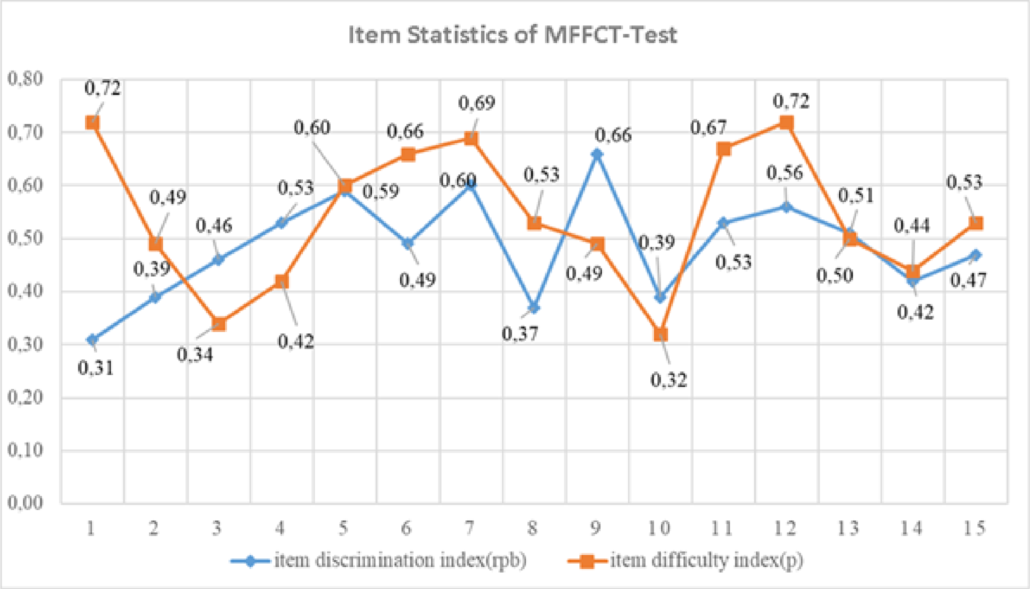


Figure 4. *The graph of item statistics of MFFCT-Test*

Figure 4 indicates item difficulty (p) values ranging from .32 to .72 for MFFCT-Test. The average item difficulty value was calculated as .54 for MFFCT-Test. Figure 4 indicating point biserial confidents for 11 items were greater than .40 and for four items were greater than .30 that these items had perfect and good discrimination values. Overall, all the items with the value of ($r_{pb} \geq .20$) were appreciated to stay in MFFCT-Test with the average discrimination value of 0.49.

Reliability analyses of the instruments were tested using both *KR-20* formula and Cronbach's alpha coefficient. The value of reliability coefficient with greater than .70 indicates the test is reliable (Frankel & Wallen, 2008). As regards to this criterion, reliability analysis of *KR-20* for the first tier of the tests produced the acceptable value of .76 and .77 for LTCT-Test and MFFCT-Test, respectively. Also, reliability analysis of Cronbach's alpha coefficient for entire tests as polytomous produced the value of .79 and .91 for LTCT-Test and MFFCT-Test, respectively.

3. 3. Exploratory Factor Analysis of Instruments

Exploratory factor analysis with parallel analysis (PA) based on the polychoric correlations matrix (PCM) was carried out independently for both tests to determine the number of dimensions. Unidimensionality of the tests were determined by the values of Unidimensional Congruence (UniCo > .95), Explained Common Variance (ECV> .85), and Mean of Item Residual Absolute Loadings (MIREAL< .30) for the overall the test as well as for each item (Ferrando & Lorenzo-Seva, 2017). Another evidence for unidimensionality considered in studies is 20% or more explained variance by the first factor with 4 to 5 times greater eigenvalue when compared to the second factor's eigenvalue (Arıcak, Avcu, Topçu, & Tutlu, 2020; Deng, Wells, & Hambleton, 2008; Hattie, 1985; Yalaki et al., 2019). Prior to factor analysis, Kaiser-Meyer-Olkin (KMO) with the critical value greater than .60 and significant result of Barlett Sphericity test ($p < .05$) were ensured for the convenience of data (Bursal, 2017).

The computed KMO value of .774 with greater than the critical value of .60 and the significant result of Barlett's test of sphericity ($\chi 2(105) = 603.9$, $p = .000010$) for LTCT-Test indicated that the data were appropriate for factor analysis. The result of the parallel analysis at 95% confidence intervals indicated that the values of UniCo = .925 (.918 ≤ UniCo ≤ .948), ECV = .769 (.754 ≤ ECV ≤ .821) and MIREAL= .224 (.199 ≤ MIREAL ≤ .220) provided evidence for the unidimensionality of the test. The parallel analysis suggested a one-factor structure that explained 30.5% of the variance of the test scores with eigenvalue of 4.57. This eigenvalue of the first factor was approximately 3 times the eigenvalue of the second factor that also confirmed the one-factor structure. Büyüköztürk (2012) suggested the cut-off point for factor loadings to be at least .30. The LTCT-Test included 14 items with loadings ranging from .357 to .727 and one item loadings with .258 retained in the test because of the content contribution. As a result of the exploratory factor analysis, the data can be threatened as essentially unidimensional for the LTCT-Test.

The computed KMO value of .834 with greater than the critical value of .60 and significant result of Barlett's test of sphericity ($\chi 2(105) = 999.1$, $p = .000010$) for MFFCT-Test indicated that the data were appropriate for factor analysis. The result of the parallel analysis at 95% confidence intervals indicated that the values of UniCo = .978 (.963 ≤ UniCo ≤ .991), ECV = .884 (.860 ≤ ECV ≤ .930) and MIREAL = .213 (.157 ≤ UniCo ≤ .258) provided the unidimensionality for the test. The parallel analysis suggested one-factor that explained 49.2% of the variance of the test scores with eigenvalue of 7.37. This eigenvalue of the first factor was approximately 6 times the eigenvalue of the second factor that also confirmed the one-factor structure. The MFFCT-Test included 15 items with loadings ranging from .561 to .823. As a result of the exploratory factor analysis, the data can be threatened as essentially unidimensional for MFFCT-Test.

3. 4. Model-Data Fit Analysis of Instruments

Confirmatory factor analysis (CFA) was used to determine whether the existing structure of the instruments confirms the one-factor model (Doğan, 2019; Whiston, 2012). The CFA tests the theory rather than producing a theory (Stevens, 2002). The path analysis was used to confirm the structure of the test as a technique of the Structure Equation Model (SEM) (Awang, 2012; Hair, Black, Babin, & Anderson, 2009). Figure 5 and Figure 6 present the results of the standardized one-factor model solution for the LTCT-Test and MFFCT-Test, respectively.

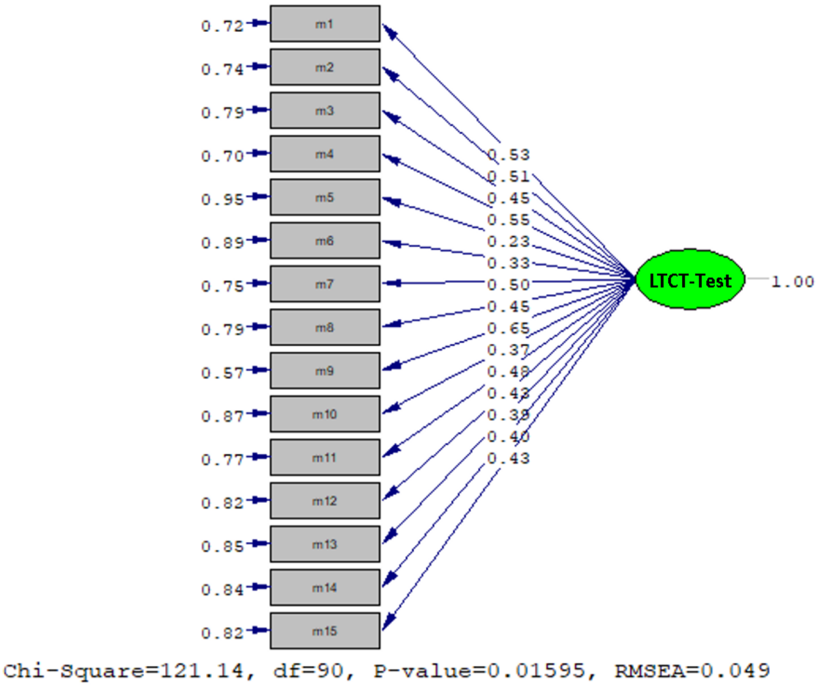


Figure 5. *CFA Diagram for standardized one-factor model solution of LTCT-Test*

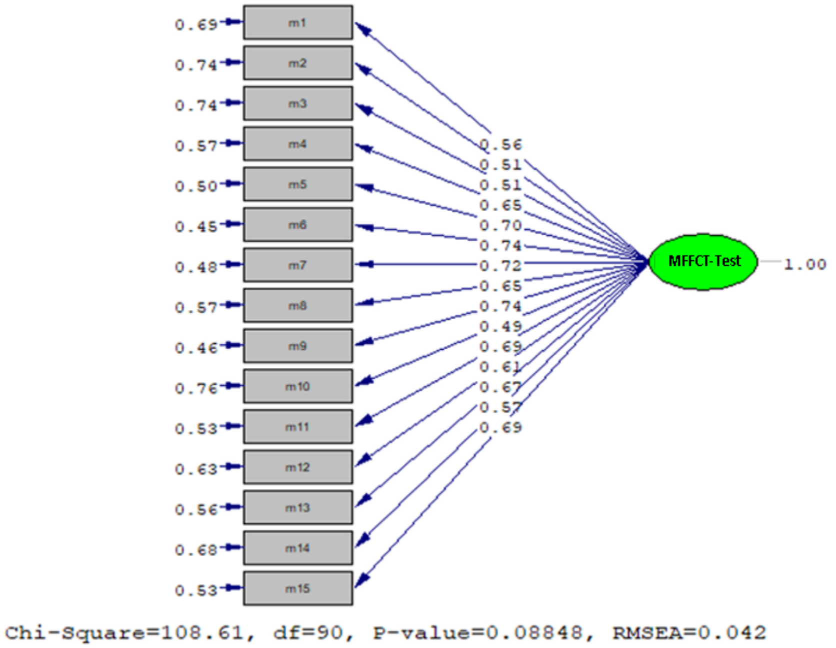


Figure 6. *CFA diagram for standardized one-factor model solution of MFFCT-Test*

The model fit indices are presented in Table 5 according to the results of CFA analysis. The critical value for χ2statistic is considered with degrees of freedom because of its sensibility to sample size. The value of χ2/df < 3 indicates the perfect fit (Kline, 2015; Tabachnick & Fidell, 2013). The Root Mean Square Error of Approximation value of (RMSE ≤ .08) and Standardized Root Mean Square Residual (SRMR ≤ .08) are considered as an acceptable fit indicator (Brown, 2015; Hair et al., 2009). According to Kline, approximation of the goodness of fit index values of Comparative Fit Index (CFI ≥ .95) and Non-Normed Fit Index (NNFI ≥ .95) indicate a good fit. As regards to these criteria, it can be interpreted that both LTCT-Test and MFFCT-Test indicated a very good fit with the one-factor model.

Table 5. The model-fit statistics for LTCT-Test and MFFCT-Test

Test	df	χ2	χ2/df	RMSEA	CFI	NNFI/TLI	SRMR
LTCT-Test	90	121.14	1.346	.049	.96	.95	.066
MFFCT-Test	90	108.61	1.206	.042	.99	.98	.051

3.5. Pre- and Post-test Comparisons of the Instruments

The mean values of the tests were evaluated by considering the total scores of the test and the three elements of the critical thinking skills including interpretation, analysis, and inference scores. The LTCT-Test and MFFCT-Test scores of the students are presented in Figure 7 and Figure 8. As seen in Figure 7, the mean values of the pre-test and post-test of LTCT-Test were 12.00 and 23.22 with the standard deviation of 6.65 and 9.83, respectively. As seen in Figure 8, the mean values of the pre-test and post-test of MFFCT-Test were 13.09 and 23.09 with the standard deviation of 7.98 and 10.02, respectively.

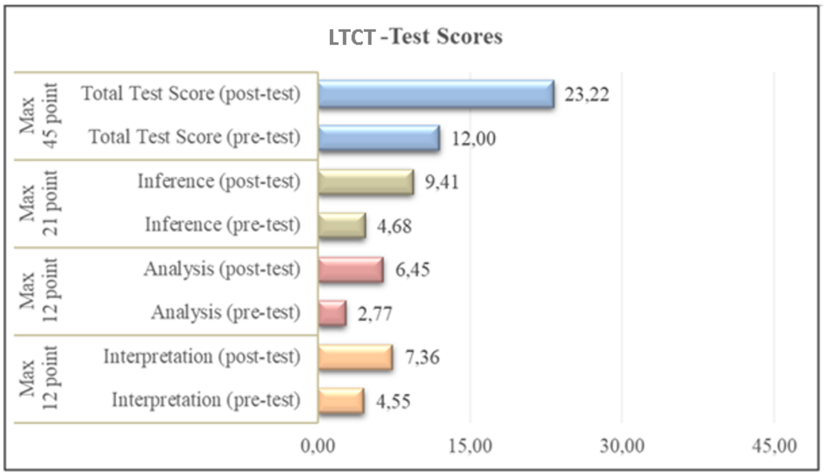


Figure 7. *Mean values for LTCT-Test*

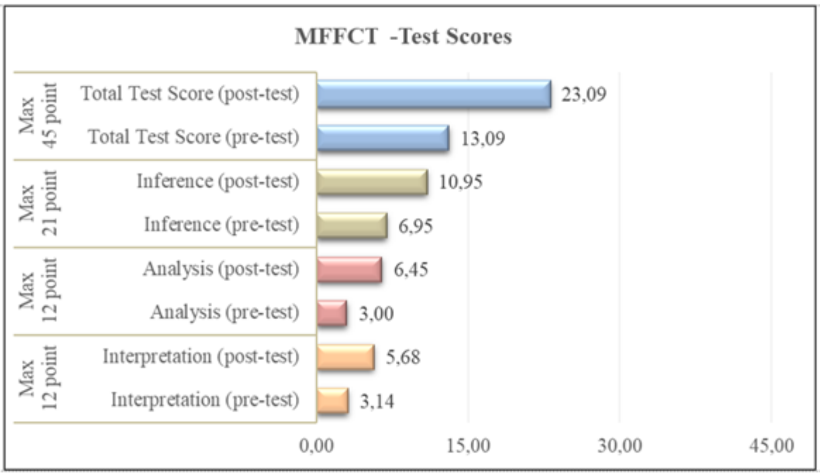


Figure 8. *Mean values for MFFCT- Test*

Then, Figure 7 and Figure 8 indicates that post-test total scores means, and the elements of the critical thinking skills of interpretation, analysis, and inference scores means for both tests are higher than pre-test scores' means. In order to determine whether there were any statistically significant mean differences in the pre- and post-test scores, the Wilcoxon test was calculated. The results for LTCT-Test and MFFCT-Test are presented in Table 6 and Table 7, respectively.

Table 6. *Wilcoxon test results of LTCT-Test*

LTCT- Test Scores	Ranks	*N*	Mean Rank	Sum of Ranks	Z	*P*
Total	Negative Ranks	1	1.00	1.00	-3.982	.000
	Positive Ranks	20	11.50	230.00		
	Ties	1				
	Total	22				
Inference	Negative Ranks	1	5.50	5.50	-3.940	.000
	Positive Ranks	21	11.79	247.50		
	Ties	0				
	Total	22				
Analysis	Negative Ranks	4	4.00	16.00	-3.598	. 000
	Positive Ranks	18	13,17	237.00		
	Ties	0				
	Total	22				
Interpretation	Negative Ranks	3	3.67	11.00	-3.271	.000
	Positive Ranks	15	10.67	160.00		
	Ties	4				
	Total	22				

Regarding the Wilcoxon test results presented in Table 6 there were significant mean differences in total test score of LTCT-Test (Z =-3.982, *p* = .00 <.05) and sub-skills of inference (Z = -3.940, *p* = .00 <.05), analysis (*Z* = -3.598, *p* = .00 <.05) and interpretation (Z =-3.271, *p* = .00 <.05) between the pre- and post-test applications. It can be concluded that the designed and implemented living things module based on STEM education approach was an effective way to develop the critical thinking skills of the participant students.

Table 7. *Wilcoxon test results of MFFCT-Test*

MFFCT-Test Scores	Ranks	*N*	Mean Rank	Sum of Ranks	Z	*P*
Total Test Score	Negative Ranks	1	1.00	0	-4.076	.000
	Positive Ranks	21	12.00	252.00		
	Ties	0				
	Total	22				
Inference	Negative Ranks	2	5.00	5.50	-3.861	.000
	Positive Ranks	21	11.79	226.50		
	Ties	0				
	Total	22				
Analysis	Negative Ranks	2	2.00	3.00	-3.818	. 000
	Positive Ranks	19	10.94	207.00		
	Ties	1				
	Total	22				
Interpretation	Negative Ranks	4	2.75	11.00	-3.392	.000
	Positive Ranks	15	11.93	179.00		
	Ties	3				
	Total	22				

Regarding the Wilcoxon test results presented in Table 7 there were significant mean differences in the total test scores of MFFCT-Test (Z = -4.076, *p* = .00 <.05) and sub-skills of inference (Z = -3.861, *p* = .00 <.05), analysis (Z = -3.818, *p* = .00 <.05) and interpretation (Z = -3.392, *p* = .00 <.05) between the pre- and post-test applications. It can be concluded that the designed and implemented force and friction module based on STEM education approach was an effective way to develop the critical thinking skills of the participant students.

4. DISCUSSION and CONCLUSION

The present study aimed to develop and validate science critical thinking skill instruments to assess the improvement in the subject of living organisms and force and friction through design-based STEM Education. In doing so, LTCT-Test and MFFCT-Test consisting of two-tier 15 multiple-choice items were developed by integrating related science content and three sub-skills of critical thinking as interpretation, analysis and inference with reference to Facione's (1990) Delphi study. In the initial phase of the instruments, 19 two-tiers items for LTCT- Test and 20 two-tier items for MFFCT-Test were written by considering the objectives of the science curriculum (MoNE, 2018) and sub-skills of critical thinking.

For content validity, both test items were evaluated by a group of experts considering the relevance of the item with the content, construct, grade level, and clarity. The modified kappa values were calculated for each item to test interrater reliability and the items with the value less than 0.60 were deleted After the minor revisions on wording in the retaining items, the 15 two-tier item LTCT-Test was applied to 147 students and the 15 two-tier item MFFCT-Test was applied to 116 students at the pilot stage. The item analysis for the first tier of dichotomous items were carried out by using item difficulty (p) and point biserial coefficient (r_{pb}) for both LTCT-Test and MFFCT-Test. The results of item analysis with TAP program pointed out that the values were in an acceptable range. Then, the tests had the average difficulty and discrimination index. In addition, reliability analysis of *KR-20* for the first tier of the tests produced the acceptable value of .759 and .767 for LTCT-Test and MFFCT-Test, respectively. Also, reliability analysis of Cronbach's alpha coefficient for entire tests as polytomous form produced the acceptable value of .789 and .908 for LTCT-Test and MFFCT-Test, respectively.

The second tier of items was evaluated by a polytomous rubric and then total scores were calculated for each item by summing the scores obtained from the first tier and second tier. Therefore, the parallel analysis was carried out based on PCM to determine the number of dimensions and the structure of tests. The results of the parallel analysis suggested extracting only one factor structure for LTCT-Test and MFFCT-Test. Further analysis to confirm the unidimensionality of the CFA was carried out. The results of the CFA confirmed the one factor structure of the tests. In other words, exploratory and confirmatory factor analysis supported the model as a one-dimensional measure for both LTCT-Test and MFFCT-Test with very good fit indices.

In conclusion, when the findings of the content and construct validity, item, and reliability analysis are considered, all items of both tests are valid to measure the critical thinking skills in the related science content as unidimensional. Much of the recent studies examining critical thinking skills have converged on the need for domain-specific teaching and assessment (Asysyifa et al., 2019; Mabruroh & Suhandi, 2017; Mapeala & Siew, 2015; Reynders et al. 2020; Sadhu & Laksono, 2018; Sya'bandari et al., 2017). In this respect, developing and validating instruments to assess the integrated critical thinking skills will contribute to the empirical examination of this construct within the context of school science learning.

The second phase of the research focused on assessing the improvement in students' integrated critical thinking skills in the subject of living organisms and force and friction through design-based STEM education. As an intervention, students participated in STEM modules enriched with critical thinking principles. To do this, the elements of critical thinking skills were reflected in the worksheets and emphasized during the implementation of the STEM modules by the teacher. Both at the beginning and at the end of the modules LTCT-Test and MFFCT-Test were conducted as pre-and post-tests. The collected data were computed with Wilcoxon test. The results were found statistically significant. In other words, participating in the STEM modules enriched with critical thinking principles improved the students' critical thinking skills such as interpretation, analysis, and inference in relation to the science content. This result is consistent with the previous literature in that increasing the critical thinking skills of students have been found positively related to STEM studies (Baharin et al., 2018; Duran & Şendağ, 2012; Mutakinat et al., 2018; Oonsim & Chanprasert, 2017; Rahmawati et al., 2019; Waddell, 2019).

As a result of this study it can be concluded that the infusing approach is an efficient way to teach critical thinking skills to students through the implementation of the science units. From this point of thought, (Willingham, 2008) suggests that thinking critically should be taught in the context of subject matter and opportunities must be given to students on their own ways to think critically. Further investigations should be required to understand instructional effectiveness and classroom dynamics to contribute designing a more effective educational environment and measure students' critical thinking skills by utilizing both qualitative and quantitative research techniques. Given the significant role of critical thinking skills in nurturing successful individuals in their daily life, teachers ought to be equipped with effective principles and strategies that enable them to sustain student engagement in critical learning activities.

Acknowledgements

This study includes a part of doctoral dissertation entitled Design, Implementation, and Evaluation of the Fifth Grade Science Course Units with an Integrated STEM Education Approach (Doğan, 2020).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). For the research, the necessary ethical approval was taken from the ethical committee and the necessary permission was obtained according to board decision, dated 05/02/2019 and 2439445 numbered from the Research Evaluation and Investigation Committee of the Antalya National Education Directorate. Written informed parental consent form was obtained for all students.

ORCID

Ayse Savran Gencer https://orcid.org/0000-0001-6410-152X

Hilmi Dogan https://orcid.org/0000-0001-7933-4115

5. REFERENCES

Abdallah, M. M. S. & Wegerif, R. B. (2014). *Design-based research (DBR) in educational enquiry and technological studies: A version for PhD students targeting the integration of new technologies and literacies into educational contexts.* ERIC: ED546471. Retrieved 3, 2019, from http://files.eric.ed.gov/fulltext/ED546471.pdf

Arıcak, O. T., Avcu, A., Topçu, F., & Tutlu, M. G. (2020). Use of item response theory to validate cyberbullying sensibility scale for university students. *International Journal of Assessment Tools in Education, 7*(1), 18-29. Retrieved October 3, 2020, from https://dx.doi.org/10.21449/ijate.629584

Asysyifa, D.S., Jumadi, Wilujeng, I., & Kuswanto, H. (2019). Analysis of students critical thinking skills using partial credit models (PCM) in physics learning. *International Journal of Educational Research Review, 4*(2), 245-253. https://doi.org/10.24331/ijere.518068

Awang, Z. (2012). *A handbook on structural equation modeling using AMOS* (6th Ed). Universiti Teknologi Mara Press: Malaysia.

Baharin, N., Kamarudin, N., & Manaf, U. K. A. (2018). Integrating STEM education approach in enhancing higher order thinking skills. *International Journal of Academic Research in*

Business and Social Sciences, 8(7), 810–822. Retrieved February 3, 2019, from http://dx.doi.org/10.6007/IJARBSS/v8-i7/4421

Bailin, S. (2002). Critical thinking and science education. *Science & Education*, *11*(4), 361-375. Retrieved February 3, 2019, from http://dx.doi.org/10.1023/A:1016042608621

Beyer, B. K. (1995). *Critical thinking.* Bloomington, IN: Phi Delta Kappa Educational Foundation.

Brooks, G. P. & Johanson, G. A. (2003). TAP: Test analysis program. *Applied Psychological Measurement*, *27*(4), 303-304.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Bursal, M. (2017). *SPSS ile temel veri analizleri.* Ankara: Anı Yayıncılık.

Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı [Data analysis handbook for social sciences]*. Ankara: Pegem Akademi.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. London: Sage Publications.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Çorlu, M. S., Capraro, R. M. & Capraro, M. M. (2014). Introducing STEM education: Implications for educating our teachers for the age of innovation. *Education and Science, 39*(171), 74-85.

Deng, N., Wells, C., & Hambleton, R. (2008). A confirmatory factor analytic study examining the dimensionality of educational achievement tests. *NERA Conference Proceedings 2008*. 31. Retrieved January 10, 2020, from https://opencommons.uconn.edu/nera_2008/31

Dilekli, Y. (2019). *Etkinliklerle düşünme eğitimi*. Ankara: Anı Yayıncılık.

Doğan, N. (2019). *Eğitimde ölçme ve değerlendirme [Measuremnet and evaluation in education].* Ankara Pegem Akademi.

Doğan, H. (2020). *Design, implementation, and evaluation of the fifth grade science course units with an integrated STEM education approach* (Unpublished doctoral dissertation). Pamukkale University, Denizli, Turkey.

Duran, M., & Şendağ, S. (2012). A preliminary investigation into critical thinking skills of urban high school students: Role of an IT/STEM program. *Creative Education*, *3*(2), 241–250. Retrieved February 3, 2019, from http://dx.doi.org/10.4236/ce.2012.32038

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs: Prentice-Hall.

Ennis, R. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18*(3), 4-10. Retrieved March, 3, 2017 from https://doi.org/10.3102/0013189X018003004

Ergün, A. & Külekci, E. (2019). The effect of problem based STEM education on the perception of 5th grade students of engineering, engineers and technology. *Pedagogical Research, 4*(3), em0037. Retrieved March, 3, 2020 from https://doi.org/10.29333/pr/5842

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations* (ERIC Document Reproduction Service No. ED315423). Retrieved March, 3, 2017, from https://eric.ed.gov/?id=ED315423

Ferrando, P. J. & Lorenzo-Seva, U. (2017). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement,* 1–19. https://doi.org/10.1177/0013164417719308

Fraenkel, J.R. & Wallen, N.E. (2008). *How to design and evaluate research in education* (7th ed.). New York: McGraw-Hill.

Gencer, A. S. & Boran, G. H. (2017). *Üst düzey düşünme becerilerinin öğretimi [Teaching higher order thinking skills*]. In S. Dal & M. Köse (Ed), Öğretim ilke ve yöntemleri (pp. 405-445). Ankara: Anı Yayıncılık

Griffard, P. B., & Wandersee, J. H. (2001). The two-tier instrument on photosynthesis: what does it diagnose? *International Journal of Science Education, 23*(10), 1039-1052. Retrieved March, 3, 2017, from https://doi.org/10.1080/09500690110038549

Hair, J. F., Black, W. C, Babin, B.J., & Anderson, R. E. (2009). Multivariate data analysis (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Haladyna, T. M. (1994). *Developing validating multiple choice test items.* Lawrence Erlbaum Associates, Publishers.

Harjo, B., Kartowagiran, B., & Mahmudi, A. (2019). Development of critical thinking skill instruments on mathematical learning high school. *International Journal of Instruction, 12*(4), 149-166. Retrieved January 10, 2020, from https://doi.org/10.29333/iji.2019.12410a

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139–64.

Honey, M., Pearson, G., & Schweingruber, H. (2014). *STEM integration in K-12 education: Status, prospects, and an agenda for research*. Washington, DC: National Academies Press.

Kline, B. R. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Kuş, M. & Çakıroğlu, E., (2020). Prospective mathematics teachers‟ critical thinking processes about scientific research: Newspaper article example. *Turkish Journal of Education, 9*(1), 22-45. https://doi.org/10.19128/turje.605456

Mabruroh, F. & Suhandi, A. (2017). Construction of critical thinking skills test instrument related the concept on sound wave. *IOP Conference Series: Journal of Physics: Conf. Series* 812 (2017) 012056 https://doi.org/10.1088/1742-6596/812/1/012056

Mapeala, R. & Siew, N. M. (2015). The development and validation of a test of science critical thinking for fifth graders. *Springer Plus, 4(*741). DOI 10.1186/s40064-015-1535-0

Massachusetts Department of Education. (2006). *Massachusetts science and technology/engineering curriculum framework*. Retrieved January 10, 2019, from http://www.doe.mass.edu/frameworks/scitech/1006.do

Ministry of National Education. (2018). *Elementary and middle school (3, 4, 5, 6, 7, and 8th grades) science curriculum*. Ankara: Board of Education and Training.

Mutakinati, L., Anwari, I., & Yoshisuke, K. (2018). Analysis of students' critical thinking skill of middle school through stem education project-based learning. *Journal Pendidikan IPA Indonesia, 7*(1), 54-65. Retrieved February 3, 2019, from http://journal.unnes.ac.id/index.php/jpii https://doi.org/10.15294/jpii.v7i1.10495

NGSS Lead States. (2013). *Next generation science standards: For states by states.* Washington, DC: The National Academies Press.

National Research Council. (2011). *Assessing 21st Century Skills: Summary of a Workshop.* Washington, DC: The National Academies Press.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McG raw-Hill, Inc.

Öner, A. T., Navruz, B., Biçer, A., Peterson, C. A., Capraro, R.M., & Capraro, M.M. (2014). T-STEM academies‟ academic performance examination by education service centers: A Longitudinal Study. *Turkish Journal of Education, 3*(4), 40-51.

Oonsim., W. & Chanprasert, K. (2017). Developing critical thinking skills of grade 11 students by STEM education: Focus on electrostatic in physics. *Rangsit Journal of Educational Studies, 4* (1), 54-59.

Patton, M.Q. (2014). *Qualitative research and evaluation methods: Integrating theory and practice.* Thousand Oaks, CA: Sage Publications.

Polit, D. & Hungler, B., P. (1999). *Nursing research: Principles and methods*. Philadelphia: Lippincott Company.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459-467.

Rahmawati, Y., Ridwan, A., Hadinugrahaningsih, T., & Soeprijanto (2019). Developing critical and creative thinking skills through STEAM integration in chemistry learning. *IOP Conference. Series: Journal of Physics: Conf. Serie*s 1156 (2019) 012033. https://doi.org/10.1088/1742-6596/1156/1/012033

Reynders, G., Lantz, J., Ruder, S. M., Stanford, C. L., & Cole, R. S. (2020). Rubrics to assess critical thinking and information processing in undergraduate STEM courses. *International Journal of STEM Education, 7*(9). Retrieved February 3, 2019, from https://doi.org/10.1186/s40594-020. 00208-5

Sadhu, S. & Laksono, E. W. (2018). Development and validation of an integrated assessment for measuring critical thinking and chemical literacy in chemical Equilibrium. *International Journal of Instruction, 11*(3), 557-572. Retrieved February 3, 2019, from https://doi.org/10.12973/iji.2018.11338a

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. New Jersey: Lawrance Erlbaum Association, Inc.

Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan, R., & Kallick, B. (2008). *Thinking-based learning: Promoting quality student achievement in the 21st century*. New York, NY: Teachers College Press.

Sya'bandari, Y., Firman, H., & Rusyat, L. (2017). The development and validation of science virtual test to assess 7th grade students' critical thinking on matter and heat topic. *Journal of Science Learning, 1*(1), 17-27.

Timmerman, M. E. & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209–220. Retrieved February 3, 2019, from https://doi.org/10.1037/a0023353

Tabachnick, B. G. & Fidell, L. S. (2013). Using multivariate statistics (6th ed.). Boston: Allyn and Bacon.

Waddell, B. (2019). Influence of STEM lessons on critical thinking (Unpublished master's thesis). The Graduate College at the University of Nebraska, Lincoln. Retrieved February 3, 2019, from https://digitalcommons.unl.edu/teachlearnstudent/103

Wang, H. H., Moore, T. J., Roehrig, G. H., & Park, M. S. (2011). STEM integration: Teacher perceptions and practice. *Journal of Pre-College Engineering Education Research, 1*(2), 1-13. Retrieved February 3, 2019, from https://doi.org/10.5703/1288284314636

Whiston, S. C. (2012). *Principles and applications of assessment in counseling* (4th ed.). Belmont, CA: Brooks/Cole, Cengage Learning.

Wicaksana, Y. D., Widoretno, S., & Dwiastuti, S. (2020). The use of critical thinking aspects on module to enhance students' academic achievement. *International Journal of Instruction, 13*(2), 303-314. Retrieved February 3, 2019, from https://doi.org/10.29333/iji.2020.13221a

Wiersma, W., & Jurs, S. G. (2005). *Research methods in education: An introduction* (8th ed.). Boston: Pearson/A and B.

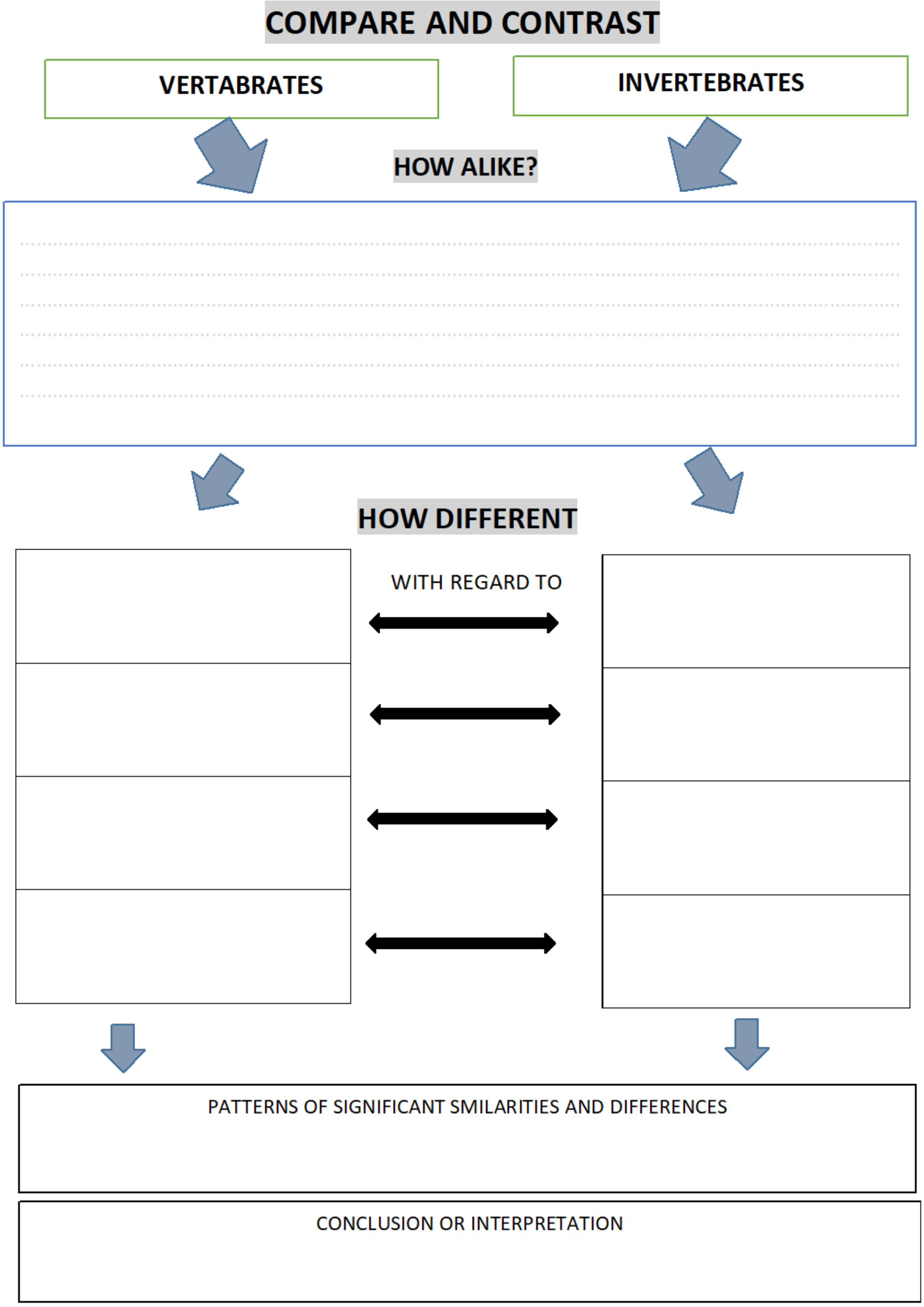
Willingham, D. T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review*, *109*(4), 21-32. Retrieved February 3, 2019, from https://doi.org/10.3200/AEPR.109.4.21-32

Yurdugül, H. & Bayrak, F. (2012). Content validity measures in scale development studies: Comparison of content validity index and kappa statics. *Hacettepe University Journal of Education, Special Issue 2,* 264-271.

Yalaki Y., Doğan, N., İrez, S., Doğan, N., Çakmakçı, G., Kara, B. E. (2019). Measuring nature of science views of middle school students. *International Journal of Assessment Tools in Education, 6(3),* 461-475. Retrieved October 3, 2019, from https://dx.doi.org/10.21449/ijate.561154

APPENDIX

Appendix I. Examples of worksheets for the integrated critical thinking skills.



COMPARE AND CONTRAST

VERTABRATES

INVERTEBRATES

HOW ALIKE?

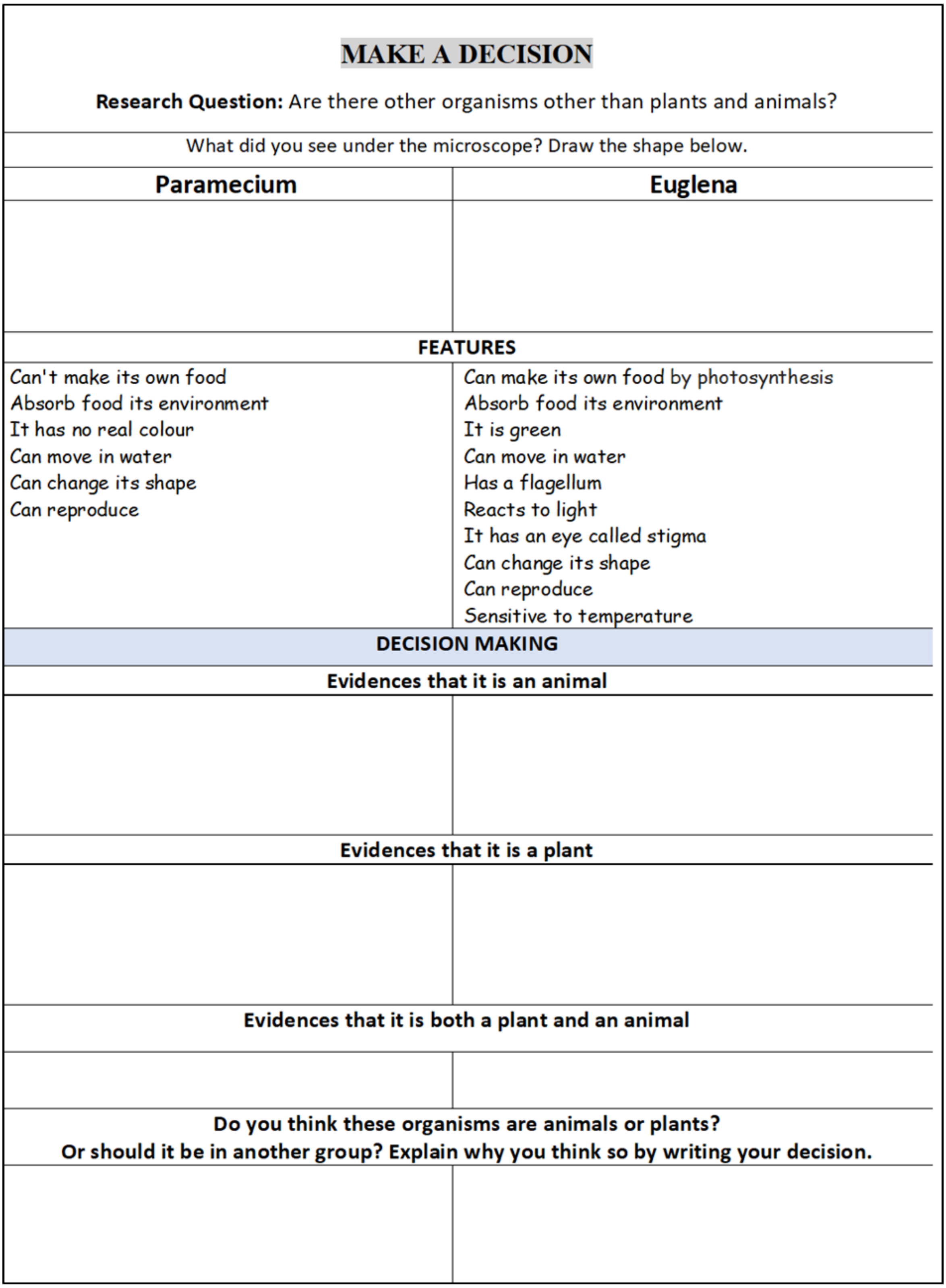
HOW DIFFERENT

WITH REGARD TO

PATTERNS OF SIGNIFICANT SMILARITIES AND DIFFERENCES

CONCLUSION OR INTERPRETATION

Adapted from Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan, R., & Kallick, B. (2008). *Thinking-based learning: Promoting quality student achievement in the 21st century*. New York, NY: Teachers College Press.



MAKE A DECISION

Research Question: Are there other organisms other than plants and animals?

What did you see under the microscope? Draw the shape below.	
Paramecium	**Euglena**
FEATURES	
Can't make its own food Absorb food its environment It has no real colour Can move in water Can change its shape Can reproduce	Can make its own food by photosynthesis Absorb food its environment It is green Can move in water Has a flagellum Reacts to light It has an eye called stigma Can change its shape Can reproduce Sensitive to temperature
DECISION MAKING	
Evidences that it is an animal	
Evidences that it is a plant	
Evidences that it is both a plant and an animal	
Do you think these organisms are animals or plants? Or should it be in another group? Explain why you think so by writing your decision.	

Adapted from Osborne, J., Erduran, S., & Simon, S. (2004). *Ideas, evidence and argument in science. Video, in-service training manual and resource pack*. London: King's College London.

DETERMINING PARTS -WHOLE RELATIONSHIP

THE WHOLE OBJECT

PARTS OF A DYNAMOMETER (What are the parts made of?)

Spring

Hook

Tube

Graduated scale

HOW DYNAMOMETER WILL WORK IF THE PARTS OF DYNAMOMETER WOULD BE BROKEN OR MISSING?

Explain how these pieces work together:

....

....

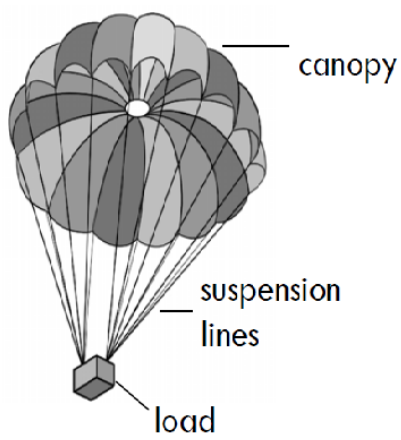
....

Adapted from Swartz, R. J., Costa, A. L., Beyer, B. K., Reagan, R., & Kallick, B. (2008). *Thinking-based learning: Promoting quality student achievement in the 21st century*. New York, NY: Teachers College Press.

PROBLEM SOLVING

Problem: Aerospace engineers want to land a robot on a newly discovered planet with a parachute. However, the atmosphere of this planet is thinner than the atmosphere of the Earth. It is known that the planet's gravitational force is greater than the Earth.

Design a parachute that can safely deliver the robot to the surface of this new planet.

Possible Solutions How can I solve the problem?	
 **A parachute that works well on the Earth**	To solve the problem, which modifications should you make on a parachute that works well on the Earth? Please explain. Write more than one solution suggestion. Solution 1 ……

Which solution should I choose?

What may happen if you follow this solution ?	Pros and Cons	How important is this result? Why?

New solution proposal

How can I improve my solution to solve this problem?

Adapted from Engineering is Elementary (n.d). *Designing Parachutes*. Museum of Science, Boston. Retrieved from http://d7.eie.org/sites/default/files/resource/file/pa_student_assessments.pdf

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 715–734

https://doi.org/10.21449/ijate.747745

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

A Study of Reliability and Validity for Citizenship Knowledge and Skill Scale

Mustafa Icen [1,*]

[1]Social Studies Education, College of Education, Yildiz Technical University, Istanbul, Turkey,

ARTICLE HISTORY

Received: Jun 19, 2020
Revised: Oct. 23, 2020
Accepted: July 27, 2020

KEYWORDS

Citizenship,
Social studies,
Citizenship education,
Reliability,
Validity

Abstract: The study aims to develop a current scale that has a higher validity and reliability and reveals high school students' perceptions of measuring their citizenship knowledge and skills. The study was conducted with two different groups. The first group is the group where data is collected to conduct Exploratory Factor Analysis (EFA) and it consists of 258 students. The second group is the group where the data is collected to carry out Confirmatory Factor Analysis (CFA) and it consists of 180 students. A total of 438 students participated in the study providing different students were in both groups. As a result of the analyzes, it was determined that the Citizenship Knowledge and Skill Scale, which includes a total of 24 items, consists of a 5-factor structure. These factors are termed as "Participation in Social Life", "Right to Education", "Individual Duties", "Duties of the State" and "Common Rights". The total variance explained by the scale is 61.79%. Additionally, there is a significant relationship between these 5 factors and there is no autocorrelation problem. Item-factor and item-test correlation coefficients were calculated for all items of the scale and it was determined that each item was consistent not only with the factor it contained but also with the whole test. Cronbach Alpha reliability of the general of Citizenship Knowledge and Skill Scale is 0.91 and Omega reliability is 0.92. It can be said that the reliability and validity of the scale are applicable and high.

1. INTRODUCTION

Citizenship is defined in the Constitution of the Republic of Turkey as follows: "Everyone bound to the Turkish state through the bond of citizenship is a Turk. The child of a Turkish father or a Turkish mother is a Turk. 'Citizenship' is acquired under the conditions determined by the law and it is lost only in cases specified in the law." (Constitution of Republic of Turkey, art. 66).

Citizenship is not just a legal formula, it refers to a social and cultural phenomenon that is becoming more and more prominent (Brubaker, 2009). It is a social identity established on the axis of citizenship, freedoms and responsibilities and open to change and restructuring (Keyman, 2008; Brubaker, 2009) It is closely related to the concepts of citizenship and identity, democracy, freedom, human dignity, human rights and respect for human rights, social justice, solidarity, and cultural, moral, mental and physical development. Teaching these concepts to individuals and internalizing by them can only be achieved through citizenship education. For

CONTACT: Mustafa Icen ✉ mustafaicen43@gmail.com Social Studies Education, College of Education, Yildiz Technical University, Istanbul, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

citizenship education approaches, it is necessary to consider and evaluate individual, society, family, and school together.

Transition to modern citizenship education in Turkey started in 1995 with the Ministry of Education's participation of the UN's "Decade of Human Rights Education" Project. It renamed the course of 'Citizenship Knowledge' as 'Citizenship and Human Rights' and revised the course content (Çayır & Bağlı, 2011; Karaman-Kepenekçi, 2005).

In this context, Turkey has increased the importance given to citizenship education since 1995, and directed to changes and innovations in education for citizenship. The title of the "Citizenship Knowledge" course has been changed to "Citizenship and Human Rights Education". This course has been abolished with the 2005 curriculum and citizenship education has started to be taught within the topics of 'Social Studies' course.

Although a new course titled "Citizenship and Democracy Education" was put into the program in 2010, this course was also abolished with the 4 + 4 + 4 education reform carried out in 2012. In the curricula published in 2018, citizenship education is once again integrated into the social studies course. In the seven learning areas in the 2018 Social Studies Curriculum, especially the acquisitions related to citizenship education have been included (Ministry of National Education of Turkey, 2018). The abolition of "Citizenship and Democracy Education", which is taught as a separate course, has been criticized by citizenship educators and still continues to be discussed. Besides its providing individuals to acquire the knowledge, understanding, talent and values that are needed by daily life, citizenship education has an important place in individuals' internalizing democratic values and expressing themselves in society by emphasizing current issues such as human rights, world problems, providing intercultural solidarity and interaction.

It is pointed out that a good citizen is a person who knows both how to govern and to be governed, which means that free people are governed by free people. Aristoteles argued that middle-class citizens, whom he thought would be able to obey the system and the logic rules more easily and comfortably, could govern the state better. According to Aristoteles, the education system in a state should be the same for everyone, and the preparation of this system should be a public activity. A good education will create citizens who desire to do their best, and a livable life will take place when all citizens establish a real community (Johnson & Morris, 2010).

Today, active and participative citizens, who are aware of global problems, are conscious about science and technology, take place in the process of preventing problems and finding solutions, and have more rights to speak in the administration process, are emphasized more and more.

The knowledge, skills and values that the countries put effort to gain to the students with the arrangements they have made in their curricula also reflect the characteristics of the citizen to be reached. Young people are hypothetically interested in ideological problems towards the future and feel the need to determine a political view or a social stance (Küçükkaragöz, 2009). On the other hand, when the European Fund 2014 Report is examined, it is determined that while participation in the politics increases in the European Union countries, there was no increase in it in Turkey and even decreased over time (Eurofound, 2014). Although the participation to the elections is higher in Turkey compared to the European Union (72%) since 88% of its voters participated in the elections, it is stated that there is low participation in the decision-making process. It is also stated that political engagement that is dealt as actively participating in a political party or trade union meeting, participating in demonstrations and communicating politicians and related stakeholders is found as 7% in Turkey. And this is indicated as a very low level compared to European Union countries.

It is seen that there are researches and measurement tools towards the level of secondary school students in 'Citizenship' course in Turkey. Citizenship education is a process that starts with preschool education and continues at all levels of education. Measuring the impact of citizenship education on students will be easier as the students move from the concrete process period to the abstract process period, their age increases and the students progress at the education level. For this reason, considering their development and learning levels, it is thought that performing the measurement of the knowledge and skills related to citizenship at the secondary education level will provide healthier and more comprehensible data. In this context, it was decided to develop a measurement tool in order to determine the citizenship knowledge and skills and perceptions of secondary school students who started to have citizenship education at the primary and secondary level.

The most obvious difference of the scale from the other scales previously developed in the literature (Doğanay, 2008; Doğanay & Sarı, 2009; Sağlam, 2000) is that it was developed for high school students. This scale was developed on high school students, and it measures the characteristics of the individuals regarding their citizenship knowledge and skills. The scales previously developed in the literature were mostly related to secondary school students and in general, they were about how the individual perceived the concept of citizenship. So they were mostly at the conceptual level. Citizenship Knowledge and Skills Scale measures the reflections of citizenship education on the individual, that is if the individual has the knowledge and skills related to citizenship and her/his ability to use these knowledge-skills in her/his life.

2. METHOD

2.1. Participants

As the study aims to develop a scale for determining the citizenship perceptions of high school students, the study group consists of high school students. The study was conducted with two different groups. The first group is the group in which data is collected to perform Exploratory Factor Analysis (EFA). The second group is the group in which data is collected to perform Confirmatory Factor Analysis (CFA) to determine to what extent the determined factors are supported by the observed variables after performing EFA.

The first group consists of 258 students. 67 (26%) of the students are males and 191 (74%) of them are females. 9 (3.5%) of the students stated that their mothers were illiterate, while 4 (1.6%) of the students stated that their fathers were illiterate. The number of students indicating that their mothers are primary school graduates is 138 (53.5%), and the number of students indicating that their fathers are primary school graduates is 75 (29.1%). The number of students whose mothers are middle school graduates is 46 (17.8%), and whose fathers are middle school graduates is 64 (24.8%). The number of mothers had university-level education is 23 (9%) and the number of fathers had university-level education is 46 (17.3). 131 (50.8) of the students are studying at Anatolian High School, 90 (34.9) of them at Vocational High School and 20 (7.8%) of them are attending Science High School and the others are at Religious Vocational High School and other high schools.

The second group of students consists of different students than in the first group. It consists of 180 students, 50 of which (27.8%) are males and 130 (72.2%) of which are females. 78 (43.3%) of these students were at Anatolian High School, 88 (48.9%) were at Vocational High School, 8 (4.4%) were at Science High School and others at Religious Vocational High School and other High Schools. 6 (3.3%) of these students stated that their mothers are illiterate, while 2 (1.1%) of the students stated that their fathers are illiterate. The number of students indicating that their mothers are primary school graduates is 57 (31.7%), and the number of students indicating that their fathers are primary school graduates is 53 (29.4%). The number of students whose mothers are middle school graduates is 50 (27.1%), and the number of students whose fathers are middle school graduates is 48 (26.7%). The number of mothers had university-level education is 19 (10.6%), and the number of fathers had university-level education is 33 (18.4%).

2.2. Item Pool Preparation Process

In order to determine the students' perception of citizenship, a literature review was conducted first. As a result of the examined literature (Doğanay, 2008; Doğanay & Sarı, 2009; Durualp & Doğan, 2017; Ersoy, 2014; İçen, Öztürk & Yılmaz, 2017; Morais & Ogden, 2011; Sağlam, 2000; 2011; Şen, 2018; 2019; Uğurlu, 2011; Üstel, 2016), a total of 27 Items were prepared including the Right to Education (6 Items), the Individual's Duties toward the State (8 Items), the Social-Political Participation (5 Items), the Duties of the State toward the Citizen (5 Items)and the Citizen Identification (3 Items). Then, a 5-point Likert scale was prepared as 1= Strongly Disagree, 2= Disagree, 3= Undecided, 4= Agree, 5= Strongly Agree.

2.3. Analysis Process

The data obtained from the second and third samples were loaded and analyzed in SPSS 23.00 and Lisrel 8.7 software to perform validity and reliability analysis of the measuring device. In addition, SPSS 23.00 was used for Exploratory Factor Analysis and Lisrel 8.7 was used for Confirmatory Factor Analysis. SPSS 23.00 and JASP 0.9.0.1 (for OMEGA reliability) software were used for reliability analysis.

After the scale items were decided, firstly, whether the scale is comprehensible was presented to the expert opinion in terms of its scope. In the second step, 5 students were asked to answer the scale one by one by reading it aloud, so that its face validity was analyzed. In order to determine the factor structure of the scale, Exploratory Factor Analysis (EFA) based on Maximum Likelihood (ML) was performed using the data obtained from the first sample (Tezci, 2016; Colton & Covert, 2007; Comrey & Lee, 1992). However, in order to apply ML, the data must show a normal distribution (MacCallum, Browne & Cai, 2007). For this reason, multivariate normality has been tested first. The reason for choosing this test was that it could be broadly generalized for situations where the data showed a normal distribution, larger correlations could be preferred, and the estimates produced less variability compared to other models were important factors (Briggs & MacCallum, 2003; Fabrigar, Wegener, MacCallum & Strahan, 1999). It is aimed to combine many interrelated measurements with typical structure or factors with factor analysis. Since factor analysis is based on the assumption that all variables are related to a certain extent (Kandemir, Tezci, Shelley & Demirli, 2019), this analysis was carried out to determine the items that are not under any factor and are not related or have a overlapping structure.

Since it was aimed to determine the minimum number of factors suitable for the original data set, ML was used (Ford, McCallum & Tait, 1986). Eigenvalue (since eigenvalue 1 and above will be significant for the factors) and scree graph were examined in determining factor numbers (Hair, Anderson, Tatham, & Black, 1995). In addition, the "direct oblimin" technique, one of the oblique rotation techniques, was used in the study. The reason for using this technique is the expectation that there may be a correlation between dimensions, especially in behavioral science areas (Byrne, 2001; Williams, Onsman, & Brown, 2020). Tabachnick & Fidell (2001) suggested that oblique rotation should be preferred if there is no significant reason-justification and if there is a coefficient of 0.32 and above in the correlation matrix. Since factor loads met the practical significance value, the value of ± 0.30 was used. The reason for using this value is its contribution to explaining the amount of the total variance calculated by a factor (Ho, 2006).

Confirmatory Factor Analysis (CFA) was performed using the data from the second sample. CFA was used to test the accuracy of the structure determined by EFA. CFA was performed in the Lisrel program with the maximum likelihood method. It was applied to test the factorial structure of the model determined by EFA (Ding, Velicer & Harlow, 1995; Gomez & Fisher, 2003). A series of indices were used to evaluate the fit of the model. Since x^2index is affected by sampling size, it is evaluated together with the degress of freedom. Apart from this, even though the values of CFI (comparative fit index), GFI (goodness of fit index), NFI (Normed Fit Index), NNFI (Non-Normed Fit Index) are demanded to be close to 1, the value of 0.90 and above can be accepted (Bentler & Bonett, 1980). Even so, Hu and Bentler (1999) stated that these values' being in 0.95 and above indicates a good fit. RMSEA (root mean square error of approximation) indicates that the value which is 0.08 or below is sufficient but 0.06 indicates a better fit (Hu & Bentler, 1999).

For the convergent validity of the scale, the analysis of the Explained Common Variance (ECV) values of each factor was determined by comparing the correlation of each factor with each other (Fornell & Larcker, 1981). Discriminant validity was evaluated by comparing the square root value of the variance explained with the square of correlations between factors. Convergence and distinctive validity is another type of validity used in testing and verifying the established model (Fornell & Larcker, 1981; Malhotra, 2011). Cronbach Alpha, Omega Reliability and Combined Reliability were calculated for reliability analysis. Composite Reliability (CR) is used to measure the internal consistency of factors and the value of 0.70 and above is considered as a good value (Hair, Black, Babin & Anderson, 2010). In the context of internal consistency, Cronbach Alpha analysis is not considered sufficient in case of multiple factor structures. It is also recommended to calculate the Omega Reliability coefficient (Dunn, Baguley & Brunsden, 2014).

3. FINDINGS

3.1. Face Validity

In order to determine whether the prepared scale items are suitable for measuring the knowledge, skills and perceptions about citizenship, 5 experts working in the social studies, citizenship education, Turkish literature and measurement fields were first presented for their opinions. As a result of the feedbacks received from the experts, it is determined that there is no need for any correction and a statement to be added. The scale form was then submitted to the opinion of 3 experts in the field of measurement and evaluation. Expressions were asked to audit in terms of scale.

As a result of this inspection, no correction was required. Then, before applying the scale, 4 different students were asked to read and answer the scale items aloud by applying face to face. Thus, it was checked whether there was any item that could not be understood or cause misunderstanding. As a result of the application, it was determined that the scale items were understandable and the data collection phase was started.

3.2. Suitability of the Data for Analysis

The first group consists of a total of 258. Descriptive analysis was conducted for the suitability of the data obtained from these students for analysis. Before applying the exploratory factor analysis, it was decided whether ML should be applied or not by examining the missing value, multicollinearity, linearity and multivariate normality in the data (Çokluk, Şekercioğlu, & Büyüköztürk, 2014). Mahalanobis' D^2, Cooks distance values for multivariate outliers were examined (Stevens, 1984). "Normality among single variables is assessed by skewness and kurtosis" (Tabachnick & Fidell, 2001, p. 613). When P-P plots, Q-Q plots, skewness and kurtosis, Mahalanobis' D^2 and Cooks ditance values were examined, it was observed that the data showed normal distribution, so ML method was applied. Whether there are extreme values, skewness and kurtosis values were examined. As a result of the descriptive analysis, it was observed that the general average of the scale was 2.94 (*SD*= .56), the Item with the lowest average (Mean= 2.56, *SD*= 1.39) was Item 6, and the item with the highest average (Mean = 3.51, *SD* = 1.23) was Item 10. It was observed that the skewness value ranged from 0.416 to -0.448, and the kurtosis value ranged from -0.686 to -1.273 (in the interval of ± 1.5). The EFA suitability of the measurements, in other words, the sampling adequacy and the test of sphericity analysis (Kaiser-Mayer-Olkin [KMO] and Bartlett 'Test) were examined to determine whether the number of samples allowed for sufficient factorization. At the end of the analysis, KMO = 0.844 was determined. This value indicates that it is a good value since it is above .700, and the sample is large enough for analysis. Bartlett test (Approx. x^2 = 3130.576, *df* = 351, *p*= .000) was observed to be significant.

In the correlation analysis, the relationship between binary and partial correlations and variables was tested. The highest binary correlation (r = .78, p <.05) was observed to be between Item 6 and Item 10 (r =-. 001, p>.05) and the lowest correlation was between Item 25 and Item 16. The highest binary correlation of Item 25 (r = .22) was observed with Item 26 and the binary correlation with other Items was observed to have been low. Similarly, the binary and partial correlations of Item 9 were observed to be very low. Correlations between the items under the factors with each other were examined. It is determined whether there is data with outliers. In addition, the Variance Inflation Factor has been tested that there is no multicollinearity between variables

3.3. Analysis of Factor Structure

Different ways can be followed to determine the factor structure of the scale. One of the commonly used approaches is the Scree Plot and the other one is to select factors with an eigenvalue above 1. In addition, the other approaches are Velicer's (1976) Minimum Average Partial Test (MAP) and Horn's (1965) Parallel Analysis. The Scree-Test) was proposed by Cattell (1978). Thompson (2004) reported that this method was criticized because, in this method, the graphic is based on the visual reading. In addition, there are opinions about the Scree plot in the data set obtained from large samples is more appropriate (Zwick & Velicer, 1986). Selection of eigenvalues greater than 1 as a factor (Kline, 1994; Tabachnick & Fidel, 2001). In this method, both of them evaluated the factor structure for reasons such as containing some sample errors, sometimes more than the number of factors (Thompson, 2004; Velicer & Jackson, 1990; Zwick & Velicer, 1986). The Scree-test is presented in Figure 1.

As seen in the Scree plot, the scale was evaluated to have 5-factor structures. There were 7 factors with an eigenvalue above 1. However, Items 25 and 9 were observed to create a separate factor. Maccallum, Widaman, Zhang and Hong (1999) suggested that the number of Items in a factor should be 3 or more for the beneficial factor. Therefore, when the factors consisting of a single Item were ignored, it was observed that it had 5-factor structures. It was also evaluated that the factor structures obtained were significant (Tatlıdil, 1992).

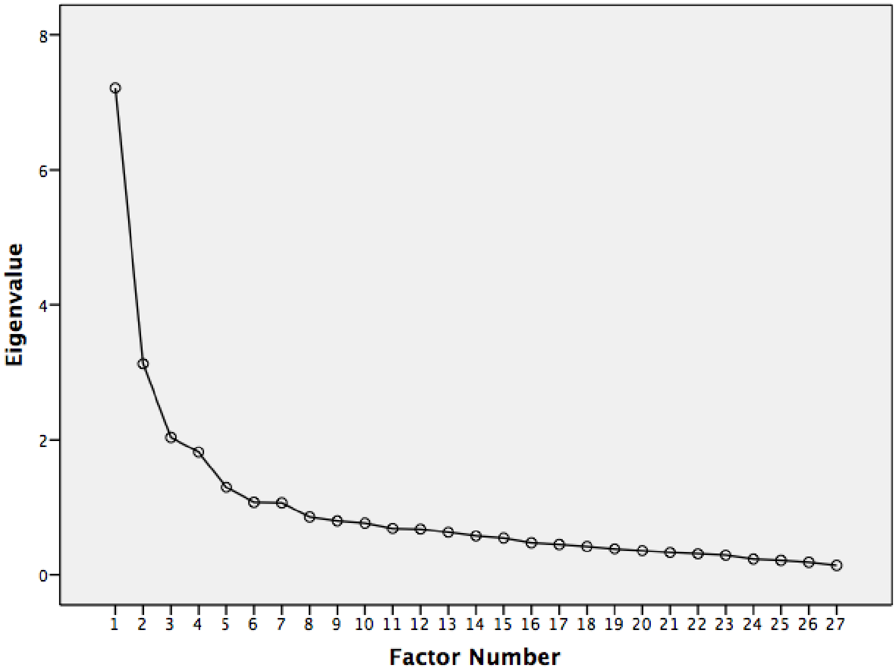


Figure 1. *Scree-Test*

Parallel analysis approach and MAP test have been determined to have had similar results in many studies (e.g. O'Connor, 2000; Yavuz & Doğan, 2015; Zwick & Velicer, 1986). In this study, the number of factors was determined by using Syntax written by O'Connor (2000). As a result of the analysis, Average Partial Correlations results are given in Table 1.

Table 1. *Eigenvalues Regarding Partial Correlations Obtained from the MAP Test*

	Squared	Power4		Squared	Power4		Squared	Power4
0	.0734	.0152	9	.0278	.0026	18	.0985	.0301
1	.0280	.0029	10	.0320	.0035	19	.1161	.0381
2	.0230	.0014	11	.0376	.0050	20	.1352	.0527
3	.0211	.0012	12	.0431	.0060	21	.1699	.0749
4	.0178	.0011	13	.0491	.0079	22	.2330	.1128
5	.0165	.0010	14	.0541	.0100	23	.3008	.1698
6	.0195	.0015	15	.0604	.0145	24	.3315	.2088
7	.0216	.0019	26	.0709	.0172	25	.4912	.3773
8	.0250	.0025	17	.0836	.0235	26	1.0000	1.0000

As a result of the analysis, it was observed that the smallest average squared partial correlation was .0165 and this value was in the 5th step. It was determined that the fourth power of the partial correlation took place in the 5th step. The fourth power of the partial correlation was included in the program by O'Conner (2000). At this point, the scree-test, eigenvalue and MAP analysis were evaluated together and the number of factors was decided.

3.4. Exploratory Factor Analysis

In determining the construct validity of the scale, EFA was performed first. As a result of the ML analysis based on the data obtained from a total of 27-item scale obtained from 258 students, it was observed that the factor load has a 5-factor-structure which is greater than ±0.30 and its eigenvalue is greater than 1. However, Item 25 which includes the definition of citizenship (Being Turkish requires being a Muslim) and Item 9 related to individual rights (Individuals should have the right to express their racist ideas as well) were excluded from the scale. Item 14 (I am thinking of joining a political party in the future) was also excluded from the scale since it was placed under two factors (Its load in Factor 1= 0.340, lts load in Factor 4= .406). The factor loads of the scale Items and the total variance explained and factors are given in Table 2.

Table 2. *Factor Loads of the Scale Items and Factors that They are Distributed*

	Factors						
	1	2	3	4	5	Mean	*SD*
M15- I believe I will do voluntary work to help people in the future.	**.884**	.302	.410	-.336	.506	2.61	1.50
M6- I want to contribute to the solutions of the people's problems in other countries.	**.884**	.324	.384		.477	2.56	1.39
M24- I believe that women and men have equal rights in the social environment I am living in.	**.833**		.418	-.328	.601	2.43	1.52
M13- I would like to take an active role in non-governmental organizations in the future.	**.689**		.319		.402	2.71	1.33
M27- Every citizen should take an active role in reducing social and economic inequalities.	**.603**		.375		.388	2.99	1.53
M4- I believe my high school education gave me the knowledge, skills, rights and responsibilities that I need to be as a good citizen.		**.846**				2.87	1.22
M1- In the high school I am studying, teachers respect students' ideas and encourage us to express our opinions.	.327	**.748**				2.73	1.11
M5- I believe that my high school education has improved my interest and ability to make a common decision by discussing a social problem.		**.731**				2.93	1.22
M2- Teachers give us the opportunity to discuss the issues that society disagrees with.		**.719**				2.86	1.06
M3- Teachers take care to explain the issue neutrally when describing a problem.	.325	**.604**				2.89	1.08
M8- Whenever I need, I go to the police department because I trust the police.	.469		**.797**			2.83	1.26
M7- If I am subjected to injustice, I confidently apply to the courts.	.352	.309	**.677**			3.16	1.18
M16- I know how to seek my rights when my citizenship rights are violated.	.450	.304	**.634**		.408	2.92	1.26
M12-The main citizenship duties are joining the military service, paying tax and go to the polls for voting.	.348		**.615**			2.92	1.28
M17- Turkey should accept more immigrants.	.308		**.507**			2.93	1.29
M18- People who were not born in Turkey but living in Turkey should have the same rights as everyone else.			**.490**		.308	3.02	1.38
M21- A quality education is necessary for men rather than women.				**.818**		3.30	1.39

M20- People who were not born in Turkey but living in Turkey must learn Turkish.				**.626**		3.34	1.36
M26- "Turkish Nation" is the common name of all citizens living in our country.				**.558**		3.23	1.36
M10- Although a law violates human rights, people must obey that law.				**.451**	-.308	3.51	1.23
M19- If there are not enough jobs to employ everyone, men should be employed more than women.	.521		.339		**.803**	2.81	1.44
M23- State officials should not be allowed to display their (religious, ethnic and sectarian identities) at work.	.412				**.704**	2.88	1.43
M11- Citizens should be able to organize peaceful protests when they find necessary.	.575				**.691**	2.60	1.36
M22- The state should provide enough financial support to the unemployed to maintain their lives.	.417				**.622**	2.86	1.28
Total Variance Explained	27.986%	10.791%	5.428%	4.636%	3.432%		

*Unvalidated translation. The scale was developed in Turkish. In order to use it in different languages, it should be re-evaluated in terms of its realiability and validity, again.

As a result of EFA, a total of 5 factor structures were determined. Factor 1 is named as "Participation in Social Life" since it contains items related to duties in individual's social life. There are 5 items under this factor. The item with the highest factor load is Item 6 and 8 with 0.884, and the item with the lowest factor load is Item 27 with 0.603. The contribution of these items to the total variance is 27.986%. The second factor is relevant to the perception of citizenship regarding education. For this reason, this factor has been named as "Right to Education". There are also 5 items in this factor. The total explained variance is 10.791%. The Item with the highest factor load is Item 4 with 0.846. And the item with the lowest factor load is Item 3 with 0.604. The third factor contains 6 items and the total explained variance is 5.428%. The Items under this factor are named as 'Individual Duties' since they include the citizenship duties toward the state. In this factor, the highest factor load is Item 8 with 0.797, and the lowest factor load is Item 18 with 0.490. There are 4 Items in the fourth factor and the total explained variance is 4.636%. The Items here reflect the notion of the social state. For this reason, it was named as "Duties of the State". The highest factor load belongs to Item 21 with 0.813, while the Item with the lowest factor load is Item 10 with 0.451. Factor 5 consists of 4 items. The contribution of this factor to variance is 3.432%. Since the items under this factor concern common rights, they are named as "Individual Rights". The highest factor load belongs to Item 19 with 0.803 and the lowest factor load belongs to Item 22 with 0.622. The total variance explained by the scale is 52.273%. It is seen that the factor load of some items is under more than one factor. For example, the factor load of Item 15 under Factor 1 is 0.884 and the load under Factor 5 is .506. Similarly, although Item 24 is under two factors, its factor loads under two factors are above .100. Therefore, it was decided that it was under the factor with a high factor load.

3.5. Descriptive Analysis Results

The second group consists of 180 students. The data obtained from these students were used in CFA, convergence validity and discriminant validity analyses. Descriptive analysis results of the data obtained from these students are given in Table 3.

Table 3. *Descriptive Analysis Results for the Second Group*

	Minimum	Maximum	Mean	*SD*	Skewness	Kurtosis
1- Right to Education	1,00	5.00	3.09	.89	-.080	-.405
2- Participation in Social Life	1.00	5.00	2.92	1.13	.084	-1.107
3- Individual Duties	1.00	5.00	3.07	.93	.093	-.899
4- Duties of the State	1.00	5.00	3.07	.95	-.049	-.804
5- Common Citizenship Rights	1,00	5.00	3.33	.99	-.145	-.819
General	1.55	4.72	3.09	.69	.132	-.821

As a result of the analysis, it was observed that the average of the scale was 3.09 (*SD* = .69). In the analysis made in terms of sub-dimensions (factors), the lowest average (Mean= 2.92, *SD* = 1.13) is in the second factor, the ‘Participation in Social Life’, and the highest average (Mean =3.33, *SD*= .99) is in the ‘Common Citizenship Rights.’ It was observed that the Skewness and Kurtosis values of the scores obtained from the scale were within the interval of ± 1.5.

3.6. CFA Analysis Results

CFA analysis was applied to the data obtained from the second sample group to which the scale was applied and the suitability of the structure determined in EFA was tested. Jöreskog (1969) suggested that EFA will generally be used to determine the construct validity of the scales. As a result of the CFA analysis applied to the data obtained from 180 students, x^2/df ratio was determined as (608.40/242) = 2.51. In addition, some fit indices (RMSEA = 0.088, RMR = .12) were observed to be higher than they should be, while others were observed lower than they should be (NFI = 0.92, RFI = 0.91, GFI = 0.80, AGFI = 0.75). Schreiber, Nora, Stage, Barlow, and King (2006) stated that it is not necessary to reconstruct the theoretical model if a sufficient fit index is achieved with the proposed modifications. For this reason, the proposed fit indices were examined. Eventually, as a result of the correction of the error variances of some variables, fit indices were obtained within acceptable limits. As a result of the modification made by based on the 4 error variances applied in the Education, Individual Duty, Common Rights factors, the ratio of x^2/df was found as (441.05 / 238) = 1.85 between the items of 4 and 5 in the 'Education' factor; between the Items of 8 and 18 in the 'Individual Duty' factor, and between the items of 20 and 21 and items of 10 and 21 in the 'Common Rights' factor. Although the value was excellent, other fit indices were also examined since it is affected by the sample size. According to this, the values of RMSEA = 0.066; NFI = .94, NNFI = 0.96, CFI = 0.97, IFI = .97; RMR = 0.09; SRMR = 0.07, GFI = 0.84 AGFI = 0.80 were determined. Some of the values (x^2/df, NNFI, CFI and IFI) were excellent, some of them (NFI, RMR, SRMR, RMSEA) were acceptable, while others (GFI and AGFI) were observed to be low (Hair et al., 1998; Hoyle, 1995). A sufficient fit index was obtained with the proposed fit indices. The Path Analysis Diagram obtained as a result of the analysis is given in Figure 2.

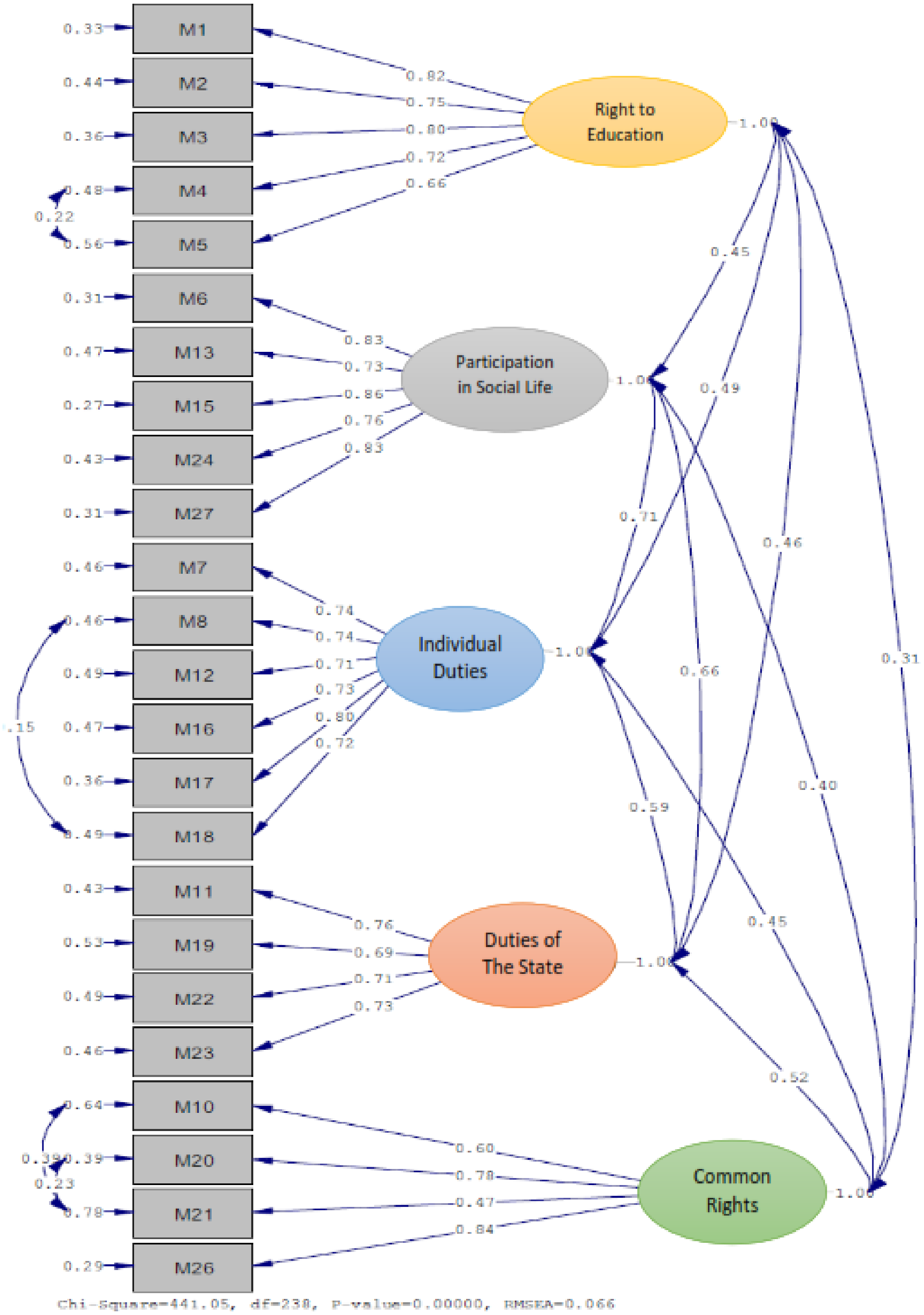


Figure 2. *Standardized path analysis diagram*

As a result of the modification made between a total of four items, it was seen that the indexes reached a level that could be called "good". In addition, no significant change was observed in the explained factor loads of these items, except for the 21st item. The factor load of the 21st item decreased from .70 to .47. However, since this value is above .30 (Harrington, 2009), it is at an acceptable level. It was determined that this item with the lowest load value was significant ($t = 6.20$, $p < .05$). As given in the Path Analysis diagram, the paths drawn for implicit variables from all observed variables were found to be significant. When the path coefficients are analyzed, it was determined that the lowest load belongs to Item 21 (0.47) in 'Common Citizenship Rights' factor, and the highest load belongs to Item 15 in 'Participation in Social Life' Factor. Ford, McCallum and Tait (1986) suggested factor loadings in social sciences to be above 0.40. Factor load values of all items ranged from 0.47 to 0.86. In addition, it was observed that the correlations between the implicit variables were positive and that the highest correlation was between 'Individual Duties' and 'Participation in Social Life' (0.71) and that the lowest correlation (0.31) was between 'Common Citizenship Rights' and 'Right to Education.' And there was no correlation above 0.85 and the correlations were significant (Bryne, 2001).

3.7. Corvergent and Discriminant Validity

Although CFA is used for construct validity, Campbell and Fiske (1959) also recommended to examine the convergence and discriminant validity to determine the 'structure' of a measuring tool. Convergence validity is the degree of confidence of the feature, which is measured well by its indicators, while the discriminant validity is the degree of measuring different features that are unrelated to each other. Or it is the relationship between observed variables that measure kthe latent variable. The discriminant validity is used to determine whether the observed variables are representative of the latent structures to which they belong to (Hair, et al., 2010). According to the Fornell-Larcker (1981) criterion, it is widely used in CFA to evaluate the degree of common variance shared among the implicit variables of the model. According to this criterion, convergent validty of the measurement model can be evaluated with Average Variance Extracted [AVE] and Combined Reliability (CR-Composite Reliability [CR]). Acceptable value of CR is 0.70 and above and acceptable value of AVE is 0.70 and above, but 0.50 and above is sufficient. In addition, the CR value should be greater than the AVE value (Gouveia & Soares, 2015; Raykov, 1997). On the other hand, the square root of the AVE value should be greater than the correlation values between the latent variables (Bagozzi & Yi, 1988; Hair et al., 2010; Hu & Bentler, 1999). The Maximum Shared Variance (MSV) and Average Shared Variance (ASV) values were examined for the discriminant validity. AVE> MSV and AVE> ASV criteria determined by Hair et al. (2010) were taken into consideration for the for the discriminant validity.

Construct validity of whether or not the citizenship scale measures the structure to which it is directed was tried to be determined by using discriminant validity, which is a version of a) convergent validity and b) divergent validity. Average Variance Extracted (AVE) and Combined Reliability (CR) values are presented in Table 4.

Table 4. *Correlations between AVE, CR Values and Factors*

	AVE	CR	MSV	ASV	1	2	3	4	5
1- Rights to Education	.57	.87	.24	.19	**(.76)**				
2- Participation in Social Life	.65	.90	.50	.33	.45	**(.81)**			
3- Individual Duties	.55	.88	.50	.32	.49	.71	**(.74)**		
4- Duties of the State	.52	.81	.44	.36	.46	.66	.59	**(.72)**	
5- Common Citizenship Rights	.48	.78	.27	.18	.31	.40	.45	.52	**(.69)**

The fact that the factor loadings and AVE values of the scale are greater than .50 is a proof of the convergence validity of that measuring tool. However, if CR values are .70 and above, it is sufficient that ASV value is 0.40 and above (Formel & Larcker, 1981; Peterson, 2000). The fact that ASV values of the scale show that 0.50 and CR are above 0.70 shows that it has convergence validity. Although the ASV value for Common Citizenship Rights is 0.48, this value is at an acceptable level considering that CR coefficient is .78. For the discriminant validity, the Formel and Larcker (1981) criteria was used. Accordingly, the correlation coefficients between the square root of the ASV value and each structure in each row-column were examined. Accordingly, the correlation between each structure is lower than the square root of the ASV value. Also, MSV and ASV values in all sub-factors are lower than the ASV value. This shows that it contributes to the positive discrimination of the measurement model. The results can be said that each structure measures different features.

3.8. Cronbach Alpha and Omega Reliability

The reliability of the data obtained from the measurement tool in terms of internal consistency was tested with the Cronbach Alpha coefficient. In addition, Omega Reliability (Zinbarg, Yovel, Revelle & McDonald, 2006) value, which is a recommended reliability in multi-factor scales and when factor loads are not equal, was calculated and compared with Cronbach Alpha Coefficient. Analysis results are presented in Table 5.

Table 5. *Cronbach Alpha and Omega Reliability*

	Alpha for N= 258	Alpha for N= 180	Alpha for N= 438	Omega for N= 258	Omega for N= 180	Omega for N=438
Rights to Education	.85	.87	.84	.85	.87	.85
Participation in Social Life	.85	.89	.86	.86	.89	.86
Individual Duties	.78	.85	.78	.79	.85	.79
Duties of the State	.81	.86	.73	.81	.86	.74
Common Citizenship Rights	.70	.84	.67	.69	.84	.68

As a result of the analysis, Cronbach Alpha reliability and Omega Reliability are high for each sub-factor of the data obtained from the first sample. However, in the 'Common Citizenship Rights' factor, both Alpha (.67) and Omega Reliability coefficients (.68) of the 438-person group, in which both the first and second sample groups were evaluated together, were found lower than the others. Cronbach Alpha reliability of the overall scale is 0.91 and Omega reliability is 0.92.

3.8. Item-Discrimination and Item-Total Correlations

As a result of the analysis regarding item-total correlations, the lowest correlation (r = .41) in 'Right to Education' factor belongs to Item 2 and the highest correlation (r = .55) belongs to Item 1. The lowest correlation (r = .56) in 'Participation in Social Life' factor belongs to Item 13 and the highest correlation (r = .67) belongs to Item 6. The lowest correlation (r =. 54) of 'Individual Duties' belongs to Item 12, and the highest correlation (r = .60) of it belongs to Item 16. The lowest correlation (r = .47) in 'Duties of the State' factor is Item 19 and the highest correlation (r = .67) of it belongs to Item 17. In the 'Common Citizenship Rights' factor, the lowest correlation (r =. 18) belongs to Item 21 and the highest correlation (r = .53) of it belongs to Item 26.

In order to determine the discrimination of the sub-factors of each item and the overall scale, a comparison was made with the upper-lower group-27% technique. As a result of the analysis, the lowest t value (t= 3.988, p <.05) belongs to Item 23 and the highest t value (t = 15.818, p <.05) belongs to Item 9. It has been observed that there is a significant difference between the upper and lower groups for all items.

3.9. Correlation between Scale Factors

In order to determine the relationship between the sub-dimensions of the scale, correlation analysis was performed. The results of the analysis are presented in Table 6.

Table 6. *Correlation Analysis between Scale Dimensions*

	Rights to Education	Social Life	Individual Duties	Duties of the State	Individual Rights
Rights to Education	1				
Social Life	.379**	1			
Individual Duties	.402**	.627**	1		
Duties of the State	.376**	.564**	.487**	1	
Individual Rights	.217**	.214*	.241**	.313**	1
Citizenship Knowledge and Skills Scale General Avg.	.644**	.796**	.778**	.775**	.541**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

As a result of the correlation analysis between the sub-dimensions of the scale, it was observed that the lowest correlation (r = .21, p <.05) was between Social Life and Individual Rights (r=.63, p <.01) and the highest correlation was between Social Life and Individual Duties. The lowest correlation between the general average of the scale and the sub-dimensions (r = .54, p<.01) is between Individual Rights and the highest correlation (r = .80, p <.01) is between Social Life.

4. DISCUSSION and CONCLUSION

A scale with the aim of determining the citizenship knowledge and skills of secondary school students in Turkey have been developed. In the study, statistical analyses of the validity and reliability of this scale, which is expected to contribute to the citizenship education literature, is included.

For the validity and reliability study of the Citizenship Knowledge and Skills Scale, high school students studying in public schools in Istanbul were selected by simple random sampling method. In order to determine the psychometric properties of the scale, Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), content validity, construct validity and statistical analyzes such as internal consistency analysis and Cronbach Alpha calculations were included. As a result of the analyzes, it was determined that the Citizenship Knowledge and Skill Scale, which includes a total of 24 items, consists of a 5-factor structure. The first factor is named as "Participation in Social Life" because it contains items related to duties in social life. There are 5 items under this factor.

The contribution of these items to the total variance is 27.986%. The second factor concerns the perception of citizenship regarding education. For this reason, this factor has been termed as "Right to Education". There are also 5 items in this factor. The total variance explained is 10.791%. The third factor contains 6 items and the total variance explained is 5.428%. The items under this factor are termed as "Individual Duties" since they include citizenship duties toward the state. There are 4 items in the fourth factor and the total variance explained is 4.636%. The items here reflect the notion of the social state. For this reason, it was termed as "Duties of the State". The fifth factor consists of 4 items and total variance explanined is 3.432%. The items under this factor are expressed as "Common Rights" since they concern general rights. The total variance explained by the scale is 52.273%. Cronbach Alpha reliability of the general of Citizenship Knowledge and Skill Scale is 0.91 and its Omega reliability is 0.92. It can be said that the reliability and validity of the scale are applicable and high.

The developed scale can be applied to secondary school (high school) students in order to measure their citizenship knowledge, skills and perceptions. However, if it is considered to be appropriate for the students' levels after taking an expert opinion, it can be used at other education levels to measure the students' citizenship knowledge and skills.

It should be noted that in this study, the numbers of samples were low. Specially, the number of data collected for Confirmatory Factor Analysis (CFA) was also low. Therefore, the measurement invariance could not be discussed. In order to discuss the measurement invariance, it should be tested in terms of gender and other variables. It can be done by increasing the numbers of samples. It is also useful to test the data by increasing the number of data in terms of different demographic data.

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).

ORCID

Mustafa İçen https://orcid.org/0000-0002-3289-6097

5. REFERENCES

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16* (Spring), 74-94.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606. https://doi.org/10.1037/0033-2909.88.3.588

Brubaker, R. (2009). *Fransa ve Almanya'da vatandaşlık ve ulus ruhu* (V. Pekel, Çev.). Ankara: Dost Kitabevi.

Briggs, N.E., & MacCallum. R.C. (2003). Recovery of weak common factors by Maximum Likelihood and Ordinary Least Squares Estimation. *Multivariate Behavioral Research. 38*(1). 25-56.

Bryne, B. M. (2001), *Structural equation modeling with AMOS Mahwah,* NJ: Lawrence Erlbaum Associates.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL* Ankara: Pegem Akademi.

Colton, D., & Covert, R. (2007). *Designing and constructing ınstruments for social research and evaluation.* San Francisco, CA: Jossey-Bass.

Comrey, A.L. & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Erlbaum.

Çayır, K., & Bağlı, M. T. (2011). 'No-one respects them anyway': secondary school students' perceptions of human rights education in Turkey. *Intercultural Education*, *22*(1), 1-14.

Ding, L., Velicer, W. F. & Harlow, L. L. (1995). Effects of estimation methods, number indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling, 2,* 119-144. https://doi.org/10.1080/10705519509540000

Doğanay, A. (2008). What does democracy mean to 14-year-old turkish children? A Comparison with Results of the 1999 IEA Civic Education Study. *Research Papers in Education, 25*(1), 51-71.

Doğanay, A., & Sarı, M. (2009). Lise öğrencilerinin vatandaşlık algılarına etki eden faktörlerin analizi [Analysis of the factors affecting high school students' perception of citizenship]. In A. Şişman, İ. Acun, C. Balkır, C. Yücel, H. Busher, T. Lawson, C. Wilkins, …, H. Ermiş (Eds.), *1st International European Union, Democracy, Citizenship and Citizenship Education* (p. 36-51). Uşak: Uşak University European Union Research Center.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A Practical solution to the pervasive problem of ınternal consistency estimation. *British Journal of Psychology, 105*, 399-412. https://doi.org/10.1111/bjop.12046

Durualp, E. & Doğan, İ. (2017). Vatandaşlık algısı ölçeği'nin faktör yapısının incelenmesi. *International Journal of Social Science, Number: 62, p. 65-83.* http://dx.doi.org/10.9761/JASSS7173

Ersoy, A. F. (2014). Active and democratic citizenship education and its challenges in social studies classroom. *Eurasian Journal of Educational Research,* 55, 1-20.

Eurofound (2014). *Quality of the life trends in Turkey: 2003–2012*. Luxemburg: Publications Office of the European Union.

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

Ford, J. K., McCallum, R. S. & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*, 291-314.

Fornell, C. & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50.

Gomez, R., & Fisher, J. W. (2003). Domains of spiritual well-being and development and validation of the Spiritual Well-Being Questionnaire. *Personality and Individual Differences, 35*(8), 1975–1991. https://doi.org/10.1016/S0191-8869(03)00045-X

Gouveia, V. V., & Soares, A. K. S. (2015). Calculadoras de validade de construto (CVC). João Pessoa, PB: BNCS/ Universidade Federal da Paraíba, [*Construct Validity Calculators* (*CVC*)] Retrieved from http://akssoares.com/psicometria/calculadora-vme-e-cc

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis with readings*, (4 th ed.), Englewood Cliffs, NJ: Prentice-Hall.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice hall.

Hair, J.F., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Harrington, D. (2009). *Confirmatory factor analysis*. Oxford university press.

Ho, R. (2006). *Handbook of univariate and multivariate data analysis and interpretation with SPSS*. NY: Chapman and Hall/CRC.

Hoyle, R. H., (1995). The structural equation modeling approach: Basic concepts and fundamental issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts issues and applications* (pp. 1-15). Thousand Oaks, CA: Sage.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185.

Hu, L. T., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. https://doi.org/10.1080/10705519909540118

İçen, M., Öztürk, C., & Yılmaz, A. (2017). Vatandaşlık duygusu ölçeği güvenirlik ve geçerlik çalışması. *Uluslararası Alan Eğitimi Dergisi*, *3*(2), 26-36.

Johnson, L., & Morris, P. (2010). Towards a framework for critical citizenship education. *The Curriculum Journal,* 21(1), 77-96. https://doi.org/10.1080/09585170903560444

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183-202.

Kandemir, M. A., Tezci, E., Shelley, M., & Demirli, C. (2019). Measurement of creative teaching in mathematics class. *Creativity Research Journal, 31*(3), 1-12. https://doi.org/10.1080/10400419.2019.1641677

Karaman-Kepenekçi, Y. (2005). A study of effectiveness of human rights education in Turkey. *Journal of Peace Education, 2(1)*, 39-55.

Keyman, F. (2008). *Kimlik, vatandaşlık ve demokratikleşme: Türkiye örneği.* İstanbul: Osmanlı Bankası Arşiv ve Araştırma Merkezi Yayınları.

Kline, P. (1994). *An easy guide to factor analysis*. New York: Routledge.

Küçükkaragöz, H. (2009). Bilişsel gelişim ve dil gelişimi (Cognitive development and language development). In B. Yeşilyaprak (Ed.), *Gelişim ve Öğrenme Psikolojisi (Development and Learning Psychology)* (s.75-107). Ankara: Pegem Akademi Yayınları.

MacCallum, R. C., Browne, M. W., & Cai, L. (2007). Factor analysis models as approximations. In R. Cudeck & R. C. MacCallum (Eds.). *Factor analysis at 100: Historical developments and future directions* (pp.153-175). Mahwah. NJ: Lawrence Erlbaum Associates. Publishers.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). *Sample size in factor analysis. Psychological Methods, 4(1), 84–99.* https://doi.org/10.1037/1082-989x.4.1.84

Malhotra, N. K. (2011). *Pesquisa de Marketing: uma orientação aplicada, (6th ed.)* São Paulo: Bookman.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika. 57*, 519-530.

Ministry of National Education of Turkey. (2018). *Sosyal bilgiler dersi öğretim programı (İlkokul ve ortaokul 4, 5, 6 ve 7. sınıflar) [Social studies curriculum (primary and middle school 4, 5, 6, and 7 grades)].* Ankara: MEB Devlet Kitapları. Retrieved from http://mufredat.meb.gov.tr/Programlar.aspx

Morais, D. B., & Ogden, A. C. (2011). Initial development and validation of the global citizenship scale. *Journal of studies in international education*, *15*(5), 445-466.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32,* 396-402.

Peterson, R. A. (2000). A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis. *Marketing letters*, *11*(3), 261-275.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173-184.

Sağlam, H. İ. (2000). Sosyal bilgiler dersinin demokratik tutum geliştirmedeki rolü. *Milli Eğitim Dergisi, 146,* 67-71.

Sağlam, H. İ. (2011). Öğretmen adaylarının etkili vatandaşlık yeterlik düzeyleri. *Kastamonu Eğitim Dergisi*, *19*(1), 39-50.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323-38.

Stevens, J.P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95,* 334-344.

Şen, A. (2018). Militarisation of citizenship education curriculum in Turkey. *Journal of Peace Education, 16*(1), 1-26. https://doi.org/10.1080/17400201.2018.1481019

Şen, A. (2019). Vatandaşlık eğitiminde değişiklik ve süreklilikler: 2018 sosyal bilgiler öğretim programı nasıl bir vatandaşlık eğitimi öngörüyor? (Changes and continuities in citizenship education: what kind of citizenship education does the 2018 social studies programme of study envisage?) *Eğitimde Nitel Araştırmalar Dergisi, 7(1),* 1-28. https://doi.org/10.14689/issn.2148-2624.1.7c1s.1m

Tabachnick, B. G., & Fidel, L. S. (2001). *Using Multivariate Statistics*. Boston, MA: Allyn and Bacon.

Tatlıdil, H. (1992). Uygulamalı çok değişkenli istatistik (Applied multivariate statistics). Ankara: Akademi Publishing.

Tezci, E. (2016). *Eğitimde ölçme ve değerlendirme* (*Measurement and evaluation in education*). Ankara: Detay Publishing.

Thompson, B. (2004). *Explaratory and Confirmatory Factor Analysis: Understanding Concepts and Applications.* Washington: American Psychological Association.

Uğurlu, C. T. (2011). Citizeinship education in European Union countries and Turkey. *Electronic Journal of Social Sciences*, *10*(37), 153-169.

Üstel, F. (2016). *Makbul vatandaşın peşinde II. Meşrutiyet'ten bugüne vatandaşlık eğitimi* (*In pursuit of the acceptable citizen citizenship education since the II. Constitutional Monarchy*) (7th ed.). İstanbul: İletişim Publishing.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 321-327.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25*, 1-28.

Williams, B., Onsman, A., & Brown, T. (2020). Exploratory factor analysis: A five step guide for novices. *Journal of Emergency Primary Health Care, 8*(3), 1-13.

Yavuz, G., & Doğan, N. (2015). Boyut sayısı belirlemede Velicer'in map testi ve Horn'un paralel analizinin kullanılması. *Hacettepe University Journal of Education*, *30*(3), 176-188.

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement*, *30*(2), 121-144.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.

6. APPENDIX

Citizenship Knowledge and Skill Scale

PARTICIPATION IN SOCIAL LIFE (First Factor-5 item)

6	I want to contribute to the solutions of the people's problems in other countries..
13	I would like to take an active role in non-governmental organizations in the future.
15	I believe I will do voluntary work to help people in the future.
24	I believe that women and men have equal rights in the social environment I live in.
27	Every citizen should take an active role in reducing social and economic inequalities.

RIGHT TO EDUCATION (Second Factor-5 item)

1	In the high school I study, teachers respect students' ideas and encourage us to express our opinions.
2	Teachers give us the opportunity to discuss the issues that society disagrees with.
3	Teachers take care to explain the issue neutrally when describing a problem.
4	I believe that my high school education gave me the knowledge, skills, rights and responsibilities that I need to be as a good citizen.
5	I believe that my high school education has improved my interest and ability to make a common decision by discussing a social problem.

INDIVIDUAL DUTIES (Third Factor-6 item)

7	If I am subjected to injustice, I confidently apply to the courts.
8	Whenever I need, I go to the police department because I trust the police.
12	The basic citizenship duties are joining the military service, paying taxes and voting.
16	I know how to seek my rights when my citizenship rights are violated.
17	Turkey should accept more immigrants.
18	People who were not born in Turkey but living in Turkey should have the same rights as everyone else.

DUTIES OF THE STATE (Fourth Factor-4 item)

11	Citizens should be able to organize peaceful protests when they find necessary.
19	If there are not enough jobs to employ everyone, men should be employed more than women.
22	The state should provide enough financial support to the unemployed to maintain their lives.
23	State officials should not be allowed to display their (religious, ethnic and sectarian identities) at work.

COMMON RIGHTS (Fifth Factor-4 item)

10	Although a law violates human rights, people must obey that law.
20	People who were not born in Turkey but living in Turkey must learn Turkish.
21	A quality education is necessary for men rather than women.
26	"Turkish Nation" is the common name of all citizens living in our country.

Vatandaşlık Bilgi ve Beceri Ölçeği (Turkish)

TOPLUMSAL HAYATA KATILIM (Birinci Faktör-5 madde)

6	Diğer ülkelerdeki insanların sorunlarının çözümüne katkı sunmak istiyorum.
13	Gelecekte sivil toplum örgütlerinde aktif rol almayı isterim.
15	Gelecekte insanlara yardım etmek için gönüllü işler yapacağımı düşünüyorum.
24	Yaşadığım sosyal çevrede kadın ve erkeklerin eşit haklara sahip olduğunu düşünüyorum.
27	Her vatandaş, sosyal ve ekonomik eşitsizlikleri azaltmak için aktif rol üstlenmelidir.

EĞİTİM HAKKI (İkinci Faktör-5 madde)

1	Okuduğum lisede öğretmenler öğrencilerin fikirlerine saygı göstererek fikirlerimizi ifade etme konusunda bizleri cesaretlendirirler.
2	Öğretmenler, toplumun görüş ayrılığına düştüğü konuları tartışmamıza fırsat verirler.
3	Öğretmenler, bir sorunu anlatırken konuyu tarafsız olarak anlatmaya özen gösterirler.
4	Lise eğitimimin iyi bir vatandaş olmam için gereken bilgi, beceri, hak ve sorumlulukları bana sunduğunu düşünüyorum.
5	Lise eğitimimin toplumsal bir sorunu tartışarak ortak bir karar verme konusunda ilgi ve yeteneğimi geliştirdiğini düşünüyorum.

BİREYSEL GÖREVLER (Üçüncü Faktör-6 madde)

7	Bir haksızlığa uğrarsam, mahkemelere güvenle başvururum.
8	İhtiyaç duyduğum zaman emniyet teşkilatına başvururum çünkü polise güvenirim.
12	Temel vatandaşlık görevleri, askere gitmek, vergi vermek ve oy kullanmaktır.
16	Vatandaşlık haklarım ihlal edildiği zaman haklarımı nasıl arayacağımı biliyorum.
17	Türkiye daha fazla göçmen kabul etmemelidir.
18	Türkiye'de doğmamış fakat Türkiye'de yaşayan kişiler herkesle aynı haklara sahip olmalıdır.

DEVLETİN GÖREVİ (Dördüncü Faktör-4 madde)

11	Vatandaşlar gerekli gördüklerinde barışçıl protesto gösterileri düzenleyebilmelidir.
19	Herkesi istihdam edecek kadar iş yok ise erkekler, kadınlardan daha çok işe alınmalıdır.
22	Devlet, işsizlere hayatlarını sürdürmeye yetecek kadar maddi destek sağlamalıdır.
23	Devlet görevlilerinin farklı kimlikleri (dinsel, etnik ve mezhepsel kimliklerini) iş yerinde sergilemesine müsaade edilmemelidir.

GENEL HAKLAR (Beşinci Faktör-4 madde)

10	Bir yasa, insan haklarını ihlal etse de insanlar o yasaya uymak zorundadır.
20	Türkiye'de doğmamış fakat Türkiye'de yaşayan kişiler Türkçe öğrenmek zorundadır.
21	Kaliteli bir eğitim, kadınlardan çok erkekler için gereklidir.
26	"Türk Milleti" ülkemizde yaşayan tüm vatandaşların ortak adıdır.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 735–752

https://doi.org/10.21449/ijate.722460

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Spatial Models for Identifying Factors in Student Academic Achievement

Filiz Akbas-Yesilyurt [1], **Huseyin Kocak** [2], **M. Ensar Yesilyurt** [1,*]

[1]Pamukkale University, Department of Economics, Denizli, Turkey
[2]Pamukkale University, Department of Business Administration, Quantitative Methods Division, Denizli, Turkey

ARTICLE HISTORY

Received: Apr. 18, 2020
Revised: Sept. 09, 2020
Accepted: Nov. 09, 2020

KEYWORDS

Student achievement,
Interaction,
Spatial models

Abstract: In the literature, estimation results of the determinants of academic achievement are controversial. There may be several reasons for these controversial results, including sample or cultural differences. Conversely, these results may arise from ignoring certain important facts, such as an interaction effect. Some studies do not consider interactions among students, and some studies may not use effective models. Surprisingly, very few studies have focused on student academic achievement using spatial models, which may be one of the most suitable models for testing interaction effects. In this study, we estimated student achievement using spatial models and a data sets from Turkey. We observed an interaction between students who live in the same neighbourhood and found evidence of an interaction among students in terms of their achievement based on a spatial error model.

1. INTRODUCTION

The question of whether students' academic performance is "contagious" is a very old debate in the literature. For example, Bronfenbrenner’s (1994) ecological systems theory, although not focused solely on student achievement, examines a child’s development within the context of the system of relationships that form the child’s environment. Similarly, some theories and discussions of “social interaction”, “correlation between students”, and the “peer effect” in schools and neighbourhoods have been presented by Manski (2000), Akerlof (1997), Bayer et al. (2008), Bernheim (2004) and Winship et al. (2011). This literature assumes that interaction effects also influence individual behaviour or outcomes due to the characteristics of students (Deitz, 2002) and agrees that students *interact with each other depending on the specific environment in which they live* (Mayer & Jencks, 1989).

Other important issues related to interaction among students are the measurement of interaction and econometric and data issues. For example, Manski (1993 and 2000), Moffitt (2001) and Brock and Durlauf (2001) discussed the estimation of social interaction effects. They suggested that the outcome for an individual in a group is affected by individual characteristics, the outcome of the group and exogenous and predetermined characteristics. However, conventional models cannot capture these effects in estimations that focus on social interaction. Many attempts have been made to capture these effects. For example, some researchers have used the

*CONTACT: M. Ensar Yeşilyurt ✉ meyyurt@pau.edu.tr Pamukkale University, Department of Economics, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

instrumental variable method (Aaronson, 1998), group fixed effects (Evans et al., 1992; Lavy & Schlosser, 2011), experimental formations (Moffitt, 2001; Zimmerman, 2003), specific dummies (Jensen & Harris, 2008; Sykes & Kuyper, 2009; Gould et al., 2009; Aslund et al., 2011; Weinhart, 2014; Gibbons et al., 2013), multilevel models (Brannstrom, 2008), logistic regression (Adejoro, 2016) and peers' features (Brooks-Gunn et al., 1993) to understand, capture and measure peer effects and interaction. However, in studies that do not consider interaction, interactions with some dissimilar characteristics are considered equivalent (Dietz, 2002), which might produce biased results.

How interaction effects can be measured and compared is an important issue (Duncan, 1994; Brannstrom, 2008). Although a powerful theoretical structure is included in the publications that consider interaction effects (Manski, 1993; Elhorst, 2010; Elhorst et al., 2013), spatial models are not common in studies of the effects of student interaction on student achievement Even though dummies represent the effect of similarities, spatial models may represent additional information, such as the type and level of interaction, direct and indirect linkages and a holistic interaction. On the other hand, as noted by Hsieh and Lin (2019), who were among to use spatial models of student achievement, "the literature on social interactions has mainly focused on the influences of peers on behaviors and decisions of an individual, rarely considering the possible formation of social preferences among the network links". Other examples use spatial models to estimate student achievement (Matlock et al., 2014; Zangger, 2016).

In this context, the present study focuses on estimating student achievement by considering interaction effects using data sets for high school students from Turkey. With the aim of investigating the influence of spillover effects on students' achievements, we test the following hypothesis associated with the assumptions using spatial models: We test whether student achievement is affected by neighbouring students' achievements as our first hypothesis and whether students' performance is influenced by neighbouring students' achievements and the performance of individual, environmental and family neighbouring effects as our second hypothesis; for our third hypothesis, after controlling for the observed characteristics, we investigate whether students' achievements are influenced by unobservable factors (Zavarrone & Vitali, 2012).

Since the present research is about Turkey Education System, a brief overview about development in Turkey Education System sheds light on the discussionis and as follows: The transformation of the education system in Turkey is not yet complete, and major changes occur frequently. Some important structures and processes are presented below (these are explained for the time period we analysed; since then, some new changes have occurred): *a)* Prior to 1999, the duration of primary school, secondary school and high school was 5, 3 and 3 years, respectively. This was changed to 4 years each. In 2005, primary school became 8 years long, and high school became 4 years long. Primary school is compulsory. These developments seem to contribute to educational output. For example, in 2013, while average years of schooling was 6.5 expected years of schooling was 11.5 (Yesilyurt et al., 2016). This finding indicates that the changes led to a significant improvement in the number of years that students were educated in schools. Improvements in quality are expected to increase with improvements in quantity. *b)* Because Turkey has a large youth population, students compete to enter university. Many students are eager to gain admission to and acquire a degree from a university. This structure has led to a demand that has surpassed the supply in higher education. After 2000, private universities were founded, and students could enter them more easily by paying a relatively high fee. Public universities are either free or require a very small fee. *c)* Two entrance exams are administered to students, who are ranked by their grade and assigned different types of points based on the public achievement index. In Turkey, university entry is based on general exams. The first-level exam is "Access to Higher Education (AHE)", and the second exam is the "University Attendance Exam (UAE)". The first exam selects and ranks students for the second exam, while the second exam directs students to the appropriate departments depending on their scores, their career plans and a portion of their first exam score. *d)* The scores from these two exams are not unique. Several types of scores constitute achievement in math, science or social science. For example, if students want to enter the faculty of medicine, they must apply with a good science score. In the first step, there are fewer types of scores. All students' scores cover the same questions and are comparable in the same score groups.

The contribution of this study is twofold. First, we use spatial models to test the above hypotheses. The use of spatial models to estimate student achievement is not common and occurs in a very limited number of studies. This approach may contribute to discussion in this field. Second, most studies in the spatial economics literature choose weight matrices subjectively. However, the results can change depending on the weight matrix. In this study, we use the Bayesian approach and traditional tests to choose the best-fitting weight matrix for the data. Recently, a similar idea was effectively used to determine factors in the development of towns in Turkey (Yesilyurt et al., 2020)

In the next sections, we introduce the education system in Turkey, the datasets, and the theoretical structure. The last two sections are dedicated to the estimation results and conclusions.

2. METHOD

In this section, we present the economic theory behind student achievement and the spatial econometric structure.

2.1. Theoretical Structure

The basic structure for determining student achievement is discussed in Hanushek (1979), Todd and Wolpin (2003) and Heck (2009) as

$$SS = \mathrm{f}(STD, FAM, SCH, ENV),$$

where SS, STD, FAM, SCH, ENV indicate the student achievement in AHE, the student's characteristics, the student's family characteristics, the student's school and the student's environment, respectively. The corresponding variable names with these categories are discussed in Section 3.2.

3.1.1. *Econometric model*

Econometric structure and their connection with the hypothesis are given below:

a) The hypothesis and spatial models:

In recent years, spatial interaction has become one of the most investigated topics in economics. In the literature, spatial heterogeneity and spatial dependence are the two sources of spatial relations. Spatial heterogeneity leads to parameter instability, and spatial dependence leads to correlation between the spatial units. Accounting for these effects with standard econometric techniques is fraught with a number of problems. In line with this study, Boarnet (1994) stated that small units create spatial dependence, which requires special treatment. Since the violation of assumptions is profoundly important, modelling spatial interaction properly is important. The basic modelling of the spatial interaction can be endogenous (Y), exogenous (X) and among the error terms (ε). The general model with all the interaction effects (Vega & Elhorst, 2015):

$$Y = \alpha + \delta WY + X\beta + WX\theta + u, \tag{1}$$

$$u = \lambda Wu + \varepsilon,$$

where W is the weight matrix that evaluates the connections between economic units. δ is the spatial autoregressive coefficient, θ is the unknown parameter and λ is the spatial autocorrelation coefficient. WY, WX and Wu indicate endogenous and exogenous interaction effects and interaction effects among the errors, respectively. If λ and θ in the equation are equal to zero, the general model is reduced to a spatial autoregressive model (SAR). If δ and θ are equal to zero, the spatial error model (SEM) will be the true model while λ is equal to zero, the spatial Durbin model (SDM) will be the true model.

Regarding the hypotheses considered in this study, if there is a spatial lag in the dependent variable, it is evidence of spatial autocorrelation (SAR) in students' achievements (i.e., there is confirmation of the first hypothesis). If spatial lags exist in the dependent and independent variables (spatial Durbin model-SDM), there is confirmation of the second hypothesis. If there is spatial lag in the error process (SEM), common unobserved factors affect student achievement, confirming the third hypothesis.

b) Specification tests:

The literature presents different approaches for choosing the correct weight matrix and the model. One of the most well-known and frequently used tests for investigating whether residuals are spatially correlated is Moran's *I* test. Another common approach to test for spatial dependencies is the Lagrange multiplier tests.

Since the weight matrix is introduced the model a priori, some researchers note that building a weight matrix that shows how the contiguity relation is determined is one of the important steps (Paelinck & Klaassen, 1979; Anselin, 1988). Alternatives can be used to achieve this outcome by comparing different weight matrices' parametric indicators (Ertur & Koch, 2007), using a weight matrix that is noted by model diagnostic statistics (Stakhovych & Bijmolt, 2009) or creating proxy variables to integrate spatial relations. However, these approaches may not prevent bias and inconsistencies due to the use of an incorrect weight matrix (Mizruchi & Neuman, 2008; Farber et al,. 2008). In recent years, Bayesian model specification tests have become another method for comparing different weight matrices and spatial models for the best fit. This approach considers the log marginal likelihoods of the models. The highest probability will lead to the appropriate specification, and the estimator is not affected by specification error (Lesage, 2014; Lesage, 2015):

$$prob(M_i|y) = \frac{prob(y|M_i)prob(M_i)}{prob(y)},$$

where $prob(M_i|y)$ indicates posterior model probabilities, $prob(y|M_i)$ is marginal likelihood, and $prob(M_i)$ is the prior probability of model i. If the result of the marginal probability calculation has a value higher than the others, this will support the model selection.

c) Weight matrix:

In spatial models, weight matrices are exogenously imposed. The weight matrix W is a nonnegative matrix that identifies the spatial relations between specified spatial units. The diagonal of the W matrix is zero, which shows the exclusion of a self-neighbouring effect and is nonnegative otherwise. To avoid the singular values of the spatial autoregressive parameter, the weight matrix is normalized. Even though there are other methods, the most commonly used is row normalization. In row normalization, each element of the weight is divided by the sum of the related row of the weight matrix. According to the literature, the misspecification of the weight matrix will yield incorrect estimates. There are different approaches to correct this problem, but there have been no convincing results so far (see Lesage & Pace, 2014). For greater objectivity, we tested different possible types of weight matrices. Some of the most commonly used weight matrices are those based on boundaries and distances, *k*-order binary contiguity matrices and *k*-order neighbour matrices: i) Binary weight matrices constructed on the neighbouring relations of the spatial units as one and zero. ii) *k*-order contiguity matrices are the generalization of the first-order contiguity weight matrix that shows if the neighbours share the same boundaries. In this study, we used a second-order contiguity weight matrix that shows the contiguity relations for the locations of neighbours to the first-order neighbour. iii) Construction of the weight as in many regional studies, we assumed that the space is Euclidean for the determination of the relation. The distance-based matrices and k-order neighbour matrices are constructed with the Euclidean distance between centroids of the two spatial units:

$$d_{mn} = \sqrt{(lat_m - lat_n)^2 + (long_m - long_n)^2},$$

where lat denotes latitude, $long$ denotes longitude and d_{mn} calculates the distance between spatial units m and n. In the distance band weight matrix, the spatial weight matrix is created if m is within the specified distance from n. More formally, $w_{mn} = 1$, when $d_{mn} \leq d_{max}$. Here, d_{max} is the specified distance, which covers many alternatives from 2 km to 30 km for this study. iv) k-order neighbour matrices are constructed on the neighbouring relations between their locations from a given spatial unit. These k-order effects help to capture the effects of all locations that are contiguous. After sorting out the distances between the spatial units, the weight matrix is constructed for a given k. In the study, from 1 to 20-nearest neighbours are used to construct a weight matrix based on whether spatial unit m is in the centroids of given nearest centroids to n, regarding the structure is given in the previous part (iii).

d) Preferred spatial model:

According to the posterior model probabilities, the model is reduced to the SEM. SEM specifications reflecting spatial dependence in the disturbances and spatial error are indicative of omitted (spatially correlated) covariates that would affect inference if left unattended. Therefore, we present only the SEM to save space (for detailed explanations of the other models, please see Elhorst, 2013). If the general model, equation (1), is reduced to the SEM, $\delta \ and \ \theta$ will be zero:

$$Y = \alpha + X\beta + (I - \lambda W)^{-1}u, \qquad (2)$$

$$u = \lambda W u + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I),$$

where λ is the autoregressive parameter for the error term Wu, ε is generally an *i.i.d.* and spatially uncorrelated error term, Y is an N × 1 vector of the dependent variable, X is an N × k matrix of observations of the explanatory variables, β is the vector of parameters, and W, which includes a contiguity relation to the model, is an N × N symmetric matrix whose diagonal elements are 0 and whose off-diagonal elements are 1. If λ is equal to zero, there will be no spatial relations between the errors, and this will yield the ordinary least squares (OLS) model.

Unbiasedness will still hold in the SEM, whereas the regression coefficients will be inefficient if we estimate the model with OLS. In this study maximum likelihood estimation is used:

$$\ln L(\beta, \lambda, \sigma^2 | y, x) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 + \ln|I - \lambda W| - \frac{1}{2\sigma^2}e'e,$$

where $e = (I - \lambda W)(y - x\beta)$. Considering all of the abovementioned economic and econometric structures, we estimated the model as follows:

$$\ln S = \alpha + \beta_1 \ln IN + \beta_2 SE + \beta_3 DM + \beta_4 ME + \beta_5 NF + \beta_6 \ln MR + \beta_7 \ln CS + \beta_8 SH$$
$$+\beta_9 \ln PE + \beta_{10} \ln NT + \beta_{11} TM + \beta_{12} \ln AS + u. \qquad (3)$$

Explanations of abbreviations used in the above equation are given in Table 1. We used MATLAB codes, which are presented on the website of James LeSage (https://www.spatial-econometrics.com), and the codes of Paul Elhorst (https://spatial-panels.com/software/ and Yesilyurt and Elhorst, 2017).

2.2. Data Sets

2.2.1. *Sample*

In this study, we reached out to 1700 of approximately 2800 students who were in the senior class in high schools in Denizli Province. However, some did not have an exam score because they had not taken the exam, and the quality of some of the questionnaires was not sufficient for the analyses. Therefore, we obtained 1083 usable questionnaires. Denizli Province is located southwest of Turkey, and its borders lie between the coordinates 37.841769, 29.043219; 37.724925, 29.150715; 37.772910, 28.995421 and 37.785811, 29.119884. It had 17 towns at the time of this study. Descriptive statistics of variables are given in Table 1.

We administered three types of questionnaires. The first questionnaire covered questions about the students and their families. The second was a questionnaire regarding the students' school qualifications. The third was a questionnaire about their teachers' qualifications.

Table 1. *Descriptive Statistics*

Variables	Number	Average	Standard deviation
Dependent variable			
ln (Entrance exam score-*S*)		5.61	4.02
Explanatory variables			
ln (Family income-*IN*)		7.51	8.01
Education level of mother-*ME* (from 1 to 5)		1.33	0.61
ln (Number of medical report days-*MR*)		2.2	1.94
ln (Class size-*CS*)		3.23	1.95
ln (Previous year's average school score-*PE*)		5.55	3.56
ln (Number of teachers-*NT*)		4.06	4.07
ln (Age of school-*AS*)		3.04	2.73
Supplementary education-SE (=1)	958		
Student lives in dormitory-DM (=1)	126		
Nuclear family-NF (=1)	744		
There is a sports hall at school-SH (=1)	413		
Teacher as manager-TM (=1)	64		

2.2.2. *Variables*

Explanations about variables are as follows:

a) Dependent variable

We used the *Access to Higher Education Equal Weight of Math and Turkish Comprehension (AHE-EW)*[†] score. Because all students answer the same questions on this exam, the scores are comparable. This exam is provided to students when they are in their senior year of high school. The second exam, the UAE, is administered after students have graduated, which is why the results of the second exam may be impossible to collect. Even though we used the scores from the first exam, *both scores represent achievement.* These scores are more objective than the scores from students' high school exams in math, science, etc. Therefore, using the latter scores may not be ideal in every situation because the tests are administered and evaluated by different

[†] Yesilyurt and Say (2016) relied on similar data sets. However, their data were for students who aimed to maximize their *math and science scores*, while the present study uses data for students who aimed to maximize their *math and Turkish comprehension scores equally*. Additionally, the authors used the OLS estimator, while the present study uses spatial models in addition to OLS.

teachers, and these results include the subjective perspective of teachers and may not be comparable for measuring student achievement.

In accordance with the majority of the literature, we used the AHE because it is general, objective, comparative and accepted by society. These types of standard and general test results/scores have been used in several studies, such as Sykes (2009) and Yesilyurt and Say (2016), to represent student achievement.

b) Independent variables

The pioneering report known as the Coleman Report (1966) attempted to understand what factors might affect students' achievement. The results suggest that school characteristics matter less than family background or peer effects. The report revealed the benefits of the integration of different groups of students on student achievement. Following this challenging work, researchers have attempted to understand and explain these effects. However, there is no consensus on the outcome and the directions of the effects.

Following this literature, the variables we tested are as follows:

I. Income of the family: One of the explanatory variables is the income of the family, which plays a role in student achievement. One of the reasons for this is that high-income families invest in resources for their children in early stages, which leads to an income-achievement gap that widens significantly after kindergarten (see Reardon, 2011). In the literature, for example, Duncan and Magnuson (2005) found a relationship between family income and achievement, especially for preschoolers and poor students. In this study, to assess family economic status, family income in Turkish lira is used.

II. The impacts of supplementary education: This is defined as all out-of-school-hours learning and is known as shadow education. It is becoming an important part of student achievement in competitive education systems. In the literature, various types of supplementary education have been found with different effects (see Tansel & Bircan 2006; Maylor et al., 2010; Bray, 2013; Tansel, 2013; Wang & Li, 2018). Supplementary education is common in countries that have nationwide examinations for access to higher levels of schooling (Tansel, 2013). We used a dummy variable for students who attended supplementary education.

III. Dormitory effect: Boarding/dormitory effects for college students are twofold. First, students are away from their homes and feel lonely and stressed, which has a negative effect (see Fisher et al., 1986). Second, these students are more likely to demonstrate non-attendance and tardiness, and the presence of their peers may produce positive effects. Particularly in Turkey, students live in a dormitory for numerous reasons. First, if a student's family is relatively poor but the student is not specifically unsuccessful, the student can live in a dormitory with most of the expenses paid by the government or certain charities. Second, if a student is granted entry to a specific school in another part of the country, he or she may live in a dormitory. In this case, families prefer dormitories, which are considered safe places for their children to live. To test this, we created a dummy variable with the value of 1 if a student lived in a dormitory during their education.

IV. Education level of mother: Researchers have found that family background is both directly and indirectly related to students' academic outcome. For example, Brenfenbronner (1994) stated that the family has vital importance for sustaining children's development. Some studies reveal that university-educated mothers and fathers have the same importance in students' performance (see Johnson et al., 1983; Aslund et al., 2011; Yesilyurt & Say, 2016), whereas some studies conclude that the mother's education has a larger effect on adolescents (see Tansel & Bircan, 2006). To test this structure, the education level of the mother was used as a categorical variable.

V. Family structure: Even though living in a nuclear family is the standard, students may live in the same house with older members of the family or with a sibling's family in some conditions. People may prefer this arrangement because it saves money or preserves social control in the family. Certain families may particularly prefer this arrangement if they have newborn children because other members of the family, such as an aunt, grandmother or grandfather, can take care of the children. Opposite effects may arise with regard to student achievement: The first is the negative effect of a crowded home, and the second is the positive effect of warm, supportive conditions at home. In this study, we observed which effect was valid, specifically in Denizli/Turkey. Parke (2003) found that children reared in nuclear families do better than those in stable blended families; however, when controlling for other effects, this effect becomes weak. To test this structure, we used a dummy variable with the value of 1 if the student lived in nuclear family and 0 otherwise.

VI. Student health: Mental and physical health has a relationship with adolescents' academic outcomes, and the number of days a student is absent due to health issues may be an indicator of an unhealthy situation. This may affect students' performance directly because poor health causes non-attendance and class failure (see Salle & Sanetti, 2016) and indirectly because they cannot form social bonds (see Way et al., 2007; Niehaus et al., 2012). The number of days absent because of health issues can be used to test the effect of student health on achievement.

Additionally, we tested several variables related to s*chool and classroom conditions.* Researchers have attempted to analyse the effect of *school conditions* on the achievement gap. Hopland (2012) found a negative relation between poor building conditions and student achievement in Norway. These variables are as follows:

VII. Class size: Some researchers have found a positive relationship between student achievement and class size (see Urquiola, 2001), whereas others have found no relation (Hoxby, 2000).

VIII. Sports hall in school: High school students are very active, and they need to release their energy. Schools with a sports hall may help students release energy safely (MacGowenn, 2007; Huesman et al., 2007; Murillo & Roman, 2011). We used a dummy variable with the value of 1 if there was a sports hall in the school and 0 otherwise.

IX. Previous year's average school score: We included this variable to determine whether achievement is continuous and to generalize whether students benefit from the school's perspective.

X. The number of teachers in the school: This control variable represents human capital, abundant expert teachers and the substitutability between teachers.

XI. Teacher as manager: This control variable is a dummy variable that represents whether at least one of a student's teachers is a manager at school (1 if so, otherwise 0). Being a manager at school may be a time-consuming job; therefore, if a teacher is a manager, he or she may not focus on education.

XII. Age of school: Studies have also investigated the relationship between the age of the school building and student achievement (O'Neill & Oates, 2001).

3. RESULT / FINDINGS

Before conducting the tests for spatial models, we performed general to specific modelling using the OLS estimator. In these estimations, we eliminated the most insignificant variable each time and estimated the model again until all variables were significant. Estimation procedures after OLS are provided in the following section, and the tests are explained step by step. Then, we performed testing to determine which weight matrix and spatial model best fit the data set based on Bayesian posterior probabilities (BPPs). We applied the test in several steps. In the first step, alternative weight matrices were investigated for different spatial models. In the second step, SAR, SEM and SDM models with their best weight matrices were compared with each other to derive the conclusion.

The weight matrices that we tested for each model were the binary matrix, second-order contiguity matrix, from 1 to 20 nearest neighbours and from 1 km to 30 km of distance band between the students' houses. Since we created many nearest neighbour weight matrices and many distance weight matrices, we estimated the best fit nearest neighbour weight matrix and spatial model (against SEM, SAR and SDM) combination and the best fit distance weight matrix and spatial model combination (against SEM, SAR and SDM) first. Then compared these combinations of distance and neighbour weight matrices to the binary contiguity and second order weight matrix.

According to the tests applied, the best fit nearest weigh matrix and spatial model combination is 16 nearest neighbours for SAR, SEM and SDM. However, for the distance matrices, different weight matrices fit different spatial models. The weight matrix of neighbours within 18 km fits the SAR model, 2 km fits the SEM and 20 km fits the SDM.

Table 2 presents these comparisons. The bold parts in the table show the alternatives tested in this step. The alternatives W3, W4, W5 and W6 were determined in the previous step to be the best fit combinations.

Table 2. *Model selection*

	SAR	SEM	SDM
W1 (binary contiguity)	**0.000**	**0.000**	**0.000**
W2 (second-order contiguity)	**0.000**	**0.000**	**0.000**
W3 (16 nearest neighbour)	0.000	***1.000***	0.000
W4 (neighbours within 18 km)	**0.000**	0.000	0.000
W5 (neighbours within 2 km)	0.000	**0.000**	0.000
W6 (neighbours within 20 km)	0.000	0.000	**0.000**

According to the BBP test that compares the best weight matrices for distance and nearest neighbours in addition to binary and second-order contiguity, 16 nearest neighbour weight matrices with SEM gives the best-combination for the data set. Therefore, we estimated spatial regressions using W3 with SEM.

The use of the 16 nearest neighbour matrices instead of a distance matrix may be understandable because it is usually not possible to restrict the interaction of people based on distance. Occasionally, certain neighbours cannot interact and encounter one another despite being very close because some streets do not intersect.

After the selection of the best model and the weight matrix, another important issue is whether there is any misspecification problem in the model. Lesage and Pace (2014) proposed a Hausman test to test whether the OLS and SEM estimates are significantly different. This test is used in the case of inefficient but consistent OLS parameter estimates and efficient estimates of SEM. To be sure that SEM is superior to OLS, we used this test. The results indicate whether there is misspecification in the model. We applied this test, and its result of 26.24 with a 0.19 marginal probability did not diagnose any problem; therefore, we could estimate the SEM. According to the regression results, even though many parameters were similar, there are some differences. For example, income was significant at the 0.1 in OLS but it is insignificant in SEM; living in the dormitory was significant at the 0.1 in OLS, but it is also significant at the 0.05 in SEM. Therefore, the test points SEM, the parameters in SEM should be considered.

Additionally, if the SEM results were ignored, the interaction between students could not be considered to help in understanding the students' reality.

Based on all the processes and tests discussed above, we did not find evidence supporting the first two hypotheses. In other words, SAR and SDM were always rejected in favour of the SEM, and we found evidence of significant spatial dependence.

Before discussing the estimation results of the explanatory variables, we discuss the general importance of interaction among students. The interaction relationships among students were confirmed by econometric tests that allowed us to investigate and discuss two important issues. First, because the SEM considers interactions among error terms, the results of the estimation confirm that there is interaction in terms of the unobservable and/or uncovered factors in the model with regard to student achievement. One of the most difficult issues when analysing social interaction and its effects is that it is nearly impossible for some variables to be measured and observed. However, according to the estimation results, the interaction parameter, *lamda,* is positive and highly significant. As a result, these data sets suggest that there is dependence in the disturbance process. Second, the SEM is able to address this issue because the error term can cover unobserved and unmeasured variables. Understanding these interaction effects may provide an alternative way for policy makers to achieve "optimal organisation of school(s), jobs and neighbourhoods" (Hoxby, 2000). Based on the preferred weight matrix, the SEM estimation results are given in Table 3.

Table 3. *Estimation results*

	SEM	OLS
	Coefficient (t-ratio)	Coefficient (t-ratio)
R^2	0.3424	0.2885
Log-likelihood	721.66	321.96
Constant	3.45 (13.99)***	3.19 (13.15)***
ln (Income-*IN*)	0.01 (1.09)	0.01 (1.64)*
Supplementary education-*SE*	0.13 (7.84)***	0.14 (7.91)***
Student lives in dormitory-*DM*	0.04 (2.27)**	0.03 (1.72)*
Education level of mother -*ME*	0.03 (2.88)***	0.03 (3.27)***
Nuclear family-*NF*	0.03 (2.66)***	0.03 (2.84)***
ln (Number of medical report days-*MR*)	-0.01 (-2.28)**	-0.01 (-2.14)**
ln (Class size-*CS*)	-0.08 (-3.26)***	-0.06 (-2.71)***
There is a sports hall at school-*SH*	0.09 (7.19)***	0.08 (5.93)***
ln (Previous year's average school score-*PE*)	0.36 (7.73)***	0.40 (8.68)***
ln (Number of teachers-*NT*)	0.07 (7.22)***	0.06 (6.19)***
Teacher as manager-*TM*	-0.05 (-1.76)*	-0.05 (-1.87)*
ln (Age of school-*AS*)	-0.04 (-4.65)***	-0.05 (-5.91)***
lamda	0.50 (9.60)***	

*significant at 10%, **significant at 5%, ***significant at 1%.

According to Table 3, average income had a positive effect on student achievement but was insignificant. We expected that this variable would positively contribute to and have a significant effect on student achievement based on the literature (Duncan et al., 1998; Stipek, 1998; Tansel, 2002; Dubow et al., 2009). On the other hand, in countries such as Turkey, councils or other governmental organizations help and transfer money to poor families to ensure balanced income levels.

Another important variable in this literature is supplementary education (Karweit & Slavin, 1981), and this variable was positive and significant in the present study. Supplementary education is one of the most controversial issues in the literature, especially in certain countries, and it is a practice that “has long been ingrained in the cultures of East Asia, and is now increasingly evident in West and Central Asia, in Europe, in North America, and in Africa” (Bray, 2013). For example, Yaylalı et al. (2006) found evidence supporting these finding results for Turkey that were similar to those Kang (2007) observed in Korea. On the other hand, it is ideal for schools to be able to provide knowledge to all students, and an education system should not necessarily be attached to an organization. Additionally, this may indicate that school syllabuses should be revised to improve the motivation and attitude of students.

One of the most attractive variables is living in a dormitory. In Turkey, political/charity organizations in addition to the government establish dorms and invite students to live in them during their education, or higher-quality schools, which are not available in every town, offer dormitory space to students who want to attend these schools. Therefore, families that live in distant villages and towns tend to allow their children to live in dormitories to prevent their students from being spoiled or to enable the students to attend better schools located far from their towns. In these dorms, senior students and other tutors help students with their lessons. Therefore, we found a positive contribution of living in a dorm to academic achievement, as expected.

We used the nuclear family as an explanatory variable for student achievement because the literature provides evidence for this link. Our positive significant results confirmed previous results (Booth & Kee, 2006). When we focus on family structure in Turkey, some cultural issues should be considered. In Turkey, particularly in less developed regions and in distant areas, some parents tend to live with brothers, sisters, fathers and mothers or some parts of the family. Elderly members of a broader family may have noticeable authority over children. This structure may restrict children’s progress and independence and thus their achievement. Conversely, a larger family may be useful because family members help each other. For example, a grandmother can take care of her grandchildren when their mother is away. Additionally, children may thrive in warmer and more supportive conditions. Consequently, according to our data set, living in a nuclear family is a positive significant sign, and it seems that the first argument is more dominant for student achievement.

Another key factor in determining student achievement is parents’ education level. In the literature, the educational level of both parents or one parent usually has a positive effect on student achievement (Brooks-Gunn et al., 1993; Dincer & Uysal, 2010; Aslund et al., 2011; Yesilyurt & Say, 2016). Particularly until the beginning of school, the family contributes to a student’s perspective, understanding and logical progress. Thus, parents can transmit signs, messages and expectations from the modern and improving world to their children, which serve as important leverage for children. In this study, we found that maternal educational level positively impacts student achievement.

We proxied the number of days students did not attend class because of physical or mental health issues, which may decrease students’ capacity. According to the literature, students who are unhealthy are at higher risk for school problems than students who are free from medical problems (Needham et al., 2000; Spernak et al., 2006; Shaw et al., 2015). Mental and physical discomfort and uneasiness restrict student capacity and targets (Fowler et al., 1992). Our results confirm that students with a higher number of missing days had lower levels of achievement.

Another group of variables is school-related factors. These are important because the literature presents some opposite results about the physical conditions of schools, even though the dominant understanding favours better physical factors (Shapson et al., 1980; Correa, 1993; Blatchford & Mortimore, 1994; Akerhielm, 1995). One school-related factor is the school’s age, which is used to measure whether the school culture contributes to student achievement. An older school may have a positive culture and discipline to help students achieve positive outcomes. On the other hand, if the school tradition has degenerated, it may negatively affect students over time. Conversely, a young school may have undisciplined practices, although it may encourage the use of new and useful technologies. A new school with new technology, good conditions and a new perspective may positively contribute to student achievement. According to our results, students in younger schools are more successful. Another school-related variable is class size, which negatively affects student achievement. This finding is understandable because when the class size is large, students may not feel comfortable in the class, and teachers may not care about engaging with students effectively (Hanushek, 1997; Hoxby, 2000; Urquiola 2001). The third school-related factor is that the presence of a sports centre at school is associated with more successful students. High school students are very dynamic and need to release their energy, and a sports centre may be an effective means of doing so (MacGowenn, 2007; Huesman et al., 2007; Murillo & Roman, 2011).

Two teacher-related factors are included in our final model. The first is the number of teachers in school with a significant positive sign. This variable is interesting and may affect students in two ways. First, if there are more teachers in a school, there are likely to be more specialist teachers. Second, competition between teachers is likely to exist. Principals who are responsible to the provincial director of education exert pressure on teachers to be more successful. Therefore, having more teachers in a school may improve teacher performance. The second teacher-related factor is being a manager in a school that requires major effort. If a teacher serves as a manager in the school in addition to being a teacher, he or she may not exhibit maximal teaching performance.

The last variable included in our final models is average university entrance exam score from the previous year, which has a positive significant sign and may determine the sustainability of achievement.

4. DISCUSSION and CONCLUSION

In this study, we estimated student academic achievement and the contribution of spatial models. The spatial weight matrix was determined objectively by utilizing Bayesian criteria. Through testing, we determined that the 16 nearest students and the SEM were the best-fitting combination for estimation.

The interaction parameter of the SEM estimation had a positive and significant result. This provides evidence that interaction is important in terms of students' achievement and investment in the student learning environment. This may be because any improvements to neighbours will contribute to focal students through their interaction with their neighbours' children. As a result, these data sets suggest that student achievement is contagious, and improving these aspects for students is crucial. Therefore, policy makers may want to invest in the learning environment to benefit students and to achieve the interaction effects among students identified in this study.

More specifically, similar to the previous literature, we found contributions of maternal education level, living in a nuclear family and some school-specific factors, such as a young school and small class size. In Turkey, women's mean years of schooling is lower than that of men. It seems that efforts to promote women's intellectual level through formal or informal education would be effective for improving student achievement. Although some literature claims that school-related factors do not affect student achievement as much as other factors, we found evidence that schools that are more comfortable, young, and endowed with new technologies promote student achievement.

Acknowledgements

The authors thank Ron Smith (Birkbeck College), two anonymous reviewers, and Eren Can Aybek (IJATE's Editor) for their useful comments.

This research was supported by Pamukkale University Scientific Research Projects Coordination Unit under grant number 2011SOBE047, for which the authors are indebted.

Declaration of Conflicting Interests and Ethics

The authors declare no conflicts 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).

ORCID

Filiz Akbaş-Yeşilyurt https://orcid.org/0000-0003-1629-4747

Hüseyin Koçak https://orcid.org/0000-0001-9683-6096

M. Ensar Yeşilyurt https://orcid.org/0000-0001-5610-3146

5. REFERENCES

Aaronson, D. (1998). Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *Journal of Human Resources*, *33*, 915-946. https://doi.org/10.2307/146403

Adejoro, O. E. (2016), Does location also matter? A spatial analysis of social achievements of young south Australians [Unpublished Master Thesis]. Department of Physical Geography and Ecosystem Science, Lund University, Lund, 2016.

Akerhielm, K. (1995). Does class size matter?. *Economics of Education Review*, *14*(3), 229-241. https://doi.org/10.1016/0272-7757(95)00004-4

Akerlof, G. A. (1997). Social distance and social decisions. *Econometrica*, *65*(5), 1005-1027. https://doi.org/10.2307/2171877

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Springer Science & Business Media.

Aslund, O., Edin, Per-Anders, Fredriksson, P., & Grönqvist, H. (2011). Peers, neighborhoods, and immigrant student achievement: evidence from a placement policy. *American Economic Journal: Applied Economics*, *3*(2), 67-95. https://doi.org/10.1257/app.3.2.67

Bayer, P., Hjalmarsson, R., & Pozen, D. (2009). Building criminal capital behind bars: Peer effects in juvenile corrections. *The Quarterly Journal of Economics*, *124*(1), 105–147. https://doi.org/10.3386/w12932

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, *102*, 841-877. https://doi.org/10.1086/261957

Blatchford, P., & Mortimore, P. (1994). The issue of class size for young children in schools: What can we learn from the research?. *Oxford Review of Education*, *20*(4), 411-428. https://doi.org/10.1080/0305498940200402

Booth, A., & Kee, H. (2006). Birth order matters: The effect of family size and birth order on educational attainment. *Journal of Population Economics*, *22*(2), 367-397. https://doi.org/10.1007/s00148-007-0181-4

Brannstrom, L. (2008). Making their mark: The effects of neighbourhood and upper secondary school on educational achievement. *European Sociological Review*, *24*(4), 463–478. https://doi.org/10.1093/esr/jcn013

Bray, M. (2013). Shadow education: comparative perspectives on the expansion and implications of private supplementary tutoring. *Procedia - Social and Behavioral Sciences*, *77*(22), 412-420. https://doi.org/10.1016/j.sbspro.2013.03.096

Brock, W. A., & Durlauf, S. N. (2001). Discrete choice with social interactions. *The Review of Economic Studies*, 68(2), 235-260. https://doi.org/10.1111/1467-937X.00168

Bronfenbrenner, U. (1994). Ecological models of human development. *Readings on the Development of Children*, *2*(1), 37-43.

Brooks-Gunn, J., Duncan, G. J., Klebanov, P. K., & Sealand, N. (1993). Do neighborhoods influence child and adolescent development?. *American Journal of Sociology*, *99*, 353-95. https://doi.org/10.1086/230268

Coleman, J. S. (1961). *The adolescent society: the social life of the teenager and its impact on* education. Free Press of Glencoe.

Correa, H. (1993). An economic analysis of class size and achievement in education. *Education Economics*, *1*(2), 129-35. https://doi.org/10.1080/09645299300000019

Dietz, R. D. (2002). The estimation of neighborhood effects in the social sciences: An interdisciplinary approach. *Social Science Research*, *31*(4), 539-575. https://doi.org/10.1016/S0049-089X(02)00005-4

Dincer, M. A. & Uysal, G. (2010). The determinants of student achievement in Turkey, *International Journal of Educational Development*, *30*(6), 592-598. https://doi.org/10.1016/j.ijedudev.2010.05.005

Dubow, E.F., Boxer, P., & Huesmann, L. R. (2009). Long-term effects of parents’ education on children’s educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations, *Merrill-Palmer Quarterly*, *55*(3), 224-249. https://doi.org/10.1353/mpq.0.0030

Duncan, G. J., Yeung, W. J., Brooks-Gunn, J., & Smith, J. R. (1998). How much does childhood poverty affect the life chances of children?. *American Sociological Review*, 63(3), 406-423. https://doi.org/10.2307/2657556

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

Duncan, G J. (1994). Families and neighbors as sources of disadvantage in the schooling decisions of black and white adolescents. *American Journal of Education*, 103(1), 20-53. https://doi.org/10.1086/444088

Elhorst, J. P. (2010). Dynamic panels with endogenous interaction effects when *T* is small, *Regional Science and Urban Economics*, *40*(5), 272-282. https://doi.org/10.1016/j.regsciurbeco.2010.03.003

Elhorst, J. P. (2013). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels.* Springer-Verlag.

Elhorst, P., Zandberg, E., & de Haan, J. (2013). The impact of interaction effects among neighbouring countries on financial liberalization and reform: A dynamic spatial panel data approach. *Spatial Economic Analysis*, *8*(3), 293-313. https://doi.org/10.1080/17421772.2012.760136

Ertur, C., & Koch, W. (2007). Growth, technological interdependence and spatial externalities: Theory and evidence. *Journal of Applied Econometrics*, *22*(6), 1033-1062. https://doi.org/10.1002/jae.963

Evans, W. N., Oates, W. E., & Schwab, R. M. (1992). Measuring peer group effects: A study of teenage behavior. *The Journal of Political Economy*, *100*(5), 966-991. https://doi.org/10.1086/261848

Farber, A., Huu Tu, N., Tran, D., & Vuong, Q. H. (2008). *The financial storms in Vietnam's transition economy: A reasoning on the 1991-2008 period*. Working Papers CEB, ULB.

Fisher, S., Frazer, N. &Murray, K. (1986). Homesickness and health in boarding school children. *Journal of Environmental Psychology*, *6*(1), 35-47. https://doi.org/10.1016/S0272-4944(86)80033-0

Fowler, M. G., Davenport, M. G., & Garg, R. (1992). School functioning of US children with asthma. *Pediatrics*, *90*(6), 939-944.

Gibbons, S., Silva, S., & Weinhardt, F. (2013). Everybody needs good neighbours? Evidence from students' outcomes in England. *The Economic Journal*, *123*(571), 831-874. https://doi.org/10.1111/ecoj.12025

Gould, E., Lavy, V., & Paserman, S. D. (2009). Sixty years after the magic carpet ride: The long-run effect of the early childhood environment on social and economic outcomes. *Review of Economic Studies*, *78*(3), 938-973. https://doi.org/10.3386/w14884

Hanushek, E. A. (1997). Assesing the effect of school resources on student performace: An update. *Educational Evoluation and Policy Analysis*, *19*(2), 141-164. https://doi.org/10.3102/01623737019002141

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, *14*(3), 351-88. https://doi.org/10.2307/145575

Heck, R. H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. Journal of Education Administration, *7*(22), 227-249. https://doi.org/10.1108/09578230910941066

Hopland, A. O. (2012). *School Building Conditions and Student Achievement: Norwegian Evidence*. Working Paper No. 2/2012, Department of Economics, Norwegian University of Science and Technology.

Hoxby, C. M. (2000). *Peer Effects in the Classroom: Learning from Gender and Race Variation*, NBER Working Paper no. 7867, National Bureau of Economic Research.

Hsieh, C. S., & Lin, X (2019), *Social Interactions and Social Preferences in Social Networks*, Working Paper.

https://spatial-panels.com/software/, (15.06.2018)

https://www.spatial-econometrics.com/, (15.06.2018)

Huesman, R L. Jr, Brown, A. K., Lee, G., Kellogg, J. P., & Radcliffe, P. M. (2007). *Modeling Student Academic Success: Does Usage of Campus Recreation Facilities Make a Difference?*. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/159766, (30.03.2017)

Jensen, B., & Harris, M. N. (2008). Neighbourhood measures: quantifying the effects of neighbourhood externalities. *The Economic Record*, *84*(264), 68-81. https://doi.org/10.1111/j.1475-4932.2008.00447.x

Johnson, R. C., Nagoshi, C. T., Ahren, F. M., Wilson, J. R., DeFries, J. C., & McClearn, G. E. (1983). Family background, cognitive ability, and personality as predictors of educational and occupational attainment. *Social Biology*, *30*(1), 86-100. https://doi.org/10.1080/19485565.1983.9988519

Kang, C. (2007). *Does Money Matter? The Effect of Private Educational Expenditures on Academic Performance*, Departmental Working Papers, National University of Singapore, Department of Economics.

Karweit, N., & Slavin, R. E. (1981). Measurement and Modelling Choices in Studies of Time and Learning. *American Educational Research Journal*, *18*(2), 157-171. https://doi.org/10.3102/00028312018002157

Lavy, V., & Schlosser, A. (2011). Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, *3*(2), 1-33. https://doi.org/10.1257/app.3.2.1

LeSage, J. P. (2014). Spatial econometric panel data model specification: A Bayesian approach. *Spatial Statistics*, *9*, 122-145. https://doi.org/10.1016/j.spasta.2014.02.002

LeSage, J. P., & Pace, K. (2014). The Biggest Myth in Spatial Econometrics. *Econometrics*, *2* (4), 217–249. https://doi.org/10.3390/econometrics2040217

LeSage, J. P. (2015). Software for Bayesian cross section and panel spatial model comparison. *Journal of Geographical Systems*, *17*(4), 297-310. https://doi.org/10.1007/s10109-015-0217-3

MacGowenn, R. S. (2007). The *Impact of School Facilities on Student Achievement, Attendance, Behavior, Completion Rate and Teacher Turnover Rate in Selected Texas High Schools*, PhD Thesis, Texas A&M Universities.

Manski, C. F. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives*, *14*(3), 115-136. https://doi.org/10.1257/jep.14.3.115

Manski, C. M. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, *60*(3), 531-542. https://doi.org/10.2307/2298123

Matlock, K., Song, J. J. & Goering, C. Z. (2014). Spatial dependency and contextual effects on academic achievement, *International Journal of Educational Administration and Policy Studies*, *6*(3), 32-42. https://doi.org/10.5897/IJEAPS2013.0327

Mayer, S., & Jencks, C. (1989). Growing up in poor neighborhoods: How much does it matter? *Science*, *243*(4897), 1441–1445. https://doi.org/10.1126/science.243.4897.1441

Maylor, U., Glass, K., Issa, T, Kuyok, K. A., Minty, S. Rose, Anthea. Ross, Alistair, Tanner, E., Finch, S., Low, N., Taylor, E., Tipping, S., Purdon, (2010). *Impact of Supplementary Schools on Pupils' Attainment: An Investigation into what Factors Contribute to Educational Improvements*, Research Reports, London: DSCF, (12.07.2018)

Mizruchi, M. S., & Neuman, E. J. (2008). The effect of density on the level of bias in the network autocorrelation model. *Social Networks*, *30*(3), 190-200. https://doi.org/10.1016/j.socnet.2008.02.002

Moffitt, R. A. (2001). *Policy Interventions, Low-level Equilibria, and Social Interactions,* in Social Dynamics, edited by S. Durlauf and P. Young. MIT Press.

Murillo, F. J., & Román, M. (2011). School infrastructure and resources do matter: analysis of the incidence of school resources on the performance of Latin American students. *School Effectiveness and School Improvement*, 22(1), 29-50. https://doi.org/10.1080/09243453.2010.543538

Needham, B. L., Crosnoe, R., & Muller, C. (2004). Academic failure in secondary school: The inter-related role of health problems and educational context. *Social Problems*, 51(4), 569-586. https://doi.org/10.1525/sp.2004.51.4.569

Niehaus, K., Rudasill, K. M., & Rakes, C. R. (2012). A longitudinal study of school connectedness and academic outcomes across sixth grade. *Journal of School Psychology*, *50*(4), 443-460. https://doi.org/10.1016/j.jsp.2012.03.002

O'Neill, D., & Oates, A. (2001). The impact of school facilities on student achievement, behavior, attendance, and teacher turnover rate in central texas middle schools. *Educational Facility Planner*, *36*(3), 14-22.

Paelinck, J., & Klaassen, L. (1979). *Spatial Econometrics*. Saxon House.

Parke, M. (2003). *Are Married Parents Really Better for Children? What Research Says About The Effects of Family Structure On Child Well-Being*. Research Paper, Center for Law and Social Policy.

Reardon, S. F. (2011). *The Widening Socioeconomic Status Achievement Gap: New Evidence and Possible Explanations*. In Whither opportunity? Rising Inequality and the Uncertain Life Chances of Low-Income Children, R. J. Murnane and G. J. Duncan (Eds.), Russell Sage Foundation.

Shapson, S. M., Wright, E. N., Eason, G., & Fitzgerald, J. (1980). An experimental study of the effects of class size. *American Educational Research Journal*, *17*(2), 141-152. https://doi.org/10.3102/00028312017002141

Shaw, S. R., Gomes, P., Polotskaia, A. & Jankowska, A. M. (2015). The relationship between student health and academic performance: Implications for school psychologists. *School Psychology International*, *36*(2), 115-134. https://doi.org/10.1177/0143034314565425

Spernak, S. M., Schottenbauer, M. A., Ramey, S. L., & Ramey, C. T. (2006). Child health and academic achievement among former head start children. *Children and Youth Services Review*, *28*(10), 1251-1261. https://doi.org/10.1016/j.childyouth.2006.01.006

Stakhovych, A., & Bijmolt, T. H. A. (2009). Specification of spatial models: A simulation study on weights Matrices. *Regional Science*, *88*(2), 389-408. https://doi.org/10.1111/j.1435-5957.2008.00213.x

Stipek, D. (1998). *Motivations to Learn: From Theory to Practice*. 4th edition, Allyn and Bacon.

Sykes, B. (2009). *Spatial Order and Social Position: Neighbourhoods, Schools and Educational Inequalit.* [Unpublished doctoral dissertation]. Amsterdam Institute for Social Science Research

Sykes, B., & Kuyper, H. (2009). Neighbourhood effects on youth educational achievement in the netherlands: Can effects be identified and do they vary by student background characteristics?. *Environment and Planning A*, *41*, 2417-2436. https://doi.org/10.1068/a41255

Tansel, A., & Bircan, F. (2006). Demand for education in Turkey: A Tobit Analysis of Private Tutoring Expenditures in Turkey. *Economics of Education Review, 25*, 303-313. https://doi.org/10.1016/j.econedurev.2005.02.003

Tansel, A. (2002). Determinants of Schooling Attainment for boys and girls in Turkey. *Economics of Education Review*, *21*(5), 455-470. https://doi.org/10.1016/S0272-7757(01)00028-0

Tansel, A., (2013). *Supplemantary Education in Turkey: Recent Developments and Future Prospects*, Koç University-TUSIAD Economic Research Forum Working Paper Series.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, *113*(485), F3-F33. https://doi.org/10.1111/1468-0297.00097

Wang, J., & Li, B. (2018). Supplementary education, student development and education equity: Evidence from primary schools in Beijing-China. *Education Economics*, *26*(5), 459-487. https://doi.org/10.1080/09645292.2018.1460653

Way, N., Reddy, R., & Rhodes, J. (2007). Students' perceptions of school climate during the middle school years: Associations' with trajectories of psychological and behavioral adjustment. *American Journal of Community Psychology*, *40*, 194-213. https://doi.org/10.1007/s10464-007-9143-y

Weinhart, F. (2014). Social housing, neighborhood quality and student performance. *Journal of Urban Economics*, *82*, 12-31. https://doi.org/10.1016/j.jue.2014.06.001

Winship, C., Harding, D. J., Gennetian, L., Sanbonmatsu, L., & Kling, J. (2011). Unpacking *Neighborhood Influences on Education Outcomes: Setting the Stage for Future Research.* In: Opportunity? Rising Inequality, Schools and Children's Life Chances; Edited by Duncan G and Whither, M. R.

Yaylalı, M., Kızıltan, A., Oktay, E., Doğan, E. M., Özer, H., Naralan, A., Özen, Ü., Özçomak, M. S., Akan, Y., & Aktürk, E. (2006). Üniversite Gençliğinin Gelir-Harcama Kalıpları Araştırması. [Survey of Income-Spending Patterns of University Students]. Ataturk University Publication, Erzurum

Yesilyurt, M. E., & Say, D. (2016). Factors affecting success of high school students in Turkey. *Ege Academic Review*, *16*(3), 541-554.

Yesilyurt, M. E., & Elhorst, J. P. (2017). Impacts of neighboring countries on military expenditures: A dynamic spatial panel approach. *Journal of Peace Research*, *54*(6), 777-790. https://doi.org/10.1177/0022343317707569

Yeşilyurt, M. E., Karadeniz, O., Gülel, F. E., Çağlar, A., & Kangallı-Uyar, S. G. (2016). Türkiye'de İllere Göre Ortalama ve Beklenen Okullaşma Yılı, [Mean and expected years of schooling for provinces in Turkey]. *Pamukkale Journal of Eurasian Socioeconomics Studies*, *3*(1), 1-7. https://doi.org/10.5505/pjess.2016.55706

Yeşilyurt, F., Koçak, H., & Yeşilyurt, M. E. (2020). Factors Determining the Development of Minimum Comparable Areas and Spatial Interaction. *Submitted.*

Zangger, C. (2016), The Spatial Structure of Educational Achievement, Researchgate.net.

Zavarrone, E., & Vitali, A. (2012). *School Performance and Network Effects among Classmates*. Population Association of America 2012 Annual Meeting.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: evidence from a natural experiment. *Review of Economics and Statistics*, *85*(1), 9-23. https://doi.org/10.1162/003465303762687677

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 753–772

https://doi.org/10.21449/ijate.693398

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Development of Teachers' Empowerment Scale: A Validity and Reliability Study

Yeliz Ozkan Hidiroglu [1,*], Abdurrahman Tanriogen [2]

[1]Ministry of National Education, Denizli, Turkey
[2]Department of Educational Science, Faculty of Education, Pamukkale University, Denizli, Turkey

ARTICLE HISTORY

Received: Feb. 24, 2020
Revised: Oct. 21, 2020
Accepted: Dec. 04, 2020

KEYWORDS

Teacher empowerment,
Scale development,
Exploratory factor analyzes,
Confirmatory factor analyzes,
Teacher

Abstract: In this research, it is aimed to develop a measurement tool to determine teachers' perceptions about empowerment in a valid and reliable way. The research data were collected from two different teacher groups of 700 people (405 + 295 teachers) who worked in the fall semester of the 2019-2020 academic year. For the content and appearance validity of the scale, seven experts were consulted in the study. Exploratory (EFA) and Confirmatory (CFA) factor analyzes were performed for the construct validity of the scale. As a result of the EFA, a structure with 37 items and 4 factors explaining 69.53% of the total variance was revealed. These factors have been named as “trust”, “status”, “professional development” and “cooperation”. Findings from CFA showed that the 37-item and four-factor structure related to teacher empowerment scale had adequate fit indices. The reliability of the measurements obtained from the teacher empowerment scale and dimensions were examined by Cronbach alpha and omega reliability method and it was determined that the calculated reliability coefficients were within the acceptable limits. Item-total correlations were examined to determine item discrimination. Findings from the item analysis showed that all of the items in the scale are distinctive. Based on these findings, it can be said that the Teacher Empowerment Scale is a measurement tool that produces valid and reliable measurements and can be used to determine teachers' perceptions about empowerment.

1. INTRODUCTION

Teachers play an important role in increasing student achievement and providing conceptual learning by designing and implementing a quality learning process in educational institutions. One of the key elements in most educational reforms is teachers (Fandino, 2010). The quality of a school is based on the quality of teachers working in that institution (Acquaah, 2004). Empowerment of teachers is closely related to the leadership of school administrators and the opportunities they provide to participate in the decision-making (Addi-Raccah, 2009). Because school administrators are the people who facilitate the empowerment of students and teachers in school (Morales-Thomas, 2015).

CONTACT: Yeliz ÖZKAN HIDIROĞLU ✉ yelizozkan09@gmail.com Ministry of National Education, Denizli, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

The concept of teacher empowerment is handled by different researchers with different definitions. According to Sharma (2014), empowering teachers is supporting teachers to become a shaper by supporting their experiences, decision-making authorities and powers, making them feel that they have a real key person in school practices and conditions. Bogler and Somech (2004) define the empowerment of teachers as a process in which teachers deal with their own development and have the ability to solve their own problems. According to Rappaport (1985), empowering teachers is controling their own personality, cognition, and motivation. Zimmerman (2000) argues that empowering teachers is both a process and a result.

The results of empirical research have shown that teacher empowerment generally plays a positive role in educational settings. For example, researchers have found that teacher empowerment increases teachers' job satisfaction (Rice & Schneider, 1994; Rinehart & Short, 1994), professional commitment and organizational citizenship behaviors (Bogler & Somech, 2004), organizational commitment (Somech, 2005), professionalism and self-confidence (Dee, Henkin, & Duemer, 2003) but decreases teachers' professional burnout (Dee et al. 2003). Therefore, it is thought that empowering teachers and awakening their sense of empowerment can lead to many positive organizational behaviors and eventually they can play an important role in teachers' organizational success and stable work (Bogler & Somech, 2004).

The concept of teacher empowerment is handled in different dimensions by different researchers. Wilson and Coolican (1996) consider teacher empowerment in two dimensions as external and internal power. Short and Rinehart (1992) discuss teacher empowerment in six dimensions: "*decision making*", "*professional development*", "*status*", "*self-efficacy*", "*autonomy and influence*". Yin, Jin and Lee (2009) consider teacher empowerment in three dimensions as "*professional development at school*", "*participation in decision-making*" and "*effect of teachers' work on other colleagues*". Al-Yaseen and Al-Musaileem (2015) reviewed teacher empowerment literature (Lichenstein, McLuaghlin & Knudsen, 1991; Lieberman & Miller, 1990; Lightfoot, 1986; Maeroff, 1988; Morris & Nunnery, 1993; Short, 1991; Sizer, 1992; Sprague, 1992) and identified 13 dimensions of teacher empowerment by scanning. These are; (1) accountability, (2) authority, (3) curriculum planning, (4) cooperation, (5) decision making, (6) impact, (7) professional development, (8) professional knowledge, (9) responsibility, (10) self-efficacy, (11) self-esteem, (12) status, and (13) new teacher training. Altınkurt, Türkkaş Anasız and Ekinci (2016) state that the concept of teacher empowerment includes two main dimensions as structural empowerment, which focuses on managerial processes and the regulation of processes, and psychological empowerment that guides teachers' perceptions. In the international literature, it has been determined that the scale of Kanter (1993) for structural empowerment and Spreitzer (1995) for psychological empowerment are frequently used. Kanter (1993) deals with structural empowerment as information, opportunity, resources, support, power and informal power dimensions and explains these dimensions. Spreitzer (1995) on the other hand, discusses and explains psychological empowerment in terms of meaning, effect, competence and autonomy.

When the literature is analyzed, it is seen that there is no common consensus in definitions and classifications about teacher empowerment. It is seen that the most widely used data collection instrument related to the subject is Short and Rinehart's (1992) teacher empowerment scale. The original name of the scale of Short and Rinehart (1992) is "*School Participants Empowerment Scale*". While this scale is considered as "*Teacher Empowerment Scale*" in some studies (Ökmen, 2018; Somech, 2005), in some studies it is considered as "*School Participants Empowerment Scale*" (Bogler, 2005; Bogler & Nir, 2012; Jiang, Li, Wang, & Li, 2019; Lintner, 2008; Sharp, 2009; Squire-Kelly, 2012; Veisi, Azizifar, Gowhary & Jamalinesari, 2015; Watts, 2009). In addition, the scale of Short and Rinehart (1992) was carried out on the Israeli sample in 1992. The "*Teacher Empowerment Scale*" developed by Yin, Jin and Lee (2009) was developed in line with the reform of the curriculum in China and teachers in China were used as a sample. "*Teacher Empowerment Scale*", which is prepared and applied directly for teachers, is not encountered. Apart from this, studies that deal with structural empowerment scales and psychological empowerment scales for teachers are discussed separately. There are two common empowerment scales in the literature: structural empowerment and psychological empowerment. Structural empowerment consists of six dimensions such as "*opportunity*", "*knowledge*", "*resources*", "*support*", "*formal power*" and "*informal power*". Psychological empowerment consists of four dimensions such as "*meaning*", "*effect*", "*competence*" and "*autonomy*". The purpose of this research is to make these two different scale types into a single scale. With this research, it is aimed to develop the "*Teacher Empowerment Scale*" prepared for the teachers directly by the researchers, to make validity and reliability calculations and to present a valid and reliable Teacher Empowerment Scale. This scale can contribute to the development of new ideas on determining the empowerment levels of teachers, revealing the current situation for the position of teachers, and taking measures for possible improvements. In addition, the interactions between teacher empowerment and various variables can be examine. It is thought that this scale will be important in determining how strong the teachers feel, and will contribute to the literature as it is an original scale for teacher empowerment.

2. METHOD

2.1. The Model of Research

This study is a scale development study. In the research conducted on the screening model, information about the sample group, measurement tool and techniques used in data analysis are given below.

2.2. Population and Sampling

2.2.1. Sampling Group 1

In the measuring instrument development process teachers who work in different branches in different regions of Turkey during 2019-2020 academic year were included in the sample. The study was first conducted with 405 teachers. In the research, extreme values were removed and the study was advanced over 368 teachers. An exploratory analysis was conducted by using this sample. In order to look at the multivariate normal distribution, "*Mahalanobis Distance Coefficient*" was examined. According to 62 (*df*) *p* values less than .001 are eliminated. The distribution of the teachers in the sampling group is given in Table 1.

2.2.2. Sampling Group 2

In order to conduct Confirmatory Factor Analysis the 37 items scale was applied again to volunteer teachers actively working in different branches during 2019-2020 academic year. In the second phase 295 teachers participated in the research. After removing outliers the data were subjected to Confirmatory Factor Analysis over 266 teachers. The demographic characteristics of the teachers in the second participant group are shown in Table 2.

Table 1. *Distribution of Teachers according to Demographic Characteristics*

Variables		*n*	*%*
Gender	Female	255	69.3
	Male	113	30.7
	Total	*368*	*100*
School Type	State	342	92.9
	Private	26	7.1
	Total	*368*	*100*
Age	20-30 ages	66	18
	31-40 ages	188	51
	41-50 ages	92	25
	51 age and over	22	6
	Total	*368*	*100*
Branch	Pre School	12	3.3
	Art	74	20.1
	Science-Math	75	20.4
	Classroom	60	16.3
	Social	49	13.3
	Sport	24	6.5
	Foreign Language	50	13.6
	Others	24	6.5
	Total	*368*	*100*
Region	Mediterranian	38	10.3
	East Anatolia	28	7.6
	Aegean	92	25
	South East Anatolia	26	7.1
	Central Anatolia	64	17.4
	Black Sea	32	8.7
	Marmara Region	88	23.9
	Total	*368*	*100*
Seniority	0-5 years	59	16,1
	6-10 years	82	22.3
	11-15 years	91	24.7
	16-20 years	56	15.2
	21 years and over	80	21.7
	Total	*368*	*100*
Working duration in the same school	0-2 years	128	34.8
	3-5 years	124	33.7
	6-8 years	67	18.2
	9 years and over	49	13.3
	Total	*368*	*100*
Educational Status	Two Years Degree	8	2.2
	Bachelor of Science	276	75
	Master's Degree	78	21.2
	PhD Degree	6	1.6
	Total	*368*	*100*

Tablo 2. *Distribution of Teachers according to Demographic Characteristics*

Variables		*n*	*%*
Gender	Female	140	52.6
	Male	126	47.4
	Total	*266*	*100*
School Type	State	242	90.9
	Private	24	9.1
	Total	*266*	*100*
Age	20-30 ages	57	21.4
	31-40 ages	148	55.6
	41-50 ages	43	16.2
	51 age and over	18	6.8
	Total	*266*	*100*
Branch	Science-Math	73	27.4
	Social	83	31.2
	Foreign Language	24	9.1
	Art	20	7.5
	Sport	17	6.4
	Classroom	30	11.3
	Pre School	7	2.6
	Others	12	4.5
	Total	*266*	*100*
Region	Mediterranian	19	7.1
	East Anatolia	12	4.5
	Aegean	145	54.5
	South East Anatolia	14	5.3
	Central Anatolia	31	11.7
	Black Sea	21	7.8
	Marmara Region	24	9.1
	Total	*266*	*100*
Seniority	0-5 years	48	18,1
	6-10 years	77	28.9
	11-15 years	66	24.8
	16-20 years	31	11.7
	21 years and over	44	16.5
	Total	*266*	*100*
Working duration in the same school	0-2 years	90	33.8
	3-5 years	105	39.5
	6-8 years	46	17.3
	9 years and over	25	9.4
	Total	*266*	*100*
Educational Status	Two Years Degree	4	1.5
	Bachelor of Science	206	77.5
	Master’s Degree	52	19.5
	PhD Degree	4	1.5
	Total	*266*	*100*

2.3. Data Collection Instrument and Data Collection

To develop a scale for teacher empowerment, a literature review was conducted on the subject. In accordance with the related literature, individual interviews were held with 32 teachers. Individual interviews play an important role in clarifying the dimensions and deciding the scale items (DeVellis, 2003). Teachers participating in the research were informed in detail about teacher empowerment and teachers were asked to answer the following interview questions accordingly:

"*What do you think is power?*", "*How do you define power?*", "Who *has power in your school? Please explain your answers with their reasons.*", "*What do you think empowerment means?*", "*Who can get you empowered?*", "*In what kind of environments do you feel empowered as a teacher?*"; "*What kind of environment does your manager provide you feel empowered?*"; "*What manager behaviors make you feel empowered as a teacher?*"; "*What kind of training would you like to make you feel empowered as a teacher?* sounding questions were included to find more detailed answers to these interview questions.

The 64 pages written response papers collected from 32 teachers by the researchers were subjected to content analysis. In content analysis, firstly, two researchers created codes in line with the theoretical framework, and then similar codes were grouped and categories were created. In the process of data analysis, these steps proposed by Yıldırım and Şimşek (2005) were followed: naming, coding and extraction, category development, ensuring validity and reliability and reporting.

After this step, a pool of 60 items was created by combining the relevant literature and teacher statements. This items pool was shaped in line with the views of two experts with PhD degrees in educational sciences. These items and the dimensions related to these items were presented to expert opinions in order to ensure scope and appearance validity and necessary arrangements were made in line with the opinions of seven experts who gave feedbacks. Accordingly, arrangements were made in the content and statements of the items and two more items were added. Finally, in order to prevent comprehension and language problems, the items were sent to four Turkish teachers and related corrections were made. Initially, the scale composed of 62 items were applied to 405 teachers who are working in different regions of Turkey. Outliers were discarded from the applied scale and the study was carried out with data gathered from 368 teachers. Kass and Tinsley (1979) state that at least 300 participant should be reached totally. According to Cattell (1978) 200 participants are acceptable and 500 participants are considered to be a very good number in factor analysis studies. Tabachnick and Field (2000) state that in order to make a healthy analysis, the sample should be at least five times of the items in the scale.

2.4. Data Analysis

The construct validity of the Teacher Empowerment Scale was studied. Item total correlation was analyzed as item statistics. Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were performed for construct validity. Cronbach Alpha (α) coefficient was calculated for the internal consistency reliability of the scale. Item total correlations were examined for item discriminations. For EFA, Cronbach Alpha and item discriminations, IBM SPSS Statistics 20.0 and Lisrel 8.7 for CFA were used.

The KMO coefficient and Barlett test result were calculated in order to determine the suitability of the data to factor analysis. The normality test of the dimensions of the scale and the entire scale was performed. The variances explained by the dimensions in the scale and the total explained variances were calculated. Screen plot graph was drawn using Jamovi program. The factors formed as a result of the exploratory factor analysis, items in the factors and factor loading distributions are included. Items with factor loadings below .50 were removed from the scale. The structure revealed by exploratory factor analysis was tested by confirmatory factor analysis. Afterwards, confirmatory factor analysis values and suitability were examined. A second-level confirmatory factor analysis was conducted in order to show that the dimensions of "*professional development*", "*trust*", "*status*" and "*cooperation*" obtained by the first-level confirmatory factor analysis of the teacher empowerment scale together represent the "teacher empowerment" variable as an upper level concept.

As a result of the second level CFA, the factorial model of the scale and standardized coefficients regarding the factor-item relationship were determined. In order to provide item analysis of the scale, item-total correlations were examined and item discrimination indixes were examined. In order to determine the reliability of the scale, Cronbach Alpha and McDonald's Omega (ω) values for the dimensions of the scale and the whole scale were calculated. Discriminant validity and convergent validity values were calculated.

3. RESULT / FINDINGS

In this section, the validity and reliability features of the "*Teacher Empowerment Scale*", which was obtained as a result of the data analysis obtained from the sample group, were emphasized.

3.1. Findings Related to Validity

3.1.1. Findings Related to Exploratory Factor Analysis

It is difficult to fully model the multivariate normal distribution for real life continuous variables (Abbott, 2011). Therefore, in multivariate analysis, it is recommended to perform univariate and multivariate extreme value examinations and then normalize the distributions with the same 'data transformations' at each variable level (Demir, Saatçioğlu & İmrol, 2016). In order to look at the multivariate normal distribution, "*Mahalanobis Distance Coefficient*" was examined. According to *62 df*, *p* values less than .001 are eliminated.

Factor analysis was performed to determine the construct validity of the scale and to determine and dimension the factor loadings of the items. Factor analysis is defined as the process of revealing new concepts (variables) called a factorization or common factorsor obtaining operational definitions of concepts using factor loading values of items (Çokluk, Şekercioğlu & Büyüköztürk, 2016). Factorization and rotation techniques are the concepts to be considered together in factor analysis (Tabachnick & Fidel, 2000). Factor analysis was performed using principal axis factoring and varimax rotation. Here, varimax rotation was preferred in order to obtain a more generalizable factor structure rather than compatibility with the data (Şencan, 2005). In order to determine the suitability of the data for factor analysis, the Kaiser-Meyer-Olkin (KMO) coefficient and the Barlett Sphericity test were calculated (see Table 3). KMO value .96 and Bartlett test result (χ^2 = 12339.121; *p* = .000) were found to be significant.

Table 3. *Teacher Empowerment Scale KMO and Bartlett's Test Statistics*

Kaiser-Meyer-Olkin Measurement of Sample Suitability		.960
Barlett Sphericity Test	*Chi-Square Value*	12339.121
	df	666
	p	.000

The skewness and kurtosis values of this are taken into consideration. According to Karagöz (2016) and Darren and Mallery (2016), the skewness and kurtosis values should be between -2 and +2 for the data to show normal distribution. In this study, skewness and kurtosis values for four factors and the entire scale are given in Table 4.

Table 4. *Skewness and Kurtosis Values Regarding Teacher Empowerment Scale and Dimensions*

	Skewness	*Kurtosis*	*Normality*
Professional Development	-1.013	1.009	Normal Distribution
Trust	-.967	1.049	Normal Distribution
Status	-.024	-.498	Normal Distribution
Cooperation	-.733	1.043	Normal Distribution
Teacher Empowerment	-.735	1.014	Normal Distribution

In the scale with 62 items which item-total correlation values below .50 and overlapping items were eliminated. The final scale consists of 37 items and four dimensions. It is a 5-point Likert type (Strongly Agree, Agree, Partially Agree, Disagree, Strongly Disagree). In Exploratory Factor Analysis, the most frequently used technique regarding the adequacy of the sample size is the sampling adequacy measurement technique of Kaiser-Meyer-Olkin (KMO). Hutcheson and Sofroniou (1999) state that the KMO value being higher than .9 indicates an excellent sample size. In this study, KMO value was calculated as .96. Therefore, it can be said that the sample size is excellent. Although the number of people in the sample is very important for factor analysis, it is known that there are many different opinions in the literature about the number. According to Tabachnick and Fidell (2000), the sample should consist of at least 300 people. Comrey and Lee (1992) argue that 100 people can be considered "*few*", 200 people can be considered "*okay*", 300 can be considered "*suitable*", 500 are considered "very suitable" and over 1000 can be considered "*perfect*". It can be said that as the sample grows, the power of the analysis will increase and the errors will decrease (Yurdabakan & Çüm, 2017). Table 5 presents the variances explained by the dimensions in the scale and the total explained variances.

Table 5. *Total Variance Table*

	Initial Eigenvalues			*Factor Loadings Total Squares*		
Factors	*Total*	*Explained Variance (%)*	*Total Variance (%)*	*Total*	*Explained Variance*	*Total Explained Variance (%)*
1	16.864	% 45.577	45.577	7.718	% 20.860	20.860
2	3.659	% 9.889	55.466	7.529	% 20.347	41.208
3	3.099	% 8.377	63.843	5.181	% 14.004	55.211
4	2.102	% 5.682	69.525	3.963	% 10.710	65.921

According to Table 5, eigenvalues of scale dimensions are 16,864 for factor 1, 3.659 for factor 2, 3.099 for factor 3, and 2.102 for factor 4. The variance explained by the first dimension is 20.860%, the variance explained by the second dimension is 20.347%, the variance explained by the third dimension is 14.004% and the variance explained by the fourth dimension is 10.710%. The scale explains 65.921% of the total variance and has a four-dimensional structure with 2% eigenvalue and 17% variance. Henson and Roberts (2006) stated that the variance rate announced in the scale studies should provide a value of 52% and above. In addition, when the Screen Plot graph is examined, the graph has become horizontal after the fourth vertical line and it is concluded that the scale is four-dimensional (see Figure 1). Screen plot chart was drawn from the Exploratory Factor Analysis menu of Jamovi program. In the Additional Output section of the analysis, options such as Screen plot, and Model fit measures to obtain fit indices similar to interdimensional correlation or structural equation modeling applications are presented (Şahin & Aybek, 2019).

Parallel analysis was used to decide the number of dimensions. The screen plot, is a graphing method to summary the results of parallel analysis. According to the Screen Plot chart, the items in the scale are collected under 4 factors. These factors are determined as "*professional* *development*", "*trust*", "*status*" and "*cooperation*" in line with the theoretical framework. Table 6 presents the items and factor loadings under these factors.

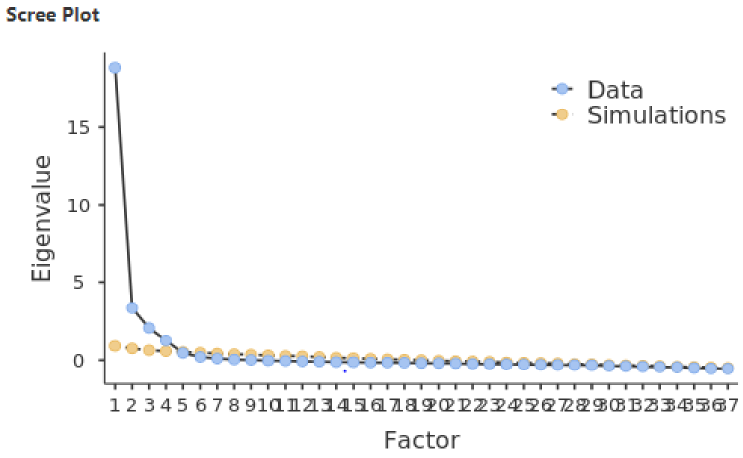


Figure 1. *Exploratory Factor Analysis Output*

Table 6. *Items in Factors and Factor Loading Distributions*

Statements	*Factors*			
	1	2	3	4
S41	.827			
S43	.810			
S40	.783			
S38	.781			
S46	.774			
S44	.757			
S19	.747			
S39	.726			
S56	.697			
S42	.675			
S45	.555			
S53	.554			
S8		.845		
S5		.814		
S9		.790		
S2		.784		
S6		.778		
S3		.773		
S1		.749		
S7		.739		
S4		.681		
S10		.678		
S13		.671		
S32			.799	
S28			.774	
S31			.773	
S30			.752	
S34			.725	
S35			.684	
S26			.683	
S29			.554	
S48				.795
S49				.793
S54				.706
S47				.656
S51				.635
S50				.613

Explanatory factor analysis results are given in Table 6. The first dimension of the scale, "*Trust*" consists of 12 items, the second dimension "*Professional Development*" consists of 11 items, the third dimension "*Status*" consists of 8 items, and the fourth dimension "*Cooperation*" consists of 6 items (see Table 7).

Table 7. *Items in the Trial Form in the Dimensions*

	Dimensions	*Items*
Teacher Empowerment Scale	Trust	19-38-39-40-41-42-43-44-45-46-53-56
	Professional Development	1-2-3-4-5-6-7-8-9-10-13
	Status	26-28-29-30-31-32-34-35
	Cooperation	47-48-49-50-51-54

3.1.2. Findings Related to Confirmatory Factor Analysis

Exploratory factor analysis is the technique of determining how many factors can be generated with the items of the instrument and the nature of relationships among them (Seçer, 2017). An inquiry is made as to whether the indicators collected under certain factors are indicators of the theoretical structure (Green, Salkind & Akey, 1997). The Confirmatory Factor Analysis is based on the examination of a structure determined in the exploratory factor analysis, whether it is verified or not (Seçer, 2017).

While interpreting the EFA results, it was adhered to the rule that the factor loadings that is expected to be theoretically included in any item to remain on the scale should be above .32 (Tabachnick & Fidell, 2000). A higher standard was set for this study and items with factor loadings below .50 were excluded from the scale. In Table 8 below, the equivalents of the scale items in the trial form in the Teacher Empowerment Scale are given.

Table 8. *The Equivalents of the Items in the Trial Form on the Scale*

Trial Form	*Scale*	*Trial Form*	*Scale*
SD1	S1	SD35	S20
SD2	S2	SD38	S21
SD3	S3	SD39	S22
SD4	S4	SD40	S23
SD5	S5	SD41	S24
SD6	S6	SD42	S25
SD7	S7	SD43	S26
SD8	S8	SD44	S27
SD9	S9	SD45	S28
SD10	S10	SD46	S29
SD13	S11	SD47	S30
SD19	S12	SD48	S31
SD26	S13	SD49	S32
SD28	S14	SD50	S33
SD29	S15	SD51	S34
SD30	S16	SD53	S35
SD31	S17	SD54	S36
SD32	S18	SD56	S37
SD34	S19		

CFA was performed to confirm the EFA results and to test the theoretically constructed measurement model. As a result of the confirmatory factor analysis, acceptable fit indices and values of the scale are given in Table 9.

Table 9. *Confirmatory Factor Analysis Values and Fit Indices*

Fit indices	*Value*	*The value of the scale*	*Fitness*	*References*
X^2/sd	Between 0 and 5	3.12	Acceptable	Wheaton, Muthen, Alwin & Summers, 1977
RMSEA	≤ 0.08	0.07	Acceptable	Hooper, Coughlan & Mullen (2008), Sümer (2000)
GFI	Between 0.85 and 1	0.72		Andersen & Gerbing, 1984; Cole, 1987
AGFI	Between 0.80 and 1	0.68		Andersen & Gerbing, 1984; Cole, 1987
CFI	≥ 0.95	0.98	Acceptable	Hu & Bentler (1999), Sümer (2000), Tabachnick & Fidell (2000)
NFI	Between 0.90 and 1.00	0.97	Acceptable	Sümer (2000), Tabachnick & Fidell (2000), Thompson (2008)
NNFI(TLI)	Between 0.90 and 1.00	0.98	Acceptable	Sümer (2000), Tabachnick & Fidell (2000), Thompson (2008)
RMR	≤ 0.08	0.05	Acceptable	Brown (2006), Hu & Bentler (1999)
SRMR	≤ 0.08	0.05	Acceptable	Brown (2006), Hu & Bentler (1999)
IFI	Between 0.90 and 1.00	0.98	Acceptable	Sümer (2000)

As a result of CFA, the structure revealed in EFA was confirmed. The model obtained with CFA is shown in Figure 2. A second-level confirmatory factor analysis was conducted in order to show that the dimensions of "*professional development*", "*trust*", "*status*" and "*cooperation*" obtained by the first-level confirmatory factor analysis of the teacher empowerment scale together represent the "teacher empowerment" variable as an upper level concept. The variances explained by the teacher empowerment variable in the first-level variables were revealed by the analysis. The factorial model of the second level CFA result and the standardized coefficients of the factor-item relationship are given in Figure 2. When the values given in Table 9 are analyzed, GFI and AGFI fit indices indicate that the data are not compatible. However, since the GFI and AGFI indices are affected by the sample size (Aybek & Cikrikci, 2018; Bayram, 2013; Hooper, Coughlan & Mullen, 2008; Raykov & Marcoulides, 2006; Sharma, Mukherjeee, Kumar & Dilor, 2005) and other fit indices are within the acceptable limits, it was concluded that the data collected in the research fit the factor structure of the scale. Since the RMSEA, CFI, NNFI (TLI), SRMR values are within the desired level ranges in the scale, it can be said that the collected data fit the factor structure of the scale.

According to the fit indices obtained, it can be said that the construct validity of the Teacher Empowerment Scale has been confirmed. The Maximum Likelihood (ML) estimation techniques have been used since the variables are measured on an interval scale and have a multivariate normal disribution. Factor loadings range from 0.73 to 0.92 in the "*Trust*" dimension, 0.67 to 0.93 in the "*Professional Development*" dimension, 0.66 to 0.90 in the "*Status*" dimension, and 0.83 to 0.88 in the "*Cooperation*" dimension.

Within the context of the compliance validity study, the correlation values of the four factors related to each other and the entire scale were examined. In order to determine the data analysis technique to be used, it was first examined whether the data showed a normal distribution.

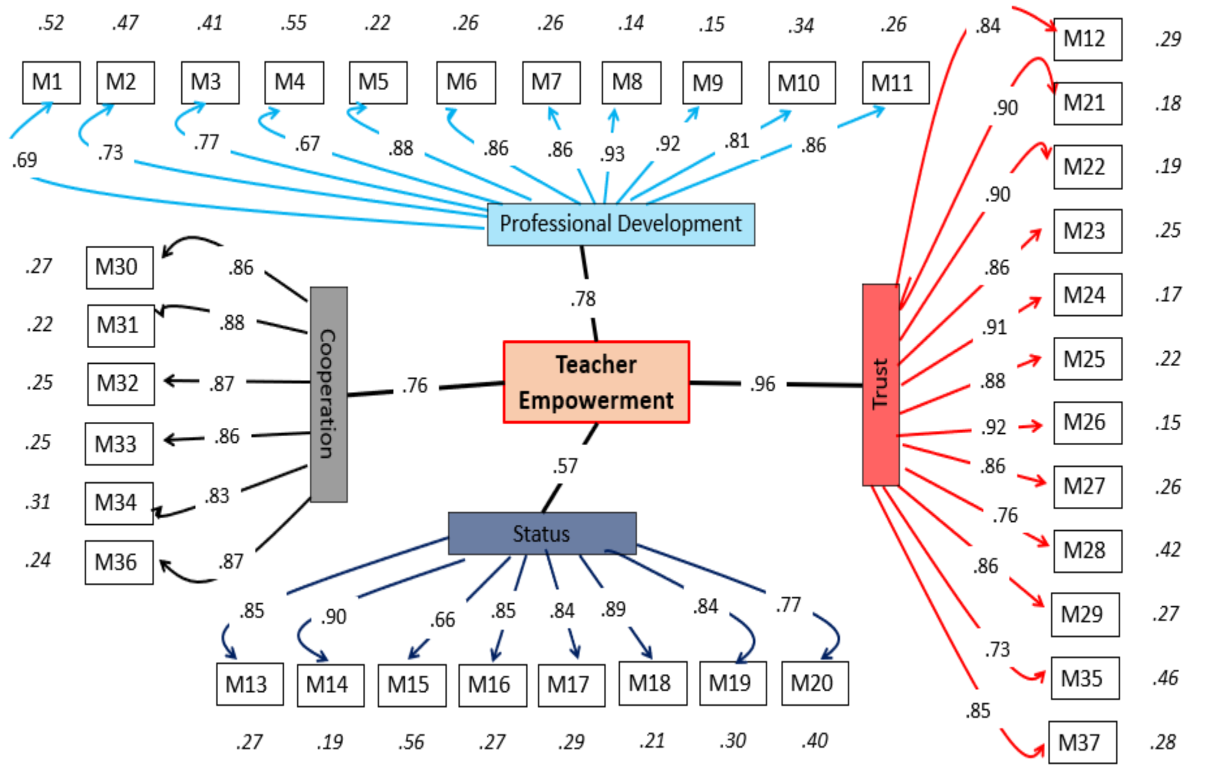


Figure 2. *Measurement Model for Teacher Empowerment Scale*

Factors and scale are normally distributed according to skewness and kurtosis values in Table 9. After the normality test, Pearson Correlation Analysis was performed to determine the correlation coefficients. Correlation coefficients of the four factors related to each other and the entire scale are given in Table 10. As a result of the correlation analysis, it was revealed that the factors had significant relationships with each other and with the entire scale.

Table 10. *Correlation Coefficients between Factors*

Dimensions	*Professional Development*	*Trust*	*Status*	*Cooperation*	*Teacher Empowerment*
Professional Development	1	.758**	.440**	.620**	.861**
Trust		1	.552**	.763**	.932**
Status			1	.520**	.730**
Cooperation				1	.828**
Teacher Empowerment					1

**p<.01

Büyüköztürk (2018) suggests that the correlation coefficient between .70-1.00 as an absolute value is high, that between .30 and .70 is medium, and between .00 and .30 indicates a low level of relationship. Total teacher empowerment score was found to be highly correlated with all dimensions of the scale. When the relations between the dimensions were examined, it was found that the "*professional development*" dimension was highly related to the "*trust*" dimension and a moderately related to "*status*" and "*cooperation*" dimensions. In addition, it was found that the "trust" dimension has a moderate relationship with the "*status*" dimension, a high level relationship with the "*cooperation*" dimension. Finally, it was found that the "*status dimension*" had a moderate relationship with the "*cooperation*" dimension.

Item total correlations were examined in order to achieve item analysis of the Teacher Empowerment Scale. Item total correlations should be greater than .30. Because Field (2005) stated that if the item total correlations were less than .30, that item did not measure the same structure as the other items, meaning that the item showed a weak correlation with the rest of the scale. The mean, standard deviation and item total correlations of the scale are given in Table 11.

Table 11. *Item-Total Statistics*

Item No	$\bar{x}$	*df*	*Item-Total Correlations*	*Item no*	$\bar{x}$	*df*	*Item-Total Correlations*
S1	4.43	.795	.628	*S20*	3.27	.967	.637
S2	4.42	.779	.640	*S21*	4.01	.971	.837
S3	4.24	.896	.720	*S22*	4.01	.948	.866
S4	3.61	1.070	.568	*S23*	3.63	1.133	.783
S5	4.13	.967	.750	*S24*	4.02	.952	.816
S6	3.97	1.026	.706	*S25*	4.04	.965	.803
S7	3.93	1.053	.732	*S26*	4.07	.915	.851
S8	4.09	.954	.787	*S27*	3.63	1.122	.802
S9	4.06	.964	.803	*S28*	3.70	1.054	.770
S10	3.86	1.052	.728	*S29*	3.63	1.119	.779
S11	3.79	1.057	.779	*S30*	3.80	.934	.704
S12	4.03	.990	.797	*S31*	3.69	.996	.738
S13	3.08	1.244	.598	*S32*	3.76	.908	.769
S14	3.11	1.103	.629	*S33*	3.44	1.023	.715
S15	3.68	1.010	.567	*S34*	3.76	.908	.648
S16	3.17	1.111	.571	*S35*	3.44	1.023	.717
S17	3.49	1.103	.606	*S36*	3.91	.855	.766
S18	3.21	1.120	.591	*S37*	3.96	.965	.811
S19	3.30	.984	.619				

If item discrimination index values are above .30, it means that item discrimination is very good (Büyüköztürk, Çakmak, Akgün, Karadeniz & Demirel, 2010; Crocker & Algina, 1986). Accordingly, it can be said that the Teacher Empowerment Scale consists of items with high discrimination.

3.1.3. Findings Related to Reliability

Cronbach's Alpha and Omega Reliability methods were used to determine reliability levels in the study. Büyüköztürk (2006); Erkuş (2014); Field (2005); Fornell and Larcker (1981); Nunnaly and Bernstein (1994); Karagöz (2016) and Seçer (2017) stated that the scale will be accepted as reliable when the Cronbach Alpha value is .70 and above. In the context of internal consistency, Cronbach Alpha analysis is not considered sufficient in case of multiple factor structures. It is also recommended to calculate the Omega Reliability coefficient (Dunn, Baguley & Brunsden, 2014). The results related to the reliability analysis of the scale are given in Table 12.

Table 12. *Reliability Values of Teacher Empowerment Scale*

Dimensions	*Cronbach's Alpha*	*McDonald's Omega*
Professional Development	.956	.957
Trust	.970	.971
Status	.944	.945
Cooperation	.946	.946
Total Scale	.973	.974

Reliability values for the dimensions of the scale and the total scale were calculated using the Jamovi program. It is also recommended to calculate the Omega Reliability coefficient (Dunn, Baguley & Brunsden, 2014). The total McDonald's ω value of the scale was calculated as 0.974 (with the Jamovi program), and the cronbach's alpha value was calculated as 0.973. The reliability of the dimensions of the scale is McDonald's ω value=0.957 for professional development dimension, Cronbach's alpha value is 0.956; McDonald's ω value for the trust dimension=0.971, Cronbach's alpha value is 0.970; McDonald's ω value for the status dimension was calculated as 0.945, the Cronbach's alpha value as 0.944, and for the cooperation dimension as 0.946, the cronbach's alpha value was calculated as 0.946. It can be interpreted that the omega coefficient is more reliable than the alpha coefficient, and according to these results, the reliability of the whole scale and all four sub-dimensions is high (Peters, 2014).

3.1.4. Evaluation of Scores from the Teacher Empowerment Scale

There are 37 items in the Teacher Empowerment Scale (see A1 Table 1). 5-point Likert type was used in the scale such as "*I strongly disagree (1), I disagree (2), I partially agree (3), I agree (4), I strongly agree (5)*". The scale is four-dimensional: "*professional development, trust, status and cooperation*". There are no inverse items in the scale. "*Professional Development*" dimension should be minimum 12, maximum 60; "*Trust*" dimension is minimum 11 and maximum 55; Minimum 8 and maximum 40 in "*status*" dimension; Minimum 6 and maximum 30 points can be obtained in the "*cooperation*" dimension. A total score can be obtained from the entire scale. The increase in the scores obtained from the Teacher Empowerment Scale means that teachers' perceptions about empowerment are at a high level.

For the convergent validity of the scale, the analysis of the Average Variance Extracted (AVE) values of each factor was determined by comparing the correlation of each factor with each other (see Table 13; Fornell & Larcker, 1981). Discriminant validity was evaluated by comparing the square root value of the variance explained with the square of correlations between factors. Convergent and distinctive validity is another type of validity used in testing and verifying the established model (Fornell & Larcker, 1981; Malhotra, 2011). Convergent validity of the measurement model can be evaluated with Average Variance Extracted (AVE) and Combined Reliability (CR). Acceptable value of CR and AVE is 0.70 (Fornell & Larcker, 1981) and value of AVE and CR of this scale is above 0.70. Also, the CR value should be greater than the AVE value (Gouveia & Soares, 2015; Raykov, 1997). AVE and CR values are presented in Table 13. CR and AVE values were calculated using the Excel program. In this study, CR value was calculated as 0.999, AVE value as 0.948. When CR and AVE values of dimensions are examined, it was calculated as CR=0.999 and AVE=0.948 for professional development; CR=0.996 and AVE=0.949 for trust; CR=0.994 and AVE=0.946 for status; CR=0.990 and AVE=0.944 for cooperation. It is seen that the entire scale and dimensions have CR and AVE values over 0.70. Therefore, it can be said that discriminant validity and convergent validity are provided. All these findings show that the data obtained are compatible with the structure revealed by EFA.

Table 13. *CR and AVE values*

Dimensions	*CR*	*AVE*
Professional Development	.999	.948
Trust	.996	.949
Status	.994	.946
Cooperation	.990	.944
Total Scale	.999	.948

4. DISCUSSION and CONCLUSION

When the national and international literature on teacher empowerment is examined, no scale was found to directly determine the perceptions of teachers about empowerment levels. This research is thought to be important in terms of filling this gap in the literature. With this research, a valid and reliable measurement tool for teacher empowerment was tried to be developed. While preparing the teacher empowerment scale, opinions of the teachers were taken first, codes and categories were determined in line with these opinions, and scale items were written in line with the literature for these codes and categories. Scale items were submitted to expert opinions to ensure scope and appearance validity. In line with the opinions of experts, arrangements were made in the item content, dimensions and expressions and two items were added to the scale. Thus, a draft measuring tool with 62 items was obtained. The items in the scale were applied to sample 1.

EFA and CFA were used to test the construct validity of the teacher empowerment scale. As a result of the EFA, a four-factor structure consisting of 37 items explaining approximately 70% of the total variance was obtained. The first factor was named as "*professional development*", the second factor as "*trust*", the third factor as "*status*" and the fourth factor as "*cooperation*" considering the item contents and theoretical structure collected in the factors. CFA was conducted to determine whether the theoretically designed model was verified by the data. The data obtained from the CFA showed that the fit indices of the four-factor structure related to teacher empowerment were sufficient.

The reliability of the measurements obtained from the teacher empowerment scale was examined by Cronbach Alpha and Omega Reliability methods. Cronbach Alpha reliability of the measurements was calculated as .956 in professional development dimension, .970 in trust dimension, .944 in status dimension and .946 in collaboration dimension. The total reliability of the scale is .973. Measurements with a reliability coefficient of .70 and above are considered reliable (Büyüköztürk, 2006; Durmuş, Yurtkoru & Zinc, 2016; Field, 2005; Fornell and Larcker, 1981; Karagöz, 2016; Nunnaly & Bernstein, 1994; Seçer, 2017 and Tezbaşaran, 1997). Omega reliability of the measurements was calculated as .957 in professional development dimension, .971 in trust dimension, .945 in status dimension and .946 in collaboration dimension. The total reliability of the scale is .974. Item analysis was conducted in order to determine the total score predictive power of the items in the teacher empowerment scale and to determine the discrimination levels. Within the scope of item analysis, the corrected item total correlations were examined. CR and AVE values were calculated using the Excel program. In this study, Combined Reliability (CR) value was calculated as 0.999, Average Variance Extracted (AVE) value as 0.948. Therefore, it can be said that discriminant validity and convergent validity are provided. When CR and AVE values of dimensions are examined, it was calculated as CR=0.999 and AVE=0.948 for professional development; CR=0.996 and AVE=0.949 for trust; CR=0.994 and AVE=0.946 for status; CR=0.990 and AVE=0.944 for cooperation. It can be said that discriminant validity and convergent validity are provided.

It is suggested that researches should be carried out to reveal the existing situation regarding the empowerment of teachers, to determine which variables affect the teacher empowerment and from which variables teacher empowerment is affected. Conducting researches in which teacher empowerment scale will be used is important in terms of contributing to the scale's measuring power and intended use.

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).

ORCID

Yeliz ÖZKAN HIDIROĞLU https://orcid.org/0000-0002-5176-1235
Abdurrahman TANRIÖĞEN https://orcid.org/0000-0002-5491-3273

5. REFERENCES

Abbott, M. L. (2011). *Understanding educational statistics using Microsoft Excel and SPSS.* United States: John Wiley & Sons, Inc.

Acquaah, M. (2004). Human factor theory, organizational citizenship behaviors and human resources management practices: An ıntegration of theoretical constructs and suggestions for measuring the human factor. *Review of Human Factor Studies Special Edition, 10*(1), 118-151.

Addi-Raccah, A. (2009). Between teacher' empowerment and supervision: A comparison of school leaders in the 1990s and the 2000s. *Management in Education, 23*(4), 161-167.

Al-Yaseen W. S., & Al-Musaileem M. Y. (2015) Teacher empowerment as an important component of job satisfaction: a comparative study of teachers' perspectives in Al-Farwaniya District, Kuwait, Compare. *A Journal of Comparative and International Education, 45*(6), 863-885.

Altınkurt, Y., Türkkaş Anasız, B., & Ergin Ekinci, C. (2016). Öğretmenlerin yapısal ve psikolojik güçlendirilmeleri ile örgütsel vatandaşlık davranışları arasındaki ilişki. *Eğitim ve Bilim, 41*(187), 79-96.

Anderson, J. C. & Gerbıng, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155-173.

Aybek, E. C., & Çıkrıkçı, R. N. (2018). Kendini değerlendirme envanterinin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliği, [Applicability of the self assessment ınventory as a computerized adaptive test]. *Türk Psikolojik Danışma ve Rehberlik Dergisi, 8*(50), 117-141.

Bayram, N. (2013). *Yapısal eşitlik modellemesine giriş AMOS uygulamaları*, 2. edition. İstanbul: Ezgi Kitabevi.

Bogler, R., & Nir, A. E. (2012). The importance bof teachers percieved organizational support to job satisfaction: What empowerment got to do it. *Journal of Education Administration, 5*(3), 287-306.

Bogler, R., & Somech, A. (2004). Influence of teacher empowerment on teacher's organizational commitment, professional commitment and organizational citizenship behavior in schools. *Teaching and Teacher Education, 20*, 277-289.

Brown T. A. (2006) *Confirmatory factor analysis for applied research*. Guilford, New York.

Büyüköztürk, Ş. (2018). *Sosyal bilimler için veri analizi el kitabı, [Data analysis handbook for social sciences]*. Ankara: Pegem Akademi.

Büyüköztürk, Ş., Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2010). *Bilimsel araştırma yöntemleri* (5th edition). Ankara: PegemA Yayıncılık.

Cattell, R. B. (1978). *The scientific use of factor analysis*. New York: Plenum.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55*, 1019-1031.

Comrey, A.L. ve Lee, H.B. (1992) *A first course in factor analysis.* Hillsdale, NJ: Erlbaum. 1992, 22-24.

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

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2016). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL Uygulamaları,* 4th edition. Ankara: Pegem Akademi.

Darren, G., & Mallery P, 2016, *IBM SPSS Statistics 23 step by step a simple guide and reference,* 14th edition. New York: Routledge.

Dee, J.R., Henkin, A.B., & Duemer, L. (2003). Structural antecedents and psychological correlates of teacher empowerment. *Journal of Educational Administration, 41*(3), 257–77.

Demir, E., Saatçioğlu, Ö., & İmrol, F. (2016). Uluslararası dergilerde yayımlanan eğitim araştırmalarının normallik varsayımları açısından incelenmesi. *Current Research in Education, 2*(3), 130-148.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (Vol. 26). Thousand Oaks, California: Sage.

Durmuş, B., Yurtkoru, E.S., & Çinko, M. (2016). *Sosyal bilimlerde Spss'le veri analizi.* İstanbul: Beta.

Erkuş, A. (2014). Psikolojide Ölçme ve Ölçek Geliştirme - I Temel Kavramlar ve İşlemler (2.Basım). Ankara: Pegem Akademi.

Fandiño, Y. J. (2010). Research as a means of empowering teachers in the 21st century. Retrieved from http://educacionyeducadores.unisabana.edu.co/index.php/eye/article/view/1624/2134

Field, A. (2005). *Discovering statistics using SPSS,* 2nd edition. London: Sage Publication.

Firestone, W. A,, & Pennell, J. R. (1993). Teacher commitment, working conditions, and differential incentive policies. *Review of Educational Research, 63*(4), 489-525.

Fornell, C., & Larcker D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50.

Gouveia, V. V., & Soares, A. K. S. (2015). Calculadoras de validade de construto (CVC). João Pessoa, PB: BNCS/ Universidade Federal da Paraíba, [Construct Validity Calculators (CVC)] Retrieved from http://akssoares.com/psicometria/calculadora-vmee-cc

Green, S. B., Salkind, N. J., & Akey, T. M. (1997). *Using SPSS for windows: Analyzing and understanding data*. NJ: Prentice Hall, Inc.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53–60.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit ındexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal. 6*(1), 1–55.

Jiang, Y., Li, P., Wang, J., & Li, H. (2019). Relationships between kindergarten teachers' empowerment, job satisfaction, and organizational climate: A Chinese model. *Journal of Research in Childhood Education, 33*(2), 257-270.

Kanter, R. (1993). *Women and men of the corporation*, 2nd edition. New York: Basic Books.

Karagöz, Y. (2016). *SPSS ve AMOS 23 uygulamalı istatistiksel analizler.* Ankara: Nobel Yayıncılık.

Kass, R. A., & Tinsley, H. E. A. (1979). Factor analysis. *Journal of Leisure Research, 11*, 120-138.

Lichenstein, G., McLaughlin, M., & J. Knudsen. (1991). *Teacher empowerment and professional knowledge*. CPRE research report series. New Brunswick, NJ: Consortium for Policy Research in Education.

Lieberman, A., & Miller, L. (1990), Restructuring schools: What matters and what works. *Phi Delta Kappan, 71*(10), 759–764.

Lightfoot, S. (1986). On goodness of schools: Themes of empowerment. *Peabody Journal of Education, 63*(3), 9–28.

Lintner, J. D. (2008). The relationship between perceived teacher empowerment and principal use of power. Auburn University. ProQuest Dissertations and Theses, Retrieved from http://search.proquest.com/docview/89141781?accountid=15099

Maeroff, G. (1988). *The empowerment of teachers: Overcoming the crisis of confidence.* New York: Teachers College Press.

Malhotra, N. K. (2011). *Pesquisa de marketing: uma orientação aplicada, (6th ed.)* São Paulo: Bookman.

Morales-Thomas, M. (2015). *Practices that promote parent engagement in an urban elementary school: A phenomenological study of Latino parents of English language learners* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3685358).

Morris, V., & Nunnery, J. (1993). Teacher empowerment in a professional development school collaborative: Pilot assessment. technical report 931101. Memphis, TN: Center for Research in Educational Policy, College of Education, Memphis State University.

Nunnaly, J., & Bernstein, I. (1994). *Psychometric theory*. New York: McGraw-Hill.

Ökmen, A. (2018). *Öğretmen güçlendirmeye ilişkin lisede görev yapan öğretmenlerin algıları, [The perceptions of high school teachers regarding teacher empowerment],* Unpublished Master Thesis, Yıldız Teknik University, Social Sciences Institute, Istanbul.

Peters, G. J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist, 16*(2), 56-69.

Rappaport, J. (1985). The power of empowerment language. *Social Policy, 16,* 15–21.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173-184.

Raykov, T., & Marcoulides, G.A. (2006). *A first course in structural equation modeling,* 2nd edition. Mahlah, New Jersey, London: Lawrence Erlbaum Associates.

Rice, E.M., & Schneider, G. T. (1994). A Decade of teacher empowerment: An empirical analysis of teacher involvement in decision making, 1980-1991. *Journal of Educational Administration, 32*(1), 43–58.

Rinehart, J., & Short, P. (1994). Job satisfaction and empowerment among teacher leaders, reading recovery teachers, and regular classroom teachers. *Education, 114*(4), 570–580.

Sahin. M. D., & Aybek, E. C. (2019). Jamovi: An easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education, 6*(4), 670-692.

Seçer, İ. (2017). *SPSS ve LISREL ile pratik veri analizi analiz ve raporlaştırma*. Ankara: Anı Yayıncılık.

Sharma, A. (2008) *Logics of empowerment: Development, gender and governance in Neoliberal India.* Minneapolis: U of Minnesota P.

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research, 58*(7), 935- 943.

Sharp, D.C. (2009). *A study of the relatıonshıp between teacher empowerment and prıncıpal effectıveness.* (Unpublished doctoral dissertation). Baker University, Faculty of Education, USA.

Short, P. (1991). Teacher commitment and job satisfaction: Which comes first? Paper presented at the Annual Meeting of American Educational Research Association, Chicago, IL, April.

Short, P. M., & Rinehart, J. S. (1992). School participant empowerment scale: Assessment of level of empowerment within the school environment. *Educational and Psychological Measurement, 52*(6), 951–960.

Sizer, T. (1992). *Horace's school*. Boston, MA: Houghton, Miffin.

Somech, A. (2005). Teachers' personal and team empowerment and their relations to organizational outcomes: Contradictory or compatible constructs?. *Educational Administration Quarterly, 41*(2), 237–266.

Sprague, J. 1992. Critical perspectives on teacher empowerment. *Communication in Education, 41* (2), 181–203.

Spreitzer, G. (1995). Psychological empowerment in the workplace: Dimensions, measurement and validation. *Academy of Management Journal, 38*(5), 1442-1465.

Squire-Kelly, V. D. (2012). *The relationship between teacher empowerment and student achievemen*t. (Unpublished PhD Dissertation). Georgia Southern University.

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar. *Turkish Psychological Articles, 3*(6), 49–74.

Şencan, H. (2005). Sosyal ve davranışsal ölçümlerde güvenirlik ve geçerlik (Reliability and validity in social and behavioural measurements). Ankara: Seçkin.

Tabachnick, B. G., & Fidell, L. S. (2000). *Using multivariate statistics* (4th edition). New York, NY: Harper & Row.

Tezbaşaran, A. A. (1997). *Likert tipi ölçek geliştirme kılavuzu.* Ankara: Türk Psikologlar Derneği.

Thompson, B. (2008). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* 3rd edition. Washington, DC: American Psychological Association.

Watts, D. M. (2009). *Enabling school structure, mindfulness, and teacher empowerment: test of a theory* (Unpublished doctoral dissertation). Retrieved from ProQuest Dissertations and Thesis database.

Wheaton, B., Muthen, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological Methodology, 8,* 84-136.

Wilson, S., & Coolican, M. J. (1996). Howhigh and lowself-empowered teacherswork with colleagues and school principals. *Journal of Educational Thought, 30*, 99-118.

Veisi, S., Azizifar, A., Gowhary, H., & Jamalinesari, A. (2015). The relationship between iranian efl teachers' empowerment and teachers' self-efficacy. *Procedia - Social and Behavioral Sciences, 185*(13), 437-445.

Yıldırım, A., & Şimşek, H. (2005). *Sosyal bilimlerde nitel araştırma yöntemleri*. Ankara: Seçkin Yayınları.

Yin, H. B., Jin, Y. L., & Lee, C. K. (2009). The impact of trust and empowerment culture on curriculum reform. *Journal of Capital Normal University, 1*, 125–132.

Yurdabakan, I., & Cum, S. (2017). Scale development in behavioral sciences (Based on exploratory factor analysis). *Turkish Journal of Family Medicine & Primary Care, 11*, 108-126.

Zimmerman, M. A. (2000). Empowerment theory: Psychological, organizational and community levels of analysis. In J. Rappaport & E. Seidman (Eds.), *Handbook of community psychology (*pp. 43-63). New York: Kluwer Academic/Plenum.

6. APPENDIX

A1 Table 1. Teacher Empowerment Scale

Dimension	*Item No*	*English Form*
Professional Development	I 01	Participation in seminars/conferences of important people in my profession is not prevented by the school administration.
	I 02	It is not prevented by the school administration to participate in any kind of training related to my branch,
	I 03	Attending personel development courses (drama, diction, personel development, effective communication, etc.) is supported by the school management.
	I 04	I have the chance to receive trainings about immigrant or problem students by the school administration.
	I 05	It is supported by the school administration to receive training on educational technology.
	I 06	I have the chance to receive trainings on new teaching methods and techniques by the school administration.
	I 07	I have the chance to participate in in-service trainings frequently and regularly by the school administration.
	I 08	It is supported by the school administration to participate in scientific training in my environment.
	I 09	It is supported by the school administration to receive trainings on classroom management.
	I 10	I have chance to participate in training (legislative training) where my Powers and rights are taught.
	I 11	The school administration provides me wqith an environment to attend the courses and trainings I need.
Trust	I 12	I feel that my administrators value me a a teacher.
	I 21	My administrators have understanding towards me.
	I 22	My administrators are supportive of my profession.
	I 23	My administrators behave fairly within the school.
	I 24	I have a healty dialogue with my administrators.
	I 25	My administrators contact me individually when there is a problem.
	I 26	My administrators respect me.
	I 27	My administrators apply school rules in the same way to everyone.
	I 28	I feel free while carrying out my duties.
	I 29	My administrators treat me empathically.
	I 35	Our administrators do not let our time g oto waste with unnecessary works.
	I 37	My administrators give me the opportunity to say my thoughts.
Status	I 13	I think I have a profession with a high social status.
	I 14	The teaching profession provides me with the social status I desire in my environment.
	I 15	Teaching makes it possible for me to deal with many cultural issues..
	I 16	The attitudes of people around me towards teachers make me strong.
	I 17	The teaching profession gives me confidence.
	I 18	The teaching profession gives me dignity.
	I 19	People around me respect the teaching profession.
	I 20	Teachers are well accepted by people in this area.
Cooperation	I 30	The teachers in our school cooperate with each other in linewith their Professional goals.
	I 31	Our school has a teaching staff to work with pleasure.
	I 32	The cooperation of the teachers in our school makes me feel safe.
	I 33	Other teachers at our school appreciate my work.
	I 34	I think that the teacher I work with have Professional ethics.
	I 36	I have a chance to cooperate with other teachers at my school.

International Journal of Assessment Tools in Education
2020, Vol. 7, No. 4, 773–799
https://doi.org/10.21449/ijate.778864
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Comparison of Classification Performances of Mathematics Achievement at PISA 2012 with the Artificial Neural Network, Decision Trees and Discriminant Analysis

Emre Toprak [1,*], **Selahattin Gelbal** [2]

[1]Erciyes University, Faculty of Education, Department of Educational Sciences, Kayseri, Turkey
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: Aug. 10, 2020
Revised: Sep. 29, 2020
Accepted: Dec. 04, 2020

KEYWORDS

Artificial Neural Netwok,
Decision Trees,
Discriminant Analysis,
PISA,
Student Achievement

Abstract: This study aims to compare the performances of the artificial neural network, decision trees and discriminant analysis methods to classify student achievement. The study uses multilayer perceptron model to form the artificial neural network model, chi-square automatic interaction detection (CHAID) algorithm to apply the decision trees method and linear discriminant analysis. The performance of each method has been investigated in different sample sizes when classifying into different numbered subgroups. The study has revealed that the artificial neural network has the best performance in large, medium and small sample sizes when classifying into six, three and two subgroups. In the very small sample size, which has homogeneous variance-covariance matrices, the discriminant analysis performs the best, while in the very small sample size, which does not have homogeneous variance-covariance matrices, it is the discriminant analysis which performs the best when classifying into six subgroups and the artificial neural network performs the best when classifying into two and three subgroups. Considering the performances of the methods with respect to sample size, it can be concluded that as the sample size gets smaller, the performance of the decision trees method gets worse, whereas the performance of the discriminant analysis method improves. No correlation of this kind has been found with regard to the artificial network method.

1. INTRODUCTION

One of the aims of the studies in the field of educational sciences is to determine the current characteristics of students and draw a road map for their development based on these characteristics. To this end, curriculum specialists try to prepare qualified curricula, education managers try to maintain order and control in the implementation of these, and educational psychologists or guidance experts try to provide guidance and counseling services where students need them. This system, which targets the development of the students, works implicitly. In order for the system to operate efficiently and for the outputs obtained at the end of the process to be interpreted reliably, the most necessary component is measurement and evaluation. After the curricula are developed, implemented and provided with guidance, it is of

CONTACT: Emre Toprak ✉ emrettoprak@gmail.com Erciyes University, Faculty of Education, Department of Educational Sciences, Kayseri, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

great importance to measure and evaluate the achievement of individuals accurately for the improvement the individual, society and the future of the country.

Following a valid and reliable assessment and evaluation of students, it is ensured that individuals who are qualified and who have the required characteristics are selected and placed in the areas required by the country with an appropriate classification. From this point of view, it can be said that the main purpose of measurement and evaluation studies conducted in education is to analyze whether students have the characteristics related to the subject being measured and to determine their achievement accurately. The answers given by students to the questions prepared for the purpose of measurement lead them to fall into one of the classification groups such as "passed/failed, adequate/inadequate, incomplete/complete, low/medium/high, bad/medium/good/very good". These classifications, which form the basis of the positions which students will have in the future, are seen in all the examinations at national and international level. Today, the High-School Entrance Examination (LGS) classifies students by their performance in placing them in different types of schools in the transition from basic education to secondary education. Similarly, the Higher Education Foundations Examination (YKS) classifies students by their performance in placing them in different education programs in the transition from secondary education to higher education. Also, in examinations such as Language Proficiency Test administered in Turkey (YDS) at the national level and Test of English as a Foreign Language (TOEFL) at the international level conducted for the purpose of assessing language proficiency, student achievements are classified by the performance at certain levels.

There are different variables that affect classification in the studies conducted on the achievement classification of students. These variables, which affect the achievement of students, can be made up of very different data that can be further increased such as students' anxiety about the lessons, attitudes towards the school, feelings of belonging to the school, motivation towards the course, interest in the course, self-efficacy of the course, the influence of the teacher, classroom atmosphere, educational opportunities at home, socioeconomic values, income levels, ages, the number of siblings, time management skills, and academic self-confidence. These data about students are important predictors of their achievement classification (Arslantaş, Özkan, & Külekçi, 2012; İbrahim & Rusli, 2007; Keser & Sarıbay, 2007; Tosun, 2007).

In recent years, predictive studies conducted for the purpose of determining student achievement and the factors that affect this achievement have increased considerably (Altun & Yazıcı, 2013; Anıl, 2008, 2009, 2011; Arslantaş et al., 2012; Çiftçi & Çağlar, 2014; Doğan, 2009; Erdil, 2010; Gelbal, 2008; Özer & Anıl, 2011; Sadi, Uyar, & Yalçın, 2014; Şahin, 2011). Determining the effect size of the factors that are related to the shaping of achievement will also guide which points should be focused on the development and higher achievement. This requires careful selection of the methods used in studies conducted for this purpose and thus ensuring the most accurate prediction and classification. However, it is difficult to classify in groups whose characteristics are similar.

Different applications have been developed in the literature for the purpose of predicting student achievement and classifying accordingly. Each model used in classification applications has its own unique algorithm. Determining the performance of these algorithms in changing conditions will make studies more efficient and increase classification achievement in studies. Comparative evaluation of the algorithms is of great importance in terms of revealing which algorithm is successful in which situations and increasing classification performances (Kuyucu, 2012). When the literature is examined, researchers generally use regression-based methods such as different regression analysis and DA and structural equation modeling in predicting and classifying student achievement and the factors affecting this achievement (Altun & Yazıcı, 2013; Ercan, Işık, & Çakır 2005; Okioga, 2013; Özdemir & Koruklu, 2011; Yıldırım, 2000). Apart from these methods, there are artificial neural networks (ANNs) that form the experience of their information processes with the information they receive from the samples given. ANNs have the ability to produce solutions to many problems by making the same decisions in the face of similar issues (Haykin, 1994; Öztemel, 2012). Compared to traditional approaches, ANNs offer more exciting alternatives (Jain, Mao, & Mohiuddin 1996). Although numerous studies have been conducted with ANNs applications in different fields such as business, statistics, mathematics, biostatistics, economics, medicine, banking, engineering, tourism, agriculture and insurance (Bayru, 2007; Burmaoğlu, 2009; Çuhadar, 2006; Kayıkçı, 2014; Kibar, 2015; Kocadağlı, 2012; Köktürk, 2012; Sabancı, 2013; Şirvan, 2010; Tolon, 2007; Torun, 2007; Yüksek, 2007), studies conducted in the field of education are limited. ANNs applications do not require any statistical assumption and they also achieve successful results with incomplete data. They even produce successful results in situations when the data is defective or multidimensional, nonlinear or has a high probability of error (Çırak, 2012; Öztemel, 2012; Tepehan, 2011). These features make the application superior to regression-based methods and stand out as important advantages that can be preferred in research.

Another classification model, which is frequently used especially in science, has easy installation and interpretation, adapts easily to databases, does not require the assumptions of parametric regression techniques to be met, has high reliability and is therefore commonly preferred, is the decision trees (DTs) method (Chang & Wang, 2006; Pehlivan, 2006). In DTs, the aim is the creation of homogeneous sub-sets of data about the dependent variable as much as possible (Kuyucu, 2012). The DTs method is a classification technique "which does not require the meeting of the assumptions of parametric regression techniques and which can establish the relationships between the dependent variable(s) and independent variables in its space without interfering in the values in the data set" (Chang & Wang, 2006; Yamauchi et al., 2001 as cited in Kayri & Günüç, 2010, p. 2472). In addition, the DTs method's visualization of the independent variables affecting the predicted variable in the form of trees by their levels of significance makes DTs more impressive than traditional regression methods (Hebert et al., 2006 as cited in Kayri & Boysan, 2008).

One of the traditional statistical methods used in classification and prediction studies is the discriminant analysis (DA) method. The DA, which serves the purpose of building models to predict membership to any group in studies, is a statistical method with strong foundations if their assumptions are met (Çokluk, Şekercioğlu, & Büyüköztürk, 2012; Johnson & Wichern, 1992; Kachigan, 1991). This decision or classification rule is commonly based on the maximum likelihood principle, where each observation is assigned to the group in which it has the greatest likelihood of occurrence (Huberty, 1994). The DA, which is one of the multivariate analysis methods, has been used in many studies in the literature (Atar, 2012; Avcılar & Yakut, 2015; Bektaş, 2012; Ceylan, 2009; Çakmak & Kara, 2011; Çankaya et al., 2009; Demircioğlu et al., 2004; Güzeller & Kelecioğlu, 2006; Oğuzlar, 2006; Öztürk, Coşkun, & Dirsehan, 2012; Serinkan & Bardakcı, 2007).

In the literature, there are different studies on binary comparison of the ANNs and DTs methods with the regression-based methods (Burmaoğlu, 2009; Çölkesen, 2009; Köktürk, 2012; Torun, 2007; Tosun, 2007). Studies conducted with ANNs applications can be exemplified as the comparison of the logistic regression analysis with the ANNs (Benli, 2005; Burmaoğlu, 2009; Güneri & Apaydın, 2004; Kurt & Türe, 2005; Naik & Ragothaman, 2004; Ocakoğlu, 2006), the comparison of the DA with the ANNs (Burmaoğlu, 2009) and the comparison of the multivariate regression analysis with the ANNs (Baş, 2006; Brown, 2007; Thigpen, 2000; Yüksek, 2007).

Studies conducted with the DTs method can be exemplified as the comparison of the logistic regression analysis with the DTs (Kuyucu, 2012; Zurada & Lonial, 2005) and the comparison of the ANNs applications with the DTs (Kuyucu, 2012; Tosun, 2007).

However, among the methods of data mining, these methods, which do not have the problem of assumption, which can be easily interpreted and which can easily integrate into database systems, are not used much in the field of educational sciences and a few studies conducted are in the form of binary method comparison. It is considered that the use of these methods in different applications in the field of educational sciences with traditional methods will be useful for researchers. Also, each of the classification methods developed can give different results under different conditions. From this point of view, the main aim of the study is the comparative examination of the classification of performances of the DA method, which is one of the traditional statistical methods, with the methods of ANNs and DTs. In addition to this main aim, when it is considered that both examinations which are conducted for the purpose of measuring cognitive characteristics and groups in which scales used for the purpose of measuring affective characteristics are administered have different sample sizes and that studies of classification are conducted in different subgroup numbers, it is believed that the examination of the performances all the three methods will produce in different sample sizes and in classifying different subgroup numbers will also be useful. Therefore, different sample sizes and different subgroup numbers have been added to the study and the main problem of the study have been determined as "Do the ANNs, the DTs and the DA performances differ in large (126,126), medium (6,186), small (603) and very small samples (102) in classifying student achievements into 6, 3 and 2 subgroups?"

It is thought that comparing these three methods with each other in the research will expand the results of existing studies and contribute to rich discussions on the subject. Also, results to be obtained from the study will form a significant model in that the three methods can be used in different applications in the field of educational sciences. The classifications to be made will be in different sample sizes and on 6, 3 and 2 subgroups separately. This is also significant in terms of revealing which method gives successful results as the number of groups and sample size of the study change.

In this section of the study, the methods used in the comparison are briefly mentioned.

1.1. The Artificial Neural Networks (ANNs)

The ANNs, which were developed based on the characteristics of the working system of the human brain, is a simulation of the biological nervous system. Considering the known structure of the biological nervous system, the ANNs, which consist of intense connections of simple computational elements in order to achieve high performance, can be defined as the generalization of a mathematical model of human perception and biological nerves (Akpınar, 2014, p. 239; Fausett, 1993 as cited in Yakut, 2012, p. 52). Considering that there are approximately ten billion nerve cells and sixty trillion connections in the human brain (Garson, 1998, p. 25), the brain can be regarded as a flawless computer running very fast (Munakata, 2008, p. 7). While Haykin (1999) describes the ANNs as "a parallel distributed processor consisting of simple units with a natural tendency to store information", Zurada (1992) describes the ANNs as systems with physical cells that receive, store and use information" (Sağıroğlu, Beşdok, & Erler, 2003, p. 25).

The ANNs have many features such as "non-linearity, learning, parallelism, adaptability, generalization, working with missing data, tolerance of error, retention of information, pattern recognition" which are effective in starting to use them widely in different fields (Bayru, 2007; Burmaoğlu, 2009; Dikmen, 2001; Haykin, 1994; Kayıkçı, 2014; Köktürk, 2012; Öztemel, 2012; Seven, 1993; Simpson, 1990; Tosun, 2007; Yüksek, 2007).

The ANNs, formed by artificial neural cells coming together, is generally composed of three layers (Demiryürek, 2009; Elmas, 2003; Öztemel, 2012; Tolon, 2007, Tosun, 2007; Yurtoğlu, 2005):

- The input layer has at least one artificial nerve cell and transmits information from the outside world to the intermediate layer.
- The intermediate layers, process the information from the input layer and send it to the output layer. There can be more than one intermediate layer in a network. The hidden layer of the ANNs is called a black box due to the fact that what is happening in this layer can not be fully explained.
- The output layer has at least one artificial nerve cell and processes the information from the intermediate layer, creates the output that the network must produce for the input set presented from the input layer. The output produced is sent to the outside world.

Learning in ANNs is the development of problem-solving ability by assimilating past experiences (Tosun, 2007, p. 41). The basis of learning, defined as the process of improving behavior through the discovery of new knowledge over time, is the process of experience (Simon, 1983 as cited in Öztemel, 2012). This process is accomplished by learning rules that modify or adjust the connection weights of the network depending on the input examples and preferably the outputs of these inputs (Durmuş, 2008, p. 44). In the learning process, it is accepted that the relationship between the inputs and outputs of each example represents the general aspect of the event from different perspectives and thus it is thought that the event is learned from different perspectives with different examples. Only examples are shown to the computer in the learning process and no other prior information is given (Öztemel, 2012).

1.2. The Decision Trees (DTs)

The DTs, which are used to classify data with the values it has by different characteristics, is a method used to divide a data set into small groups over certain decision steps and make the elements that come together in groups more similar to each other after each division process (Berry & Linoff, 2004; Sun & Hui, 2008).

The DTs consist of roots, branches and leaves in the form of a natural tree. In the DTs with a structure similar to a flowchart, each of the attributes is represented by a node. The last structure in the tree is called "leaf", the top structure is called "root" and the structures between them are called "branches" (Akpınar, 2014; Quinlan, 1993 as cited in Özkan, 2013, p. 53). In the DTs application, many questions are asked about the data and the results are trying to be reached based on the answers received. DTs are derived from top-down, general-specific data, starting from the root node (Oğuzlar, 2004). The basic logic in this process is based on the division of the relevant group into two more homogeneous subgroups at each stage. Decision rules are formed based on the answers received during the question and answer process and the classification process starting from the root node continues until nodes or leaves without branches are found, in other words, until a statistically significant difference is reached (Köktürk, 2012; Thomas, 2000).

There are "decision nodes" in the DTs indicating the test used for classification. The decision nodes perform the tests and branch successfully. Each branch of the tree is a candidate for completing the classification process and a decision node is formed when the classification process does not take place. If a class has been reached, at the end of that branch there is a leaf, one of the classes to be determined on the data. These processes start from the top root node and continue until the lowest leaves are reached (Özekes, 2003).

The DTs application operates in an order with a two-tiered structure, the first step is "learning" and the second step is "classifying". A previously known training data in the learning step is analyzed by the classification algorithm in order to form a model. The model learned is shown as classification rules or DTs. In the classification step, the test data is used to determine the accuracy of the classification rules or the DTs. If the accuracy obtained as a result of the analysis is at an acceptable rate, the rules are used to classify the new data. This acceptability ratio is obtained by comparing the known class and the estimated class (Argüden & Erşahin, 2008; Kıran, 2010; Köktürk, 2012; Özekes, 2003; Silahtaroğlu, 2013).

When the DTs algorithms are used for classification, they are generally called classification trees and when they are used for regression purposes, they are called regression trees (Rokach & Maimon, 2008). In order for the DTs algorithms to be applied to a problem, events and objects must be expressed with certain property values and there must be distinctive properties that affect the determination of classes (Altıntaş, 2010).

1.3. The Discriminant Analysis (DA)

The concept of discriminant was first used by James Joseph Sylvester in 1851 (Akpınar, 2014, p. 189). As a statistical method, while the DA, which was first introduced by Fisher (1936), could initially ensure the division of only two groups, today with its increasing computational power, it ensures the division of data series into more categories (Albayrak, 2006; Akpınar, 2014, p. 189).

The DA, which is one of the multivariate analysis methods, is a method that ensures the division of the variables in the X data set into two or more real groups and that derives discrimination functions that enable the optimal assignment of the units to the real groups in the natural environment by considering the p feature to be examined (Özdamar, 2010). The DA (Çokluk et al., 2012), which is used in the models created to predict membership to any group in studies, is used when it is desired to subdivide the p units with known features into subcategories by their characteristics.

What is aimed at the DA is to determine one or more functions consisting of a linear combination of variables that maximize the differences between individuals in groups (Çakmak, 1992). By means of the DA, which has basically two purposes, first, functions that separate the groups from each other are found, and then, through the functions found, it is ensured that a newly observed unit is assigned to one of the groups in such a way that the classification error is minimum (Güzeller & Kelecioğlu, 2006). Due to these two functions of the DA, it has been considered appropriate by some authors to be named differently. For example, if the DA is applied to determine a discrimination function, it refers to as the Descriptive DA, and if it is applied for classification purposes, it refers to the Predictive DA (Özdamar, 2010). The descriptive DA identifies discrimination functions and through these functions, it enables us to determine the discriminating variables which reveal the difference between the groups the most. The predictive DA allows predicting which group a unit whose group is unknown will be included in (Grimm & Yarnold, 1995; Tatlıdil, 1996; Özdamar, 2002 as cited in Güzeller & Kelecioğlu, 2006). The DA, which is the process of revealing the differences between two or more groups by means of discrimination variables, is a broad concept that covers several closely related statistical approaches (Klecka, 1980). The DA has many assumptions such as normal distribution, sample size, variable selection, homogeneity of variance-covariance matrices, extreme values and multiple connections that affect the performance of the discrimination performance.

2. METHOD

This section includes the type of research, the study group, the data collection instruments, data collection process and data analysis.

2.1. The Type of the Research

This research, in which the data obtained from PISA 2012 application is used, is a field research in terms of taking it from real life and daily life and not being artificial due to any intervention in the research (Kaptan, 1995); it is a cross-sectional study in terms of examining the event, which is the subject of the research for a specific time unit only (Büyüköztürk et al., 2008) and it is descriptive research because an existing event or situation is defined as it is (Karasar, 2014).

2.2. Study Group

The number of the subgroups to be classified and the sample sizes were determined in a way that they were interconnected by taking into account the main problem and sub-problems of the research. The achievement levels determined by PISA were utilized in the formation of classification groups and the samples were selected in a way that they fit the classification into 6, 3 and 2 subgroups;

- The classification for the 6 subgroups was made as Level 1 / Level 2 / Level 3 / Level 4 / Level 5 / Level 6. Each achievement level included in the PISA classification constituted a group.
- The classification for the 3 subgroups was made as Lower Level / Medium Level / Upper Level. The Sub-Level Group consists of the students in Level 1 and Level 2 in PISA. The Intermediate Group consists of the students in Level 3 and Level 4 in PISA. The Top-Level Group consists of the students in Level 5 and Level 6 of PISA.
- The classification for 2 subgroups was made as Lower Level / Upper Level. The Sub-Level Group consists of the students in Level 1, Level 2 and Level 3 in PISA. The Top-Level Group consists of the students in Level 4, Level 5 and Level 6 in PISA.

The population of the study consists of 15-year-old students studying in all the countries participating in PISA 2012 Mathematics practice. There is a sample of 485,490 students from all the countries participating in this practice. The distribution of the students in the sample by their achievement levels is given in Table 1.

Table 1. *The distribution of all the students participating in PISA 2012 by their achievement levels*

Level	*f*	*%*
Less than Level 1	69,691	14.4
Level 1	91,369	18.8
Level 2	109,383	22.5
Level 3	99,016	20.4
Level 4	67,520	13.9
Level 5	34,652	7.1
Level 6	13,859	2.9
Total	485,490	100.0

However, by taking into account the independent/predictive variables used for classification purposes, the students with missing data were excluded from the study and the number of the students constituting the largest sample was determined as 126,126. In addition, in order to form 6, 3 and 2 subgroups, the students below the first level were not included in the research. The distribution of the students representing the large sample by their achievement level is given in Table 2.

Table 2. *The distribution of 126,126 people by their achievement levels*

6 Subgroups	*f*	*%*
Level 1	26,292	20.8
Level 2	32,893	26.1
Level 3	30,424	24.1
Level 4	20,917	16.6
Level 5	11,105	8.8
Level 6	4,495	3.6
Total	126,126	100.0

As can be seen in Table 2, the largest sample was selected as 126,126 people for comparison on different sample sizes. In addition, 6,186 medium-sized, 603 small-sized and very small samples of 102 for each where homogeneity was and was not ensured in variance-covariance matrices were prepared.

The systematic sampling method was utilized in the preparation of large, medium and small samples. All the students were ranked from the lowest achievement level to the highest achievement level before making the selection. Following the ranking, a study was conducted in a way that there were 1,000 students at each achievement level and a sample of 6,186 people was first formed. This sample is given in Table 3.

Table 3. *The creation of 6,186 sample and distribution by achievement levels*

6 Subgroups	*f*	*%*	*Systematic sampling*
Level 1	1,012	16.4	Starting from the 1st person, one person in every 26th person
Level 2	1,028	16.6	Starting from the 1st person, one person in every 32th person
Level 3	1,015	16.4	Starting from the 1st person, one person in every 30th person
Level 4	997	16.1	Starting from the 1st person, one person in every 21th person
Level 5	1,010	16.3	Starting from the 1st person, one person in every 11th person
Level 6	1,124	18.2	Starting from the 1st person, one person in every 4th person
Total	6,186	100.0	

Following the ranking, a study was conducted in a way that there were 100 students at each achievement level and a small sample of 603 people was formed. This sample is given in Table 4.

Table 4. *The creation of 603 sample and distribution by achievement levels*

6 Subgroups	*f*	*%*	*Systematic sampling*
Level 1	100	16.6	Starting from the 1st person, one person in every 262th person
Level 2	100	16.6	Starting from the 1st person, one person in every 328th person
Level 3	101	16.7	Starting from the 1st person, one person in every 304th person
Level 4	101	16.7	Starting from the 1st person, one person in every 209th person
Level 5	101	16.7	Starting from the 1st person, one person in every 111th person
Level 6	100	16.6	Starting from the 1st person, one person in every 45th person
Total	603	100.0	

Apart from these samples, in order to see the effect of homogeneity on the performances of the methods in variance-covariance matrices, two very small samples of 102 people where homogeneity was ensured and was not ensured in variance-covariance matrices were prepared. These samples are also given in Table 5 and Table 6.

Table 5. *The formation of very small and homogeneous sample (102 persons) in variance-covariance matrices and distribution by achievement levels*

6 Subgroups	*f*	*%*	*Purposive sampling*
Level 1	100	16.7	17 people
Level 2	100	16.7	17 people
Level 3	101	16.7	17 people
Level 4	101	16.7	17 people
Level 5	101	16.7	17 people
Level 6	100	16.7	17 people
Total	603	100.0	

Table 6. *The formation of very small and homogeneous sample (102 persons) in variance-covariance matrices and distribution by achievement levels*

6 Subgroups	*f*	*%*	*Purposive sampling*
Level 1	100	16.7	17 people
Level 2	100	16.7	17 people
Level 3	101	16.7	17 people
Level 4	101	16.7	17 people
Level 5	101	16.7	17 people
Level 6	100	16.7	17 people
Total	603	100.0	

The sizes and distribution of the samples formed by the classification groups after the selections are given in Table 7.

Table 7. *All the sample sizes and distribution by classification groups*

	6 Subgroups	*f*	*%*	*3 Subgroups*	*f*	*%*	*2 Subgroups*	*f*	*%*
126126	Level 1	26,292	20.8	Low level	59,185	46.9	Low level	89,609	71.0
	Level 2	32,893	26.1						
	Level 3	30,424	24.1	Medium level	51,341	40.7			
	Level 4	20,917	16.6				Top level	36,517	29.0
	Level 5	11,105	8.8	Top level	15,600	12.4			
	Level 6	4,495	3.6						
6186	*6 Subgroups*	*f*	*%*	*3 Subgroups*	*f*	*%*	*2 Subgroups*	*F*	*%*
	Level 1	1,012	16.4	Low level	2,040	33.0	Low level	3,055	49.4
	Level 2	1,028	16.6						
	Level 3	1,015	16.4	Medium level	2,012	32.5			
	Level 4	997	16.1				Top level	3,131	50.6
	Level 5	1,010	16.3	Top level	2,134	34.5			
	Level 6	1,124	18.2						
603	*6 Subgroups*	*f*	*%*	*3 Subgroups*	*f*	*%*	*2 Subgroups*	*f*	*%*
	Level 1	100	16.6	Low level	200	33.2	Low level	301	49.9
	Level 2	100	16.6						
	Level 3	101	16.7	Medium level	202	33.5			
	Level 4	101	16.7				Top level	302	50.1
	Level 5	101	16.7	Top level	201	33.3			
	Level 6	100	16.6						
102	*6 Subgroups*	*f*	*%*	*3 Subgroups*	*f*	*%*	*2 Subgroups*	*f*	*%*

	Level 1	17	16.7	Low level	34	33.3	Low level	51	50.0
	Level 2	17	16.7						
	Level 3	17	16.7	Medium level	34	33.3			
	Level 4	17	16.7				Top level	51	50.0
	Level 5	17	16.7	Top level	34	33.3			
	Level 6	17	16.7						
102 (Unhomogenized)	*6 Subgroups*	*f*	*%*	*3 Subgroups*	*f*	*%*	*2 Subgroups*	*f*	*%*
	Level 1	17	16.7	Low level	34	33.3	Low level	51	50.0
	Level 2	17	16.7						
	Level 3	17	16.7	Medium level	34	33.3			
	Level 4	17	16.7				Top level	51	50.0
	Level 5	17	16.7	Top level	34	33.3			
	Level 6	17	16.7						

2.3. The processing and analysis of the data

In the study, the multilayer perceptron model was used in the creation of the ANNs model, the CHAID algorithm and the linear DA, which is one of the DA types, were used in the application of the DTs method. The dependent variable included in the models consists of 6, 3 and 2 subgroups made up of the students' achievement levels and the independent variables consist of 17 variables used in the classification.

In the selection of an activation function in the ANNs analyses, the hyperbolic tangent function was applied to the cells in the hidden layer and the Softmax function was applied to the cells in the output layer. 70% of the data set in the analyses was chosen as the sample used in education and 30% was selected as the sample used in the test. Since the ANNs determined a different 70% of the data set as the training set and a different 30% as the test set, 50 different experimental analyses were performed. As a result of the 50 different analyses, the performance of the method was revealed by selecting the highest classification percentage obtained by the method.

The SPSS program was used in all the analyses. Before proceeding with the analysis of the data used within the scope of the research, some arrangements were made for the sub-problems of the research. With the arrangements made, it was tried to meet these assumptions of the DA, which has different assumptions. The variables which do not have missing data within the data, which do not have a distribution closest to normal, which do not have extreme values, whose variance-covariance matrices are homogeneous and which do not have multiple connections were tried to be selected. The examinations are presented respectively.

- **Missing data**

Firstly, independent/predictive variables that are thought to be used in the study were examined and the data of the students who did not have missing data in these variables were included in the study. Since the comparison was made according to the sample sizes in the research, the number of data was requested to be large enough. At this point, 17 variables with the best data were determined. In the PISA sample of 485,490 students, the data of the students with missing data in the variables included in the study were cleared and a total of 126,126 students remained.

- **Sample size**

The sample size assumption was met in all the subgroups to be classified. Each subgroup included students at least as many as the number of the independent variables. The size of the subgroups in the samples created and their distribution by the classification groups are given in Table 7.

- **Normal distribution**

When the kurtosis and skewness values of the variables in the study are examined, it can be said that the distributions are close to normal. However, the significance of skewness and kurtosis values is evaluated by dividing by their standard errors. When the values were divided by their standard errors, it was observed that the results obtained were significant and some of the variables did not have a normal distribution in different samples. Considering the state of the variables in all the sample sizes, it was observed that as the sample size decreased, the skewness and kurtosis of the distributions approached the normal distribution; however, the normal distribution assumption was not fully met.

- **Variable Selection**

The selection of excessive and unnecessary variables was avoided and 17 variables to be used in classification were determined. In the selection of the variables, the variables which do not have missing data, which are closest to the normal distribution, which do not have extreme values and which do not have multiple connections were selected. The dependent and independent variables used in the research are given in Table 8.

Table 8. *The dependent and independent variables used in the research*

The dependent variables	*The independent variables*
6 subgroups 3 subgroups 2 subgroups	Math anxiety Attitudes towards school: Learning outcomes Attitude towards school: Learning activities Sense of belonging to school Math teacher's classroom management Cognitive activities in mathematics Class climate Education at home Math Motivation Mathematics interest Math behavior Mathematics self-efficacy Mathematical intention Math teacher support Openness to problem solving Self-perception of mathematics Teacher Student Relations

- **The homogeneity of variance-covariance matrices**

The homogeneity of variance-covariance matrices was tested for each sample size. Comments were made depending on whether the samples met this assumption.

- **Extreme values**

Whether there are extreme values in the data set was examined and all the values belonging to the variables were converted to standard values. When the sample size is 100 or less, if the z-score of any observation is not in the range of (-3, +3) and when the sample size is more than 100, if the z-score of any observation is not in the range of (-4, +4), the observation is the extreme value (Mertler & Vannatta, 2005 as cited in Çokluk et al., 2012). As a result of the examinations carried out, it was determined that there were no extreme values. This assumption was met.

- **Multicollinearity**

Multicollinearity is a problem arising in case that there are very high correlations between variables. The inclusion of variables with multicollinearity in analyses increases errors and weakens the analysis. A correlation of 0.90 and over between two variables indicates that there is multicollinearity (Çokluk et al., 2012). When correlations between variables were examined in different sample sizes, while it was determined that there was no multicollinearity in a large, medium and small sample and in a very small sample which was homogenized, multicollinearity between variables was found in a very small sample which was unhomogenized. An examination of correlation coefficients with two variables is not enough to determine the problem of multicollinearity because the problem is not just that the correlation between two variables is high, but that an independent variable has a high degree of correlation with all the other independent variables. For this reason, different examinations were made to determine the multicollinearity status, and tolerance, VIF and CI statistics (Tabachnick & Fidell, 2007).

The variables do not have multicollinearity in a large, medium and small sample and in a very small sample where homogeneity was ensured in variance-covariance matrices. In addition, while it was established that in a very small sample where homogeneity was not ensured in variance-covariance matrices, the tolerance values belonging to the variables "sense of belonging to school, mathematics motivation, mathematics interest, mathematics teacher's support and teacher-student relationship" were smaller than 0.10 and that the VIF values were higher than 10, it was observed that the CI values were within normal limits.

As a result of the examinations conducted, the independent/predictive variables which are in Table 8 were used. The reasons for the selection of these variables for classification and prediction in the research can be summarized as follows:

1. There is no missing data in variables.
2. They are the variables that have the closest level to the normal distribution even though they can not meet the assumption of normal distribution.
3. The variables do not have extreme values.
4. There is no problem with multiple connections between variables except for a very small sample where homogeneity was not ensured in variance-covariance matrices.
5. The comparisons have been made by creating separate samples in which the assumption of the homogeneity of variance-covariance matrices has been met and has not been met.
6. In addition, these variables are the ones which are frequently used in studies aimed at predicting student achievement in the literature (Altun & Yazıcı, 2013; Anıl, 2008, 2009, 2011; Arslantaş et al., 2012; Çiftçi & Çağlar, 2014; Erayman, 2004; Ercan et al., 2005; Erdil, 2010; Gelbal, 2008; Kaysılı, 2008; Keser & Sarıbay, 2007; Özabacı & Acat, 2005; Özer & Anıl, 2011; Sadi et al., 2014; Şahin, 2011).

3. RESULTS

The results of the methods' classification performance of student achievements into 6, 3 and 2 subgroups in the large, medium, small and very small sample sizes are given in Figure 1, Figure 2 and Figure 3.

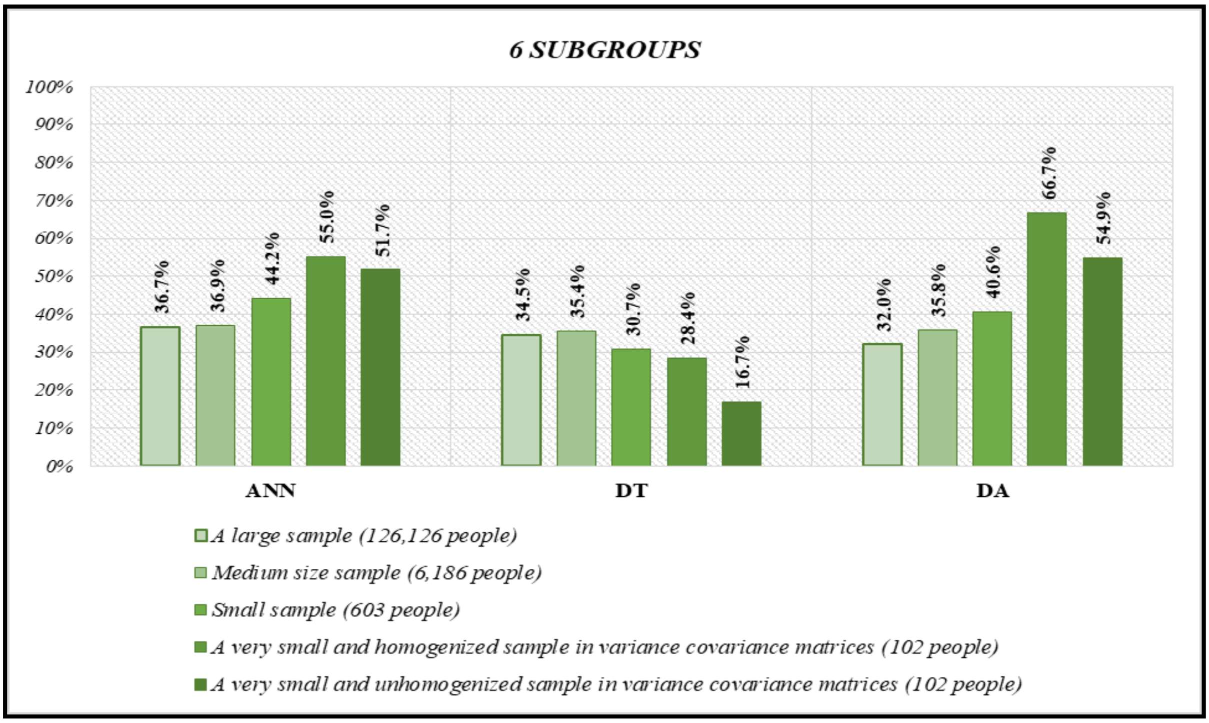


Figure 1. *The performance of the methods to classify student achievement into 6 subgroups in different sample sizes*

When the findings obtained are examined, the most successful method for making the classification into 6 subgroups is the ANNs in large, medium and small sample sizes. While the second-best method is the DTs in the large sample, the DA is the second-best method in the medium and small samples. While the DA is the most successful method for classifying 6 subgroups in the very small sample size where homogeneity was ensured and was not ensured in variance-covariance matrices; the second-best method is the ANNs.

As the sample size changes in classification into 6 subgroups, different findings regarding the performance of the ANNs have been obtained. While the performance averages of 50 different trial analyses conducted for the purpose of determining the best performance of the ANNs indicate that the best performance average is 39.9% and in 50 different trial analyses, the highest classification performance was achieved with 55.0% in the very small sample size where homogeneity was ensured in variance-covariance matrices. While this performance was followed by the small sample with an average performance of 38.4%; the second-highest performance was obtained in the very small sample size which homogeneity was not ensured in variance-covariance matrices. While the third-best average performance was obtained in the large sample, the third-highest performance was observed in the small sample. While the fourth-best average performance was obtained in the medium-sized sample, the fourth-highest performance was again observed in the medium-sized sample. It was observed that the lowest average performance was in the very small sample size where homogeneity was not ensured in variance-covariance matrices and that the lowest performance was in the large sample. These findings suggest that increasing the number of trials for the purpose of achieving the best performance in applications to be made with the ANNs in the classification of 6 subgroups will increase the opportunity to achieve the best performance. At the same time, ensuring the homogeneity of variance-covariance matrices increases the performance of the ANNs classification into 6 subgroups.

The DTs showed the highest performance in the classification of 6 subgroups in the medium-sized sample, while the performance of the DTs in the large sample was approximately 0.9%

lower than that in the medium-sized sample, this loss was further increased in small and very small samples. While the classification performance was 35.4% in the medium-sized sample where the method achieved the highest performance, the performance of the method decreased to 16.7%, especially in the very small sample size where homogeneity was not ensured in variance-covariance matrices. In addition, although the DTs method lost performance when the sample size decreased, if homogeneity was ensured in variance-covariance matrices, it showed higher performance in the very small sample than the very small sample where homogeneity was not ensured in variance-covariance matrices.

The DA is very sensitive to the homogeneity of variance-covariance matrices. The high performance of the DA in the very small sample and in the sample where homogeneity was ensured in variance-covariance matrices reveals this situation. The classification performance which increased to 32.0% in the large sample, 35.8% in the medium-sized sample and 40.6% in the small sample for classifying into 6 subgroups, approximately doubled and was 66.7% in the very small sample size where homogeneity was ensured in variance-covariance matrices. Also, the classification performance in the very small sample where homogeneity was not ensured in variance-covariance matrices is 54.9% for classifying into 6 subgroups. This supports the fact that the performance of the method increases as the sample size decreases.

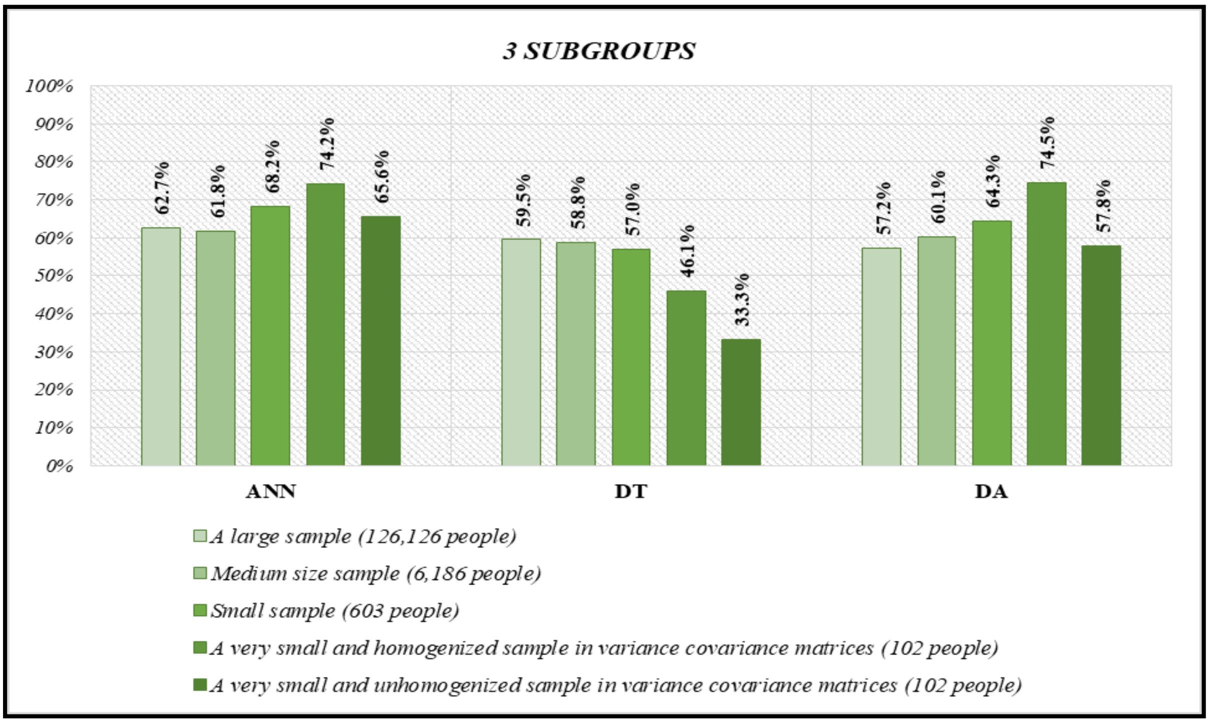


Figure 2. *The performance of the methods to classify student achievement into 3 subgroups in different sample sizes*

The ANNs are also the most successful method in the classification of 3 subgroups in large, medium and small sample sizes. While the second-best method is the DTs in the large sample, it is the DA method in medium and small samples. While the DA is the most successful method in classifying into 3 subgroups in the very small sample size where homogeneity was ensured in variance-covariance matrices, the second-best method is the ANNs. While the ANNs are the most successful method in classifying into 3 subgroups in a very small sample size where homogeneity was not ensured in variance-covariance matrices; the second-best method is the DA.

As the sample size changes in classifying into 3 subgroups, different findings regarding the performance of the ANNs have been obtained. In order to determine the best performance of the ANNs, the performance averages of 50 different trial analyses conducted in each sample size showed that the best performance average was in the small sample with 63.1%. In 50 different trial analyses, the highest classification performance was achieved in the very small sample size where homogeneity was ensured in variance-covariance matrices with 74.2%. While this performance was followed by a very small sample size where homogeneity was ensured in variance-covariance matrices where the average performance was 62.3%; the second-highest performance was obtained in the small sample. While the third-best average performance was obtained in the large sample, the third-highest performance was observed in the very small sample size where homogeneity was not ensured in variance-covariance matrices. While the fourth-best average performance was obtained in the medium sample, the fourth-highest performance was observed in the large sample. It was observed that the lowest average performance was in the very small sample size where homogeneity was not ensured in variance-covariance matrices and that the lowest performance was in the medium size sample. These findings suggest that increasing the number of trials in order to achieve the best performance in applications to be made with the ANNs in classifying into 3 subgroups will increase the opportunity to achieve the best performance. Also, ensuring the homogeneity of variance-covariance matrices seems to increase the performance of the ANNs.

The classification performance of the DTs method decreased in classifying into 3 subgroups as the sample size decreased. The DTs showed the highest performance in the large sample. While the performance of the DTs was lost approximately 0.7% in the medium-sized sample compared to the larger sample, this loss was further increased in small and very small samples. While the classification performance of the DTs method was 59.5% in the medium-sized sample where it achieved the highest performance, the performance of the method decreased to 33.3%, especially in the very small sample size where homogeneity was not ensured in variance-covariance matrices. In addition, although the DTs method lost performance when the sample size decreased, if homogeneity was ensured in variance-covariance matrices, it showed higher performance in the very small sample than the very small sample where homogeneity was not ensured in variance-covariance matrices.

The DA is very sensitive to the homogeneity of variance-covariance matrices. The highest performance shown by the DA in the very small sample and in the sample where homogeneity was ensured in variance-covariance matrices reveals this. The classification performance in classifying into 3 subgroups which increased to 57.2% in the large sample, 60.1% in the medium sample and 64.3% in the small sample, was 74.5% in the very small sample size where homogeneity was ensured in variance-covariance matrices. The fact that the classification performance is 57.8% in the very small sample size where homogeneity was not ensured in variance-covariance matrices suggests that the method's classification performance into 3 subgroups is low in cases where homogeneity was not ensured in variance-covariance matrices. When this assumption is not fulfilled, the sample where the performance is the highest is determined as a small sample.

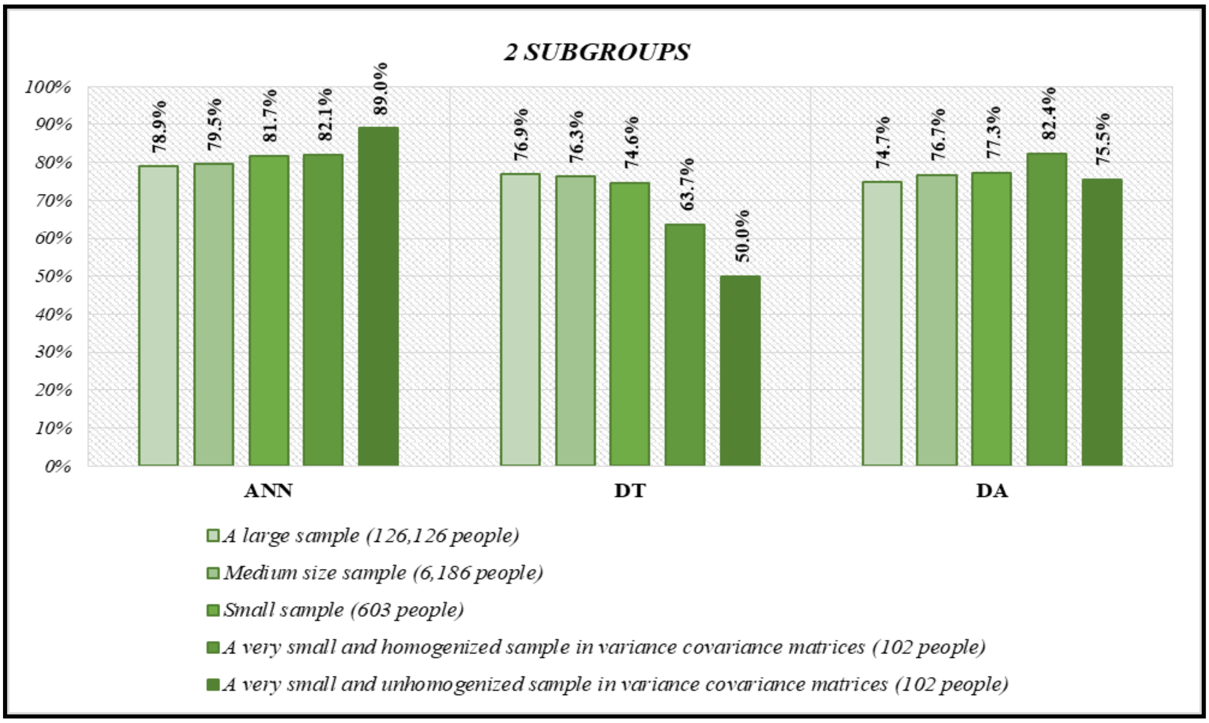


Figure 3. *The performance of the methods to classify student achievement into 2 subgroups in different sample sizes*

The ANNs are the most successful method in classifying into 2 subgroups in large, medium and small sample sizes. While the second-best method is the DTs in the large sample, the DA is the second-best method in the medium and small samples. While the most successful method for making 2 subgroupings in a very small sample size where homogeneous was ensured in variance-covariance matrices is the DA; the second-best method is the ANNs. While the ANNs are the most successful method for classifying 2 subgroups in a very small sample size where homogeneity was not ensured in variance-covariance matrixes; the second-best method is the DA.

As the sample size changes in classifying into 2 subgroups, different findings regarding the performance of the ANNs have been obtained. While the performance averages of 50 different trial analyses in each sample size for the purpose of determining the best performance of the ANNs, showed that the best performance average was in the large sample with 78.2%; in 50 different trial analyses, the highest classification performance was achieved with 89.0% in the very small sample size where homogeneity was not ensured in variance-covariance matrices. While this performance was followed by a small sample where the average performance was 78.0%, the second-highest performance was obtained in the very small sample size where homogeneity was ensured in variance-covariance matrices. While the third-best average performance was obtained in the medium sample, the third-highest performance was observed in the small sample. While the fourth-best average performance was obtained in the very small sample size where homogeneity was ensured in variance-covariance matrices, the fourth-highest performance was observed in the medium-sized sample. It was observed that the lowest average performance was in the very small sample size where homogeneity was not ensured in variance-covariance matrices and that the lowest performance was in the large sample. These findings suggest that increasing the number of trials in order to achieve the best performance in applications to be performed with the ANNs in classifying into 2 subgroups will also increase the opportunity to achieve the best performance. Also, ensuring the homogeneity of variance-covariance matrices increases the performance of the ANNs.

The classification performance of the DTs method decreased in classifying into 2 subgroups as the sample size decreased. The DTs showed the highest performance in the large sample. While the performance of DTs was lost approximately 0.6% in the medium-sized sample compared to the large sample, this loss was further increased in small and very small samples. While the classification performance of the DTs method was 76.9% in the medium-sized sample where it achieved the highest performance, especially in a very small sample size where homogeneity was not ensured in variance-covariance matrices, the performance of the method decreased to by 50.0%. In addition, although the DTs method lost performance when the sample size decreased, if homogeneity was ensured in variance-covariance matrices, the method showed higher performance in the very small sample than the very small sample size where homogeneity was not ensured in variance-covariance matrices.

The DA is very sensitive to the homogeneity of variance-covariance matrices. The high performance that the DA showed in a very small sample and in a sample where homogeneity was ensured in variance-covariance matrices reveals this. The classification performance which increased to 74.7% in the large sample, 76.7% in the medium sample and 77.3% in the small sample in classifying into 2 subgroups was 82.4% in the small sample where homogeneity was ensured in variance-covariance matrices. The fact that the classification performance into 2 subgroups is 75.5% in the very small sample where homogeneity was not ensured in variance-covariance matrices reveals that the performance of the method is low when homogeneity was not ensured in variance-covariance matrices. If this assumption is not fulfilled, the sample with the highest performance is determined as a small sample.

4. DISCUSSION and CONCLUSION

In this study, the classification performances of different classification methods were evaluated under varying conditions. Apart from comparing the performances of the methods used in the study in classifying different sample sizes and subgroup numbers with each other, results regarding the conditions under which each method performed the best were obtained. A result that can be said in general for each of the methods; when the number of subgroups classified was less, the methods showed higher performance. When the number of subgroups is high, classification becomes difficult as expected.

As a result of 50 different trial analyses performed to achieve the highest performance with ANNs in each sample size, the highest performance was reached with 89.0% in a very small sample where there was no homogeneity in variance-covariance matrices. However, when the average of 50 different trial analyses was examined, it was determined that the highest performance average was obtained with a large sample with 78.2%. Again, the second-highest performance with 82.1% was observed in a very small sample where homogeneity was achieved in variance-covariance matrices; the fact that the average of 50 different trial analyses in this sample was lower than the performance average obtained in the large sample showed a similar situation.

ANNs, use a different part of the research data as a training set and another part as a test set in each analysis. For this reason, the findings obtained in each analysis differ. These results show that increasing the number of trials analyzes increases the chance of achieving the highest performance. As a result, it is possible to say that ANNs show the highest performance in a very small sample where homogeneity is not provided in variance-covariance matrices. This situation also revealed that there is no need to meet the assumption of ensuring homogeneity in the variance-covariance matrices in the classification studies to be performed with ANNs.

The highest classification performance in terms of DTs method was obtained in classifying 2 subgroups in the large sample. However, it is possible to say that the performance of the method decreases linearly as the sample size decreases, except that the performance of classifying into

6 subgroups in the large sample is approximately 0.9% lower than the performance of classifying into 6 subgroups in the medium sample. In addition, the fact that the classification performance for all subgroups is higher in case of homogeneity in variance-covariance matrices in a very small sample, compared to a very small sample where homogeneity is not provided, shows that the performance of the method will increase if this assumption is met.

The DTs method does not give good results in estimating the values of continuous variables. In addition, it does not give good results in a/the model building when the number of classes is very high and learning set examples are low (Akpınar, 2014; Büyükışıklar, 2014; Köktürk, 2012). The findings obtained in this study are similar to those obtained in the literature. At the same time, it was determined that ANNs showed higher classification performance in the studies conducted to compare the classification performances of ANNs and DTs, and it was observed that this result coincided with the research findings.

Tosun (2007) used the ANNs and the DTs in classifying student achievements into 2 subgroups in a small sample of 424 people. As a result of the analysis, he found that the performance of the ANNs was higher. İbrahim and Rusli (2007) compared the performance of the DTs, the regression analysis and the ANNs in classifying students' academic achievement in a small sample of 206 students. As a result of the analysis, they determined that the ANNs yielded more achievementful results than the other two methods. Çölkesen (2009) compared the achievement of the classification of satellite images into 6 subgroups in medium-sized samples of 6000 and 3750 people with the ANNs, the DTs and the k-star algorithms. As a result of the analysis, he determined that advanced classification techniques were a better and more effective alternative than conventional classifiers in the classification of remotely sensed images. Köktürk (2012) compared the achievements of the classification of the K-nearest neighborhood, the ANNs and the DTs in a small sample of 240 patients who applied to the gynecology and obstetrics clinic. She compared the data obtained from pregnant women in classifying into two subgroups. As a result of the analysis, he determined that the achievement of the classification of the ANNs technique was better than the other two methods. Kuyucu (2012) compared the classification performance of the logistic regression, the ANNs and the DTs in classifying into two subgroups in a small sample of 236 people. As a result of the analysis, he determined that the method which showed the highest performance was the ANNs.

The performance of DA was significantly affected by the sample size and homogeneity in variance-covariance matrices, which is one of the basic assumptions of the method. As the sample size decreased, the performance of the DA increased linearly. In addition, ensuring homogeneity in variance-covariance matrices made the performance of the method even stronger. DA achieved the highest performance in classifying 2 subgroups in the sample where homogeneity was achieved in variance-covariance matrices. On the other hand, even if the classification performance into 2 and 3 subgroups in very small samples was higher than in the large sample when homogeneity could not be achieved, it remained behind the performances obtained in medium and small samples. These results show that if the homogeneity is achieved in variance-covariance matrices, DA reaches a very high performance. DA showed higher performance than ANNs in classifying into 6, 3 and 2 subgroups in a very small sample where homogeneity was achieved in variance-covariance matrices. On the other hand, in the classification of large, medium and small samples into 6, 3 and 2 subgroups, ANNs have performed better than DA, and these results are similar to the literature.

Türe et al. (2005) compared the ANNs, logistic regression analysis and flexible DA methods in the prediction of primary hypertension in classifying into two subgroups in a small sample of 276. As a result of the analysis, they determined that the performance of the neural networks was higher than that of the DA. Burmaoğlu (2009) compared the achievement of the DA, the logistic regression analysis and the classification of the ANNs into two subgroups in a small sample of 120 where homogeneity was not ensured in variance-covariance matrices using United Nations Development Program Human Development Index Data. As a result of the analysis, he determined that the multilayer perceptron model made better classification than the DA. Avcılar and Yakut (2015) compared the classification performance of the ANNs, the logistic regression and the DA methods in classifying into three subgroups in a small sample of 500 in determining voter preferences in local elections. As a result of the analysis, they found that the performance of the ANNs was higher than the DA. Wheeler (1993), compared the methods of ANNs and DA in classifying the achievement of the law school students' attorney exams into 2 subgroups in a small sample of 460 people. As a result of the analysis, he determined that the ANNs showed higher performance.

The results obtained from this study were similar to the results of the studies in the literature. In addition, the performance results of the methods used in the study in different sample sizes and in classifying into different subgroup numbers added new findings to the literature.

The suggestions developed based on the results obtained from the research can be listed as suggestions based on the results of this research and suggestions for the researchers in new studies.

As a result of the research, the highest performance of ANNs in the classification of student achievement into 6, 3 and 2 subgroups in large, medium-sized and small samples revealed that the method can be used reliably in these sample sizes and in these subgroup numbers. For this reason, it is recommended that ANNs should be preferred to classify 6, 3 and 2 subgroups in these sample sizes. The highest performance of ANNs in classifying into 2 and 3 subgroups in very small sample sizes where homogeneity is not provided in variance-covariance matrices is an important reason for the method to be preferred in classifying these subgroup numbers.

DTs lost their classification performance significantly as the sample size decreased. In addition to the decrease in the sample size, the performance loss of the method increased even more when the variance-covariance matrices were not homogenous. For this reason, reaching large samples in classification studies to be carried out with DTs is important and necessary for the high performance of the method.

The performance of the DA method increased as the sample size decreased. In addition, DA showed a much higher performance than ANNs and DTs in the condition that homogeneity is provided in variance-covariance matrices. According to the results obtained from the research, it can be said that it would be appropriate to prefer DA in case of homogeneity in variance-covariance matrices in classification studies to be made in very small samples.

The ANNs findings obtained in the research are limited to the multilayer perceptron model. It may be useful to examine different network models in other studies in order to obtain more information about the performance of ANNs in different sample sizes and classification into subgroup numbers.

In the analysis made with ANNs in the research, 70% of the data was separated as a training set and 30% as a test set. In new researches, the performance of the method can be handled under different conditions by changing the ratios of training and test sets.

DTs findings obtained in the research are limited by the CHAID algorithm. It may also be useful to examine different algorithms in other studies in order to obtain more information about DTs' performance in classifying different sample sizes and subgroup numbers.

DA findings obtained in the research are also limited to linear DA. It may be useful to use different types of DA such as quadratic DA and flexible DA in other studies in order to obtain more information about DA's performance in classifying different sample sizes and subgroup numbers.

In the study, variables with no missing data, the closest to normal distribution, no extreme values, no multicollinearity problems among variables as much as possible and the most frequently used variables in the literature were used in the classification of student achievement. New research with different variables other than these variables can contribute to the generalizability of the performance of the methods.

It is possible to work with data sets with missing data in researches in the field of education. Therefore, examining the performance of the methods with a different study in case of missing data in the data sets may contribute to the literature.

PISA 2012 mathematics test and survey results were used in the study. Working with data sets from different fields other than mathematics can gain new perspectives to researchers by revealing the performance of the methods in other fields.

In PISA applications, there are a total of 7 levels, which are lower than level one, level one, level two, level three, level four, level five and level six in order to determine student achievement. In this study, in order to create 6, 3 and 2 subgroups, students with a level lower than one were excluded from the study and the studies were carried out on 6, 3 and 2 subgroups. In different studies, 4, 5, 7 etc. comparisons can be made in samples with different subgroup numbers.

In this study, the performances of the methods in large, medium, small and very small sample sizes were discussed. New research plans can be prepared by creating different sample sizes to compare methods.

Acknowledgements

This paper was extracted from the first author's Ph.D. thesis.

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).

ORCID

Emre TOPRAK https://orcid.org/0000-0002-4131-4888
Selahattin GELBAL https://orcid.org/0000-0001-5181-7262

5. REFERENCES

Akpınar, H. (2014). *DATA, veri madenciliği veri analizi [DATA, data mining data analysis]*. İstanbul: Papatya Publishing.

Albayrak, A. S. (2006). *Uygulamalı çok değişkenli istatistik teknikleri [Applied multivariate statistical techniques]*. Ankara: Asil Publishing.

Altıntaş, Y. (2010). *Veri madenciliğinin tıpta kullanımı ve bir uygulama: Hemodiyaliz hastaları için risk seviyelerine göre risk faktörlerinin etkileşimlerinin incelemesi [The usage of data mining in medicine and an application: Analysis of risk factors' interactions according to risk levels for hemodialysis patients]* (Unpublished Master Thesis). Gazi University, Institute of Science, Ankara.

Altun, F., & Yazıcı, H. (2013). Ergenlerin benlik algılarının yordayıcıları olarak: Akademik öz-yeterlik inancı ve akademik başarı [As predictors of self-perception of adolescents: Academic self-efficacy belief and academic success]. *Kastamonu Education Journal*, *21*(1), 145-156.

Anıl, D. (2008). The analysis of factors affecting the mathematical achievement of Turkish students in the PISA 2006 evaluation program with structural equation modeling. *American-Eurasian Journal of Scientific Research, 3*(2), 222-227.

Anıl, D. (2009). Uluslararası öğrenci başarılarını değerlendirme programı (PISA)'nda Türkiye'deki öğrencilerin fen bilimleri başarılarını etkileyen faktörler [Factors effecting science achievement of science students in programme for international students' achievement (PISA) in Turkey]. *Education ve Science, 34*(152), 87-100.

Anıl, D. (2011). Türkiye'nin PISA 2006 fen bilimleri başarısını etkileyen faktörlerin yapısal eşitlik modeli ile incelenmesi [Investigation of factors affecting PISA 2006 science success of Turkey with structural equation model]. *Educational Sciences: Theory & Practice*, *11*(3), 1253-1266.

Argüden, Y., & Erşahin, B. (2008). *Veri madenciliği: Veriden bilgiye, masraftan değere [Data mining: From data to information, from cost to value]*. İstanbul: ARGE Consultancy.

Arslantaş, İ. H., Özkan, M., & Külekçi, E. (2012). Eğitim fakültesi öğrencilerinin akademik başarı düzeylerinin bazı demografik değişkenler açısından incelenmesi [The analysis of academic achievement for some of demographic variables in education faculty of students]. *Electronic Journal of Social Sciences, 11*(39), 395-407.

Atar, H. Y. (2012). Resim-iş öğretmenliği özel yetenek sınavlarının sınıflama doğruluğu üzerine bir çalışma [A study on the classification accuracy of art teaching special aptitude exams]. *Education ve Science*, *37*(163), 283-296.

Avcılar, M. Y., & Yakut, E. (2015). Yapay sinir ağları çoklu lojistik regresyon ve çoklu diskriminant analiz yöntemlerinden yararlanarak yerel seçimlerde seçmen tercihlerinin belirlenmesi: Osmaniye ili uygulaması [Determination of the voter preferences by using ANNs, multiple logistic regression and multiple discriminant analysis techniques: An investigation local elections in Osmaniye province]. *International Journal of Alanya Faculty of Business Administration*, *7*(2), 207-224.

Baş, N. (2006). *Yapay sinir ağları yaklaşımı ve bir uygulama [Artificial neural networks approach and an application]* (Unpublished Master Thesis). Mimar Sinan Fine Arts University, Institute of Science, İstanbul.

Bayru, P. (2007). *Elektronik basında tüketici tercihleri analizi: yapay sinir ağları ile lojit modelin performans değerlendirilmesi [Electronic media consumer choice analysis: Artificial neural networks to evaluate the performance of the model with lojit]* (Unpublished Doctoral Dissertation). İstanbul University, Institute of Social Sciences, İstanbul.

Bektaş, S. (2012). Çok şeritli bölünmüş karayollarında kaza tahmin modeli [A crash prediction model for multilane divided highway]. *Journal of Advanced Technology Sciences*, *1*(1), 27-34.

Benli, Y. K. (2005). Bankalarda mali başarısızlığın öngörülmesi lojistik regreyon ve yapay sinir ağı karşılaştırması [Prediction of financial failure in banks, comparison of logistic regression and artificial neural network]. *The Journal of the Industrial Arts Education Faculty of Gazi University*, 16, 31-46.

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: For marketing, sales, and customer relationship management* (2nd ed). USA: Wiley.

Brown, J. D. (2007). *Neural network prediction of math and reading proficiency as reported in the educational longitudinal study 2002 based on non-curricular variables* (Unpublished Doctoral Dissertation). Duquesne University, Pennsylvania, ABD.

Burmaoğlu, S. (2009). *Birleşmiş milletler kalkınma programı beşeri kalkınma endeksi verilerini kullanarak diskriminant analizi, lojistik regresyon analizi ve yapay sinir ağlarının sınıflandırma başarılarının değerlendirilmesi [Evaluating classification success of discriminant analysis, logistic regression analysis and neural network models using UN Developing Programme's Human Development Index]* (Unpublished Doctoral Dissertation). Atatürk University, Institute of Social Sciences, Erzurum.

Büyükışıklar, A. (2014). *Karar ağaçları sınıflandırma algoritması ile toprak özgül direnci tespitinde jeolojik veri kullanımı [Use of geological data to determine soil resistance with decicson tree classification algorithm]* (Unpublished Master Thesis). Bilecik Şeyh Edebali University, Institute of Science, Bilecik.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2008). *Bilimsel araştırma yöntemleri [Scientific research methods].* Ankara: Pegem Academy.

Ceylan, E. (2009). PISA 2006 sonuçlarına göre Türkiye'de fen okuryazarlığında düşük ve yüksek performans gösteren okullar arasındaki farklar [Differences between low-and high-performing schools in scientific literacy based on PISA 2006 results in Turkey]. *Van Yuzuncu Yil University Journal of Education*, *6*(2), 55-75.

Chang, L. Y., & Wang, H. W. (2006). Analysis of traffic injury: an application of non-parametric classification tree techniques. *Accident Analysis Prevention*, *38*, 1019-1027.

Çakmak, Z. (1992). *Çoklu ayırma ve sınıflandırma analizi: Eğitimde öğrencilerin meslek seçimine uygulanması [Multiple discriminant and classification analysis: Application to students' choice of profession in education]*. No: 658, Eskişehir: Anadolu University Publications.

Çakmak, Z., & Kara, H. (2011). Yöneticilerde benlik algılamalarının belirlenmesi: Sanayi örgütlerinde bir araştırma [Determination of self-perception of managers: A research in industrial organizations]. *Dumlupınar University Journal of Social Sciences*, 30, 301-310.

Çankaya, A. B., Taşdemir, G., Taşdemir, S., & Zilelioğlu, O. (2009). Delici göz yaralanması olgularımızın uzun dönem sonuçları ve görsel prognozu etkileyen faktörlerin analizi [Long term results of our penetrating eye injury cases and factors influencing final visual outcome]. *Turkish Journal of Ophthalmology*, *39*, 220-226.

Çırak, G. (2012). *Yükseköğretimde öğrenci başarılarının sınıflandırılmasında yapay sinir ağları ve lojistik regresyon yöntemlerinin kullanılması [The usage of artifical neural network and logistic regression methods in the classification of student achievement at higher education]* (Unpublished Master Thesis). Ankara University, Institute of Educational Sciences, Ankara.

Çiftçi, C., & Çağlar, A. (2014). Ailelerin sosyo-ekonomik özelliklerinin öğrenci başarısı üzerindeki etkisi: Fakirlik kader midir? [The effect of socio-economic characteristics of parents on student achievement: Is poverty destiny?]. *International Journal of Human Sciences*, *11*(2), 155-175.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları [Multivariate statistics for social sciences SPSS and LISREL applications]* (2nd ed). Ankara: Pegem Academy.

Çölkesen, İ. (2009). *Uzaktan algılamada ileri sınıflandırma tekniklerinin karşılaştırılması ve analizi [Comparing and analyzing of advanced classifier techniques in remote sensing]* (Unpublished Master Thesis). Gebze Technical University, Institute of Engineering and Science, Kocaeli.

Çuhadar, M. (2006). *Turizm sektöründe talep tahmini için yapay sinir ağları kullanımı ve diğer yöntemlerle karşılaştırmalı analizi (Antalya ilinin dış turizm talebinde uygulama) [Forecasting tourism demand by artificial neural networks and time series methods (A comparative analysis in inbound tourism demand to Antalya)]* (Unpublished Doctoral Dissertation). Süleyman Demirel University, Institute of Social Sciences, Isparta.

Demircioğlu, N., Ayan, S., Avanoğlu, B., & Sıvacıoğlu, A. (2004). Kastamonu-Taşköprü orman fidanlığında üretilen 2+0 yaşlı sarıçam (Pinus sylvestris L.) fidanlarının TSE normlarına göre değerlendirilmesi [Evaluation of 2+0 aged nursery of the scotch pine (Pinus sylvestrisl.) raised in Kastamonu-Taşköprü forest nursery as to TSE quality classification]. *Pamukkale University Journal of Engineering Sciences*, *10*(2), 243-251.

Demiryürek, O. (2009). *Polyester/viskon karışımlı open-end rotor iplik özelliklerinin yapay sinir ağları ve istatistiksel modeller kurularak tahmin edilmesi [Predicting the properties of polyester/viscose blended open-end rotor spun yarns by establishing artificial neural networks and statistical models]* (Unpublished Doctoral Dissertation). Çukurova University, Institute of Science, Adana.

Dikmen, İ. (2001). *Strategic decision making in construction companies: An artificial neural network based decision support system for international market selection* (Unpublished Doctoral Dissertation). Middle East Technical University, Institute of Science, Ankara.

Doğan, N. (2009). Bilgisayar destekli istatistik öğretiminin başarıya ve istatistiğe karşı tutuma etkisi [The effect of computer-assisted statistics instruction on achievement and attitudes toward statistics]. *Education and Science, 34*(154), 3-16.

Durmuş, G. (2008). *Çimentolu harç özelliklerine yüksek sıcaklık etkisinin belirlenmesi ve yapay sinir ağı ile modellenmesi [Examining of the effect of high temperature on cement mortar properties and modelling by artificial neural network]* (Unpublished Doctoral Dissertation). Gazi University, Institute of Science, Ankara.

Elmas, Ç. (2003). *Yapay sinir ağları [Artificial neural networks]*. Ankara: Seçkin Bookstore.

Erayman, Y. (2004). *KSÜ öğrencilerinin sosyo-ekonomik yapılarının başarıları üzerine etkisi [The effect of socio-economical structure of the students in KSU on their success]* (Unpublished Master Thesis). Kahramanmaraş Sütçü İmam University, Institute of Science, Kahramanmaraş.

Ercan, S., Işık, O., & Çakır, V. (2005). *HHO öğrencilerinin akademik başarılarına etki eden faktörlerin çoklu regresyon yöntemiyle incelenmesi [Investigation of factors affecting academic success of HHO students with multiple regression method]* V. National Production Research Symposium (25-27 November 2005), İstanbul Commerce University, İstanbul.

Erdil, Z. (2010). Sosyoekonomik olarak risk altında bulunan çocuklara yönelik erken müdahale programları ve akademik başarı ilişkisi [Relationship of academic achievement and early intervention programs for children who are at socio-economical risk]. *Hacettepe University Faculty of Health Sciences Nursing Journal*, *17*(1), 72-78.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, *7*(2), 179-188.

Garson, D. G. (1998). *Neural networks, an ıntroductory guide for social scientists*. London: Sage Publications, 25.

Gelbal, S. (2008). Sekizinci sınıf öğrencilerinin sosyoekonomik özelliklerinin Türkçe başarısı üzerindeki etkisi [The effect of socio-economic status of eighth grade students on their achievement in Turkish]. *Education and Science, 33*(150), 1-13.

Grimm, L. G., & Yarnold, P. R. (1995). *Reading and understanding multivarite statistics*. Washington D. C.: American Psychological Association.

Güneri, N., & Apaydın, A. (2004). Öğrenci başarılarının sınıflandırılmasında lojistik regresyon analizi ve sinir ağları yaklaşımı [Logistic regression analysis and neural networks approach in the classification of students' achievement]. *Gazi University Journal of Commerce and Tourism Education Faculty, 1*, 170-188.

Güzeller, C., & Kelecioğlu, H. (2006). Ortaöğretim kurumları öğrenci seçme sınavının sınıflama geçerliği üzerine bir çalışma [The study on classification validity of secondary education student selection & placement exam]. *H. U. Journal of Education*, *30*, 140-148.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. New York: Mcmillan Press.

Huberty, C. J. (1994). *Applied discrimination analysis.* New York: John Wiley and Sons.

İbrahim, Z., & Rusli, D. (2007). *Predicting students' academic performance: Comparing neural network, decision tree and linear regression*. 21. Annual SAS Malasia Forum (5 September 2007), Kuala Lumpur, Malezya.

Jain, A. K., Mao, J., & Mohiuddin, K. (1996). Artificial neural networks: A tutorial. *IEEE Computer Society*, *29*(3), 31-44.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). New Jersey: Prentice-Hall, Englewood Cliffs.

Kachigan, S. K. (1991). *Multivariate statistical analysis: A conceptual introduction* (2nd ed.). New York: Radius Press.

Kaptan, S. (1995). *Bilimsel araştırma ve istatistik teknikleri [Scientific research and statistical techniques]*. Ankara: Tekışık Web Offset.

Karasar, N. (2014). *Bilimsel araştırma yöntemi [Scientific research method]* (27th ed). Ankara: Nobel Academic Publishing.

Kayıkçı, Ş. (2014). *Web sayfalarının yapay sinir ağları ile sınıflandırılması [Classification of web pages using neural networks]* (Unpublished Doctoral Dissertation). Marmara University, Institute of Social Sciences, İstanbul.

Kayri, M., & Boysan, M. (2008). Bilişsel yatkınlık ile depresyon düzeyleri ilişkisinin sınıflandırma ve regresyon ağacı analizi ile incelenmesi [Assesment of relation between cognitive vulnerability and depression's level by using classification and regression tree analysis]. *H. U. Journal of Education*, *34*, 168-177.

Kayri, M., & Günüç, S. (2010). Türkiye'deki ortaöğretim öğrencilerinin internet bağımlılık düzeyini etkileyen bazı faktörlerin karar ağaçları yöntemleri ile incelenmesi [An analysis of some variables affecting the internet dependency level of Turkish adolescents by using decision tree methods]. *Educational Sciences: Theory & Practice*, *10*(4), 2465-2500.

Kaysılı, B. (2008). Akademik başarının arttırılmasında aile katılımı [Parent involvement to improve academic achievement]. *Ankara University Faculty of Educational Sciences Journal of Special Education*, *9*(1), 69-83.

Keser, İ., & Sarıbay, E. (2007). İzmir'deki özel ve devlet üniversitelerindeki öğrencilerin başarılarını etkileyen faktörlerin belirlenmesi ve karşılaştırılması [Determining and comparing the factors effecting the performance of the students at the state and private universities in İzmir]. *Muğla University Journal of Institute of Social Sciences,* 18. 39-48.

Kıran, Z. (2010). *Lojistik regresyon ve C&RT analizi yöntemleriyle sosyal güvenlik kurumu ilaç provizyon sistemi üzerinde bir uygulama [An application on Pharmacy Provision System data of Social Security Institution by logistic regression and CART analysis technics]* (Unpublished Master Thesis). Gazi University, Institute of Science, Ankara.

Kibar, F. (2015). *Türkiye'de kamyon kazaları ile trafik ve karayolu geometrik özellikleri arasındaki ilişkinin istatistiksel ve yapay sinir ağları yöntemleri ile modellenmesi [Modeling the relationship between truck accidents and traffic and highway geometric characteristics in turkey with statistical and artificial neural networks methods]* (Unpublished Doctoral Dissertation). Karadeniz Technical University, Institute of Science, Trabzon.

Klecka, W. (1980). *Discriminant analysis*. London: Sage Publications.

Kocadağlı, O. (2012). *Genetik algoritmalar ve bulanık üyelik fonksiyonlarıyla hibrit bayes yapay sinir ağları [Hybrid bayesian neural networks with genetic algorithms and fuzzy membership functions]* (Unpublished Doctoral Dissertation). Mimar Sinan Fine Arts University, Institute of Science, İstanbul.

Köktürk, F. (2012). *K-en yakın komşuluk, yapay sinir ağları ve karar ağaçları yöntemlerinin sınıflandırma başarılarının karşılaştırılması [Comparing classification success of k-nearest neighbor, artifical neural network and decision trees]* (Unpublished Doctoral Dissertation). Bülent Ecevit University, Institute of Health Sciences, Zonguldak.

Kurt, İ., & Türe, M. (2005). Tıp öğrencilerinde alkol kullanımını etkileyen faktörlerin belirlenmesinde yapay sinir ağları ile lojistik regresyon analizinin karşılaştırılması

[Comparison of artificial neural networks and logistic regression analysis in determining factors affecting alcohol consumption among medicine students]. *Medical Journal of Trakya University*, *22*(3), 142-153.

Kuyucu, Y. E. (2012). *Lojistik regresyon analizi (LRA), yapay sinir ağları (YSA) ve sınıflandırma ve regresyon ağaçları (C&RT) yöntemlerinin karşılaştırılması ve tıp alanında bir uygulama [Comparison of logistic regression analysis (LRA), artificial neural networks (ANNs) and classfication and regression trees (C&RT) methods and an aplication in medicine]* (Unpublished Master Thesis). Gaziosmanpaşa University, Institute of Health Sciences, Tokat.

Munakata, T. (2008). *Fundamentals of the new artificial intelligence, neural, evolutionary, fuzzy and more*. Springer-Verlag London Limited.

Naik, B., & Ragothaman, S. (2004). Using neural networks to predict MBA student achievement. *College Student Journal*, *38*(1), 143-149.

Ocakoğlu, G. (2006). *Lojistik regresyon analizi ve yapay sinir ağı tekniklerinin sınıflama karşılaştırması ve bir uygulama [Logistic regression analysis and comparison of classification characteristics of artifical nueural network techniques and an application]* (Unpublished Master Thesis). Uludağ University Institute of Health Sciences, Bursa.

Oğuzlar, A. (2004). CART analizi ile hanehalkı işgücü anketi sonuçlarının özetlenmesi [Summary of results of household labor force with CART analysis]. *Atatürk University Journal of Economics and Administrative Sciences*, *18*(3-4), (79-90).

Oğuzlar, A. (2006). Hanehalkı tipi ve kır-kent ayırımının diskriminant analizi ile incelenmesi [Assessment of household type and rural-urban area distinctions by means of discriminant analysis]. *Akdeniz University Faculty of Economics & Administrative Sciences Faculty*, 11, 70-84.

Okioga, C. K. (2013). The impact of students' socio-economic background on academic performance in universities, a case of students in Kisii University College. *American International Journal of Science*, *2*(2), 38-46.

Özabacı, N., & Acat, M. B. (2005). Sosyo ekonomik çevreye göre ilköğretim öğrencilerinin başarısızlık nedenleri [Causes of academic underachievement among socio-economic level for secondary school student]. *Eskişehir Osmangazi University Journal of Social Sciences*, *6*(1), 145-170.

Özdamar, K. (2010). *Paket programlar ile istatistiksel veri analizi (Çok değişkenli analizler) [Statistical data analysis with package programs (Multivariate analysis)]* (7th ed), Eskişehir: Kaan Bookstore.

Özdemir, Y., & Koruklu, N. (2011). Üniversite öğrencilerinde değerler ve mutluluk arasındaki ilişkinin incelenmesi [Investigating relatinship between values and happiness among university students]. *Van Yuzuncu Yil University Journal of Education*, *8*(1), 190-210.

Özekes, S. (2003). Veri madenciliği modelleri ve uygulama alanları [Data mining models and application areas]. *Journal of İstanbul Commerce University*, 3, 65-82.

Özer, Y., & Anıl, D. (2011). Öğrencilerin fen ve matematik başarılarını etkileyen faktörlerin yapısal eşitlik modeli ile incelenmesi [Examining the factors affecting students' science and mathematics achievement with structural equation modeling]. *H. U. Journal of Education, 41*, 313-324.

Özkan, Y. (2013). *Veri madenciliği yöntemleri [Data mining methods]* (2nd ed). İstanbul: Papatya Publishing.

Öztemel, E. (2012). *Yapay sinir ağları [Artificial neural networks]* (1st ed). İstanbul: Papatya Publishing.

Öztürk, S., Coşkun, A., & Dirsehan, T. (2012). Fırsat sitelerine yönelik e-sadakati belirleyen boyutların incelenmesi [Analyzing dimensions determining e-loyalty towards daily deal sites]. *Eskişehir Osmangazi University Journal of Economics and Administrative Sciences*, *7*(2), 217-239.

Pehlivan, G. (2006). *CHAID analizi ve bir uygulama [CHAID analysis and an application]* (Unpublished Master Thesis). Yıldız Technical University, Institute of Science, İstanbul.

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: Theory and applications. Series in Machine Perception and Artificial Intelligence.* Vol. 69, USA: World Scientific Publishing Co. Pte. Ltd.

Sabancı, K. (2013). *Şeker pancarı tarımında yabancı ot mücadelesi için değişken düzeyli herbisit uygulama parametrelerinin yapay sinir ağlarıyla belirlenmesi [Determination of variable rate herbisit application parameters with artificial neural networks for weed contention in agriculture of sugar beet]* (Unpublished Doctoral Dissertation). Selçuk University, Institute of Science, Konya.

Sadi, Ö., Uyar, M., & Yalçın, H. (2014). Lise öğrencilerinin biyoloji dersi başarılarında, cinsiyet, sınıf düzeyi ve aile yapısının rolü [The role of gender, grade level and family environment in high school students' biology achievement]. *Journal of Research in Education and Teaching*, *3*(2), 138-151.

Sağıroğlu, Ş., Beşdok, E., & Erler, M. (2003). *Mühendislikte yapay zeka uygulamaları-I, yapay sinir ağları [Artificial intelligence applications in engineering-I, artificial neural networks]* (1st ed). Kayseri: Ufuk Publishing.

Serinkan, C., & Bardakcı, A. (2007). Pamukkale Üniversitesinde çalışan öğretim elemanlarının iş tatminlerine ilişkin bir araştırma [Job satisfaction: An empirical research towards academicians working in Pamukkale University]. *Selçuk University Karaman Journal of FEAS*, *12*, 152-163.

Seven, A. (1993). *Yapay sinir ağları ile doku sınıflandırma [Tissue classification using artificial neural networks]* (Unpublished Master Thesis). İstanbul Technical University, Institute of Science, İstanbul.

Silahtaroğlu, G. (2013). *Veri madenciliği kavram ve algoritmaları [Data mining concepts and algorithms]* (2nd ed). İstanbul: Papatya Publishing.

Simpson, P. K. (1990). *Artificial neural systems foundations, paradigms, application and implementation*. Elmsford NY: Pergamon Press.

Sun, J., & Hui L. (2008). Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems*, *21*(1), 1-5.

Şahin, A. (2011). İlköğretim 6. sınıf öğrencilerinin dinleme becerisi farkındalıklarının sosyo-ekonomik düzeye göre incelenmesi [A study on 6th grade students' self-awareness on listening skills according to their socio-economic level]. *Çankırı Karatekin University Journal of Institute of Social Sciences, 2*(1), 178-188.

Şirvan, O. (2010). *Yapay sinir ağları kullanılarak retina görüntülerinden hastalık tanılama sistemi tasarımı ve gerçekleştirimi [Design and implementation of disease recognition system in retinal images using artificial neural networks]* (Unpublished Doctoral Dissertation). Ege University, Institute of Science, İzmir.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn & Bacon/Pearson Education.

Tatlıdil, H. (1996). *Uygulamalı çok değişkenli istatistiksel analiz [Applied multivariate statistical analysis]*. Ankara: Cem Web Offset.

Tepehan, T. (2011). *Türk öğrencilerinin PISA başarılarının yordanmasında yapay sinir ağı ve lojistik regresyon modeli performanslarının karşılaştırılması [Performance comparison of artificial neural network and logistic regression model in predicting Turkish students' PISA success]* (Unpublished Doctoral Dissertation). Hacettepe University, Institute of Social Sciences, Ankara.

Thomas, L. C. (2000). A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumer. *International Journal of Forecasting, 16*(2), 149-172.

Thigpen, M. K. (2000). *Data mining techniques in education: a comparison of conventional statistical linear regression and neural network based tools* (Unpublished Doctoral Dissertation). University of Alabama, Alabama, ABD.

Tolon, M. (2007). *Tüketici tatmininin yapay sinir ağları yöntemiyle ölçülmesi ve Ankara'daki perakendeci mağazaların müşterileri üzerinde bir uygulama [Measuring customer satisfaction with artificial neural networks and an application of retail consumers in Ankara]* (Unpublished Doctoral Dissertation). Gazi University, Institute of Social Sciences, Ankara.

Torun, T. (2007). *Finansal başarısızlık tahmininde geleneksel istatistiki yöntemlerle yapay sinir ağlarının karşılaştırılması ve sanayi işletmeleri üzerinde uygulama [Comparison of traditional statisticial techniques with artificial neural networks in financial failure prediction and an application on industry firms]* (Unpublished Doctoral Dissertation). Erciyes University, Institute of Social Sciences, Kayseri.

Tosun, S. (2007). *Sınıflandırmada yapay sinir ağları ve karar ağaçları karşılaştırması: Öğrenci başarıları üzerine bir uygulama [Artificial neural networks and decision tree comparison in classification analysis: An application on students' success]* (Unpublished Master Thesis). İstanbul Technical University, Institute of Science, İstanbul.

Türe, M., Kurt, İ., Yavuz, E., & Kürüm, T. (2005). Hipertansiyonun tahmini için çoklu tahmin modellerinin karşılaştırılması (Sinir ağları, lojistik regresyon ve esnek ayırma analizleri) [Comparison of multiple prediction models for hypertension (Neural network, logistic regression and flexible discriminant analyses)]. *The Anatolian Journal of Cardiology, 5*(1), 24-28.

Wheeler, M. C. (1993). *A comparative case study of neural network analysis and statistical discriminant function analysis for predicting law students passing the bar examination* (Unpublished Doctoral Dissertation). Gonzaga University Spokane, WA, USA.

Yakut, E. (2012). *Veri madenciliği tekniklerinden c5.0 algoritması ve destek vektör makineleri ile yapay sinir ağlarının sınıflandırma başarılarının karşılaştırılması: İmalat sektöründe bir uygulama [The comparison of the classification successes of the artifical neural networks through data mining techniques of C5.0 algorithm and supporting vector machines: An application in manufacturing sector]* (Unpublished Doctoral Dissertation). Atatürk University, Institute of Social Sciences, Erzurum.

Yıldırım, İ. (2000). Akademik başarının yordayıcısı olarak yalnızlık, sınav kaygısı ve sosyal destek [Loneliness, exam anxiety and social support as predictors of academic success)]. *H. U. Journal of Education*, *18*, 167-176.

Yurtoğlu, H. (2005). *Yapay sinir ağları metodolojisi ile öngörü modellemesi: Bazı makroekonomik değişkenler için Türkiye örneği [Predictive modeling with artificial neural network methodology: Turkey examples for some macroeconomic variables]* (Unpublished Master Thesis). SPO General Directorate of Economic Models and Strategic Research, Ankara.

Yüksek, A. G. (2007). *Hava kirliliği tahmininde çoklu regresyon analizi ve yapay sinir ağları yönteminin karşılaştırılması [Comparation of multiple regression analysis and neural network methods for predicting air pollution]* (Unpublished Doctoral Dissertation). Cumhuriyet University, Institute of Social Sciences, Sivas.

Zurada, J., & Lonial, S. (2005). Comparison of the performance of several data mining methods for bad dept recovery in the healthcare industry. *The Journal of Applied Business Research*, *21*(2), 37-54.